Title	Unique Peripheral Antennas in the Photosystems of the Streptophyte Alga Mesostigma viride
Author(s)	Aso, Michiki; Matsumae, Renon; Tanaka, Ayumi; Tanaka, Ryouichi; Takabayashi, Atsushi
Citation	Plant and Cell Physiology, 62(3), 436-446 https://doi.org/10.1093/pcp/pcaa172
Issue Date	2021-01-08
Doc URL	http://hdl.handle.net/2115/83775
Rights	This is a pre-copyedited, author-produced version of an article accepted for publication in Plant and Cell Physiology following peer review. The version of record is available online at: https://academic.oup.com/pcp/article/62/3/436/6070176 , http://doi.org/10.1093/pcp/pcaa172.
Type	article (author version)
File Information	Plant and Cell Physiology62(3)_436-446.pdf

This is a “clear copy” of the revised manuscript. The content is exactly the same as “marked copy” described above.

Cover page

Title: Unique peripheral antennas in the photosystems of the streptophyte alga *Mesostigma viride*.

Running Title: Unique Photosystems in *Mesostigma viride*

Corresponding Author: Atsushi Takabayashi
Institute of Low Temperature Science, Hokkaido University, N19 W8 Kita-Ku, Sapporo 060-0819, Japan
E-mail: takabayashi@pop.lowtem.hokudai.ac.jp
Tel: +81-75-706-5493; Fax: +81-75-706-5493

Subject Area: (5) photosynthesis

Number of black and white figures: 0

Number of color figures: 6

Number of tables: 0

Type and number of supplementary materials: 2 figures and 4 tables

Title: Unique peripheral antennas **in the photosystems of the streptophyte alga *Mesostigma viride.***

Running Title: The Unique Photosystems in *Mesostigma viride*

Michiki Aso[1], Renon Matsumae[1], Ayumi Tanaka[1], Ryouichi Tanaka[1], Atsushi Takabayashi[1]

[1]Institute of Low Temperature Science, Hokkaido University, N19 W8 Kita-ku, Sapporo 060-0819, Japan

Abbreviations: CN, clear-native; α-DDM, α-dodecyl maltoside; Isoseq, isoform sequencing; LHC, light-harvesting complex; PAGE, polyacrylamide gel electrophoresis

[Abstract]

Land plants evolved from a single group of streptophyte algae. One of the key factors needed for adaptation to a land environment is the modification of the peripheral antenna systems of photosystems. Here, the photosystems of *Mesostigma viride*, an earliest-branched streptophyte alga, were analyzed to gain insight into their evolution. Iso-seq and phylogenetic analyses of Light-Harvesting Complexes (LHCs) revealed that *M. viride* possesses three algae-specific LHCs, including algae-type LHCA2, LHCA9, and LHCP; while the streptophyte-specific LHCB6 was not identified. These data suggest that the acquisition of LHCB6 and the loss of algae-type LHCs occurred after the *M. viride* lineage branched off from other streptophytes. Clear-native (CN)-PAGE resolved the photosynthetic complexes, including the PSI-PSII megacomplex, PSII-LHCII, two PSI-LHCI-LHCIIs, PSI-LHCI, and the LHCII trimer. Results indicated that the higher-molecular weight PSI-LHCI-LHCII likely had more LHCII than the lower-molecular weight one, a unique feature of *M. viride* photosystems. CN-PAGE coupled with mass spectrometry strongly suggested that the LHCP was bound to PSII-LHCII, while the algae-type LHCA2 and LHCA9 were bound to PSI-LHCI, both of which are different from those in land plants. Results of the present study strongly suggest that *M.*

viride photosystems possess unique features that were inherited from a common ancestor of streptophyte and chlorophyte algae.

[Introduction]

Photosystems (PSs) in oxygenic photosynthetic organisms are composed of core and peripheral antenna complexes. While the peripheral antenna complex in these organisms is highly divergent, the core complex is highly conserved (Green and Durnford, 1996; Neilson and Durnford, 2010). The diversity of the peripheral antenna system contributes to the ability of photosynthetic organisms to adapt to different light environments due to the multiple roles it plays in photosynthesis, including the harvesting of light energy and transferring it to the core complex. The variety of light spectra available in different environments is one of the driving forces behind the evolution of different peripheral antenna systems in photosynthetic organisms. (Croce and Amerongen, 2014; Stomp et al., 2007). The thermal dissipation of excess light energy, which occurs mainly in peripheral antennas (Ruban, 2018), is also necessary to avoid photooxidative damage. Therefore, photosynthetic organisms have also evolved a variety of thermal dissipation mechanisms in response to different environments (Giovagnetti and Ruban, 2018; Goss and Lepetit, 2015; Niyogi and Truong, 2013; Wobbe et al., 2016).

Notable differences in peripheral antenna systems have been reported, even within the green plants of photosynthetic organisms, and especially for PSI (Pan et al., 2020; Suga and Shen, 2020). In vascular plants, PSI retains four LHCI (Lhca1-Lhca4) proteins as a peripheral antenna and transiently gains a few extra LHCII proteins (Mazor et al., 2017; Pan et al., 2018; Qin et al., 2015). PSI in chlorophyte algae typically have a greater number of LHCI proteins forming their peripheral antenna than vascular plants. For example, PSI in *Chlamydomonas reinhardtii* has ten LHCI (two Lhca1 and one Lhca2-Lhca9) proteins and one mobile LHCII trimer that can transiently bind to the antenna (Su et al., 2019; Suga et al., 2019). A similar antenna structure has also been reported in *Bryopsis corticulans* (Qin et al., 2019). Chlorophyte algae may typically have a larger size PSI antenna than vascular plants as an adaptation to aquatic environments where light intensity is weaker than in terrestrial environments (Suga and Shen, 2020). In contrast, *Dunaliella salina* has a “mini-PSI” that does not possess a second LHCI belt unlike other reported chlorophyte algae (Perez-Boerema et al. 2020). Since *D. salina* has the ability to survive high salinity and high light environments (Perez-Boerema et al. 2020), the unique PSI structure may represent an adaptation to an extreme habitat. In the model moss plant *P. patens*, PSI can possess five (Pinnola et al., 2018) or eight (Iwai et al., 2018) more LHCI proteins than occurs in PSI-LHCI in vascular plants, plus one additional mobile LHCII trimer. The number of LHC proteins in *P. patens* PSI can change in response its photosynthetic status (Iwai et al., 2015; Pinnola et al., 2018). The changes in the PSI antenna size reflect the variable light environment that is typically experienced by mosses, which is characterized by low irradiance complemented by sunflecks (Pinnola et al., 2018). These reports reveal that the peripheral antenna system of PSI ~~I~~ in green plants exhibit large species-specific differences that apparently reflect their adaptation to different light environments. However, to the best of our knowledge, detailed investigations on the peripheral antenna system of PSI in streptophyte algae have not been reported so far.

Significant diversity is also present in thermal dissipation mechanisms in green plants. Thus far, PsbS- and LHCSR-dependent thermal dissipation mechanisms have been reported (Niyogi and Truong, 2013). The LHCSR-dependent mechanism is the primary process used for thermal dissipation in chlorophyte algae, although they possess both PsbS and LHCSR proteins (Correa-Galvis et al., 2016; Tibiletti et al., 2016). In contrast, vascular plants only possess PsbS-dependent thermal dissipation, having lost the LHCSR protein during their evolution (Giovagnetti and Ruban, 2018; Niyogi and Truong, 2013; Pinnola, 2019). Notably, both PsbS and LHCSR play an essential role in thermal dissipation in the model moss plant, *P. patens* (Pinnola, 2019).

Phylogenetically, *P. patens* represents an early-branched land plant and possesses thermal dissipation characteristics that are intermediate between chlorophyte algae and vascular plants. In addition, the contribution of PsbS to the thermal dissipating capacity in land plants is much higher than it is in chlorophyte algae. How thermal dissipation mechanisms have changed over the course of evolution from ancestral green algae to land plants, however, has not been fully elucidated. This can be mainly attributed to the absence of a suitable streptophyte alga species that could serve as a model for photosynthesis research.

M. viride is one of the earliest-branched freshwater streptophyte algae, although its exact phylogenetic position in green plants is still under debate (Lemieux et al., 2007, 2000; Li et al., 2020; Wang et al., 2020). Nevertheless, it is worth testing whether this basally branching alga retains a feature present in a common ancestor of streptophytes and chlorophytes. Therefore, in the present study, the photosystems of the earliest-branched *M. viride* were characterized to gain insight into the structure of photosystems of a common ancestor of streptophytes and chlorophytes. Results revealed that considerable changes occurred in photosystems after the *M. viride* lineage branched off from other streptophyte algae.

[Results]

Separation of photosystems by ClearNative (CN)-PAGE

Clear-Native (CN)-PAGE is a powerful technique that enables one to separate protein complexes while retaining their structure. Here, the *M. viride* photosystems were separated using amphipol-based CN-PAGE (Furukawa et al., 2019) after solubilization with a mild detergent, dodecyl maltoside (α-DDM). As a result, the PSI-PSII megacomplex, the PSII-LHCII supercomplexes, the PSI-LHCI-LHCII bands, the PSI-LHCI, and the LHCII trimer were resolved (Fig. 1A). The identification of the separated bands was accomplished using 2D-CN/SDS-PAGE followed by immunoblot analysis (Fig. 1B) and silver-staining (Fig. 1C), as described in previous studies (Järvi et al., 2011; Takabayashi et al., 2011). The identification of two PSI-LHCI-LHCII bands were confirmed by further analysis described in a later section of this report. The overall band profile (Fig. 1A) was similar to the profile for *P. patens* presented by Furukawa et al. (2019), however, a substantial difference was evident for PSI-LHCI-LHCII. Two PSI-LHCI-LHCII bands were found in *M. viride* (Fig. 1A), whereas only one PSI-LHCI-LHCII band was found in the profile of *P. patens* presented by Furukawa et al. (2019). The presence of two PSI-LHCI-LHCII bands appears to be a unique characteristic of the PSs of *M. viride* since land plants also exhibit one PSI-LHCI-

LHCII band in BN-PAGE and CN-PAGE gels (Järvi et al., 2011; Pesaresi et al., 2009; Pinnola et al., 2018).

Iso-seq analysis to provide a protein database for identification of *M. viride* LHC proteins by MS

No information is available on whether commercially-available LHC antibodies react with *M. viride* LHCs. Therefore, mass spectrometry (MS) analysis of the protein complexes resolved in the CN-PAGE was conducted, as described in our previous papers (Takabayashi et al., 2017, 2013), to elucidate the composition of the LHC proteins in the resolved PSI and PSII supercomplexes. This approach can be used to estimate the positions of the bands of protein complexes separated by CN-PAGE, especially for the relatively high expressed proteins. The number of *M. viride* photosynthetic protein sequences available in public databases was limited at the time of our MS analysis. Therefore, Isoform sequencing (Iso-seq) analysis was used to obtain a transcriptome using PacBio sequencing, which provides longer and more complete sequence information relative to short-read sequencing platforms such as Illumina (Zhao et al., 2019). Full-length cDNA sequences enabled us to estimate full-length amino acid sequences with high reliability, a feature that is advantageous for protein identification by MS.

The iso-seq analysis (see Materials and Methods) allowed us to identify nine PSI core proteins, 11 PSII core proteins, and 14 LHC proteins (Table S1). The small subunits of PSI and PSII were not identified because the TransDecoder software identified ORFs that are at least 100 amino acids long using the default settings. Notably, the rate of false positives drastically increased when a shorter minimum length was used. Therefore, default settings were used to predict ORFs in our analysis to obtain reliable results.

Phylogenetic analysis of LHC proteins

A BLAST query identified 14 putative *M. viride* LHC proteins (Table S1). The annotation of the LHC proteins derived from the BLAST query, however, may not be reliable as considerable sequence similarity has been reported among different LHC proteins. Therefore, a phylogenetic tree of LHC sequences was constructed based on the alignment shown in Fig. S1 and each LHC sequence of *M. viride* was annotated based on the nomenclature of *A. thaliana* LHC proteins, except for LHCBMs (corresponding to major LHCII (LHCB1, LHCB2, and LHCB3)) and three algae-specific LHCs (algae-type LHCA2, LHCA9, and LHCP). Algae-type LHCA2 is conserved among chlorophytes, however, it is not closely related to plant-type LHCA2 (Fig. 2). Therefore, we annotated it as algae-type LHCA2 in this study to avoid any misrepresentation. Based on the phylogenetic analysis, *M. viride* possesses LHCBMs, LHCB4, LHCB5, LHCA1, two LHCA2, LHCA3, and three algae-specific LHCs as peripheral antenna proteins. The *M. viride* LHC sequences identified by the Iso-seq analysis exhibited good correspondence with previously reported *M. viride* LHCs (Koziol et al., 2007) (Fig. 2). Some differences, however, were observed relative to the previous classification (Koziol et al., 2007). These differences are likely due to the considerable sequence similarities among LHCs, as the LHCs in green lineages rapidly diversified during the early evolution of green algae. LHCs that were classified differently in the previous report (Koziol et al., 2007) and our present study have been marked with an asterisk (Fig. 2).

The Iso-seq analysis did not detect an LHCB6 gene, even though LHCB6 is highly conserved among streptophyte algae and land plants (Kouřil et al., 2016). This result is consistent, however, with a previous study by Koziol et al. (2007) in which LHCB6 sequences were also not identified in *M. viride,* and with the *M. viride* RNA-seq data generated in the 1000 plant transcriptomes (1kP) project (Carpenter et al., 2019; Leebens-Mack et al., 2019). In contrast, LHCP proteins were found in *M. viride*

(Fig.2), which is also consistent with the previous report by Koziol et al. (2007). LHCP is a unique LHC protein found in prasinophyte algae, including *Ostreococcus tauri* and *Micromonas pusilla*. Prasinophyte algae is a group of early-diverging chlorophyte algae, which gave rise to core chlorophyte algae including *C. reinhardtii*. It is very notable that *M. viride*, an early-divergent streptophyte alga, also encodes LHCP. Previous studies of a model prasinophyte alga, *O. tauri*, reported that it contains LHCP as the peripheral antenna of PSI-LHCI (Six et al., 2005; Swingley et al., 2010), although no detailed studies have been conducted on LHCP-containing photosystems. A possible loss of LHCB6 and the presence of LHCP in *M. viride* suggest that the LHC composition in *M. viride* is likely inherited from a common ancestor of streptophyte and chlorophyte algae.

The distribution of LHC proteins in PSI and PSII

No studies have been conducted on the distribution of *M. viride* LHCs in PSI and PSII, although they have been identified and classified (Koziol et al., 2007). Therefore, we determined the protein composition of the separated photosystems using MS. As expected, PSI, PSII, and LHC proteins were detected in the PSI-PSII megacomplex band (Table S2). Also as expected, PSI and LHC proteins were detected in the PSI-LHCI band (Table S3). Unexpected proteins were rarely detected in these two photosystems, suggesting that contamination from the other protein complexes was limited. The PSII-LHCII band also contained PSII and LHC proteins, although some contamination from PSI subunits and the NDH-like complex subunits were observed to some extent (Table S4).

Importantly, the bands representing protein complexes on electrophoresis gels are often distorted by tailing at both ends. The protein complexes in the tailings are also detected by high-resolution of MS, which is the main source of contamination in a band of interest. To predict unknown protein complexes and to estimate the protein compositions of known protein complexes, a protein migration profile was used after the native-PAGE coupled with MS (Takabayashi et al., 2017, 2013). A protein migration profile is a plot where the y-axis represents the amount of protein estimated from MS data, while the x-axis represents the migration distance on a native-PAGE gel. A peak in the protein migration profile has been demonstrated to correspond to the position of the band (Helbig et al., 2009; Remmerie et al., 2011; Wessels et al., 2009). This approach allows one to determine if the proteins in a band detected by MS are contaminants from other protein complexes by verifying their peak positions in protein migration profiles (Helbig et al., 2009; Müller et al., 2016; Takabayashi et al., 2017; Wessels et al., 2009).

In the present study, migration profiles were constructed for PSI, PSII, and LHC proteins based on MS data. A normalized spectral abundance factor (NSAF) method, which is a label-free quantification method (Zybailov et al., 2006), was used to estimate the amount of protein from MS data. First, the migration profiles of PSI and PSII were compared. The distance between PSII-LHCII and PSI-LHCI-LHCII on the CN-PAGE gel was relatively close, and PSI-LHCI-LHCII bands contained considerable amounts PSI-LHCI-LHCII proteins. Nevertheless, it was possible to isolate those bands using their migration profiles (Fig. 3A). The PSII-LHCII peak was observed in the position of gel slices 8 and 9 (Fig. 3A), where the band was distinctly visible in the CN-PAGE gel (Fig. 1A). The PSI-LHCI-LHCII and the PSI-LHCII peaks were observed in the positions of gel slices 12 to 15 (Fig. 1A), where those bands were also visible in the CN-PAGE gel (Fig. 1A). These data indicate that it is possible to distinguish PSI-bound LHCs and PSII-bound LHCs by comparing their migration profiles with the migration profiles of PSI and PSII.

Comparing the migration profiles of LHCs with those of PSI and PSII resulted in the classification of LHCs into two groups. One group included LHCA1, LHCA2, LHCA3, algae LHCA2, and LHCA9 proteins, whose peaks overlapped with the peaks of PSI-LHCI and PSI-LHCI-LHCII (Fig. 3B). The other group included LHCB4, LHCB5, and LHCP proteins, whose peaks overlapped with the peak of PSII-LHCII (Fig. 3C). These data strongly suggest that PSI-LHCI includes the former LHCs, whereas PSII-LHCII includes the latter LHCs. Notably, the migration profile of LHCBM proteins appeared to be different from the profiles of other LHCs. Its highest peak corresponded with LHCII trimers, suggesting that LHCBM is the major component of the LHCII trimer.

Separation of PSI-LHCI-LHCII complexes by sucrose density gradient

Two PSI-LHCI-LHCII bands were observed on CN-PAGE gels. A “two-step” separation of those bands was performed to improve the purity of the bands observed on CN-PAGE using sucrose density gradient centrifugation as the first “rough” separation step prior to further separation by CN-PAGE. As a result, three bands (B1-B3) containing photosynthetic pigments were identified after separation by sucrose density gradient centrifugation (Fig. 4). After subsequent separation by CN-PAGE, PSI-LHCI-LHCII, in addition to PSI-LHCI, were found to be present in the middle band (B2) (Fig. 4), while the bottom band (B3) primarily contained PSII-LHCII, and the upper band (B3) contained LHCII trimer (Fig. 4).

The three PSI-containing bands present in the CN-PAGE gel were further separated by SDS-PAGE after they were cut out of the CN-gel to compare their composition. Two replicates for each PSI-LHCI-LHCII band were subjected to 2D-SDS-PAGE (Fig. 5). Silver-staining of the resulting gels revealed the PsaA/PsaB heterodimer, the PsaA and the PsaB monomers, LHC monomers (c.a. 20kDa), and the small PSI subunits. A portion of PsaA and PsaB remained as a heterodimer and migrated slowly in the gel despite the inclusion of SDS (Fig. 5). A comparison of the three CN-PAGE bands containing PSI indicated that the stoichiometry of the PSI and LHCI subunits were similar, while the amounts of LHCII were greatest in the top band and lowest in the bottom band. These results indicate that variations in the size of the three CN-PAGE bands containing PSI were attributed to differences in the number of LHCII (mainly LHCBM) subunits bound to each PSI-LHCI complex. Importantly, the identification of the LHCII band was based on the separation pattern of the B1 and B3 bands containing LHCII trimers by the subsequent 2D-SDS-PAGE (Fig. S2). These data suggest that the larger PSI-LHCI-LHCII possesses more LHCII than the smaller one.

[Discussion]

Unique peripheral antenna systems of *M. viride*

We classified the LHCs in the peripheral antenna of PSI and PSII. LHCB4 and LHCB5 were identified as the peripheral antenna of PSII-LHCII, similar to land plants. In contrast, LHCB6 was not identified in the Iso-seq analysis in this study. The absence of LHCB6 is consistent with the reported absence of LHCB6 sequences in *M. viride* LHC sequences (Koziol et al., 2007) and is also consistent with the *M. viride* RNA-seq data generated in the 1KP project (Carpenter et al., 2019; Leebens-Mack et al., 2019). Therefore, it is likely that *M. viride* does not possess LHCB6, although we do not exclude the possibility that the expression level of LHCB6 is relatively low compared to other LHCs, and thus was undetected. LHCB6 is widely distributed among streptophyte algae and land plants but has not been found in chlorophyte algae, including *C. reinhardtii* (Grebe et al., 2019; Kouřil et al., 2016). The putative loss of LHCB6 in *M. viride* suggests that LHCB6 was acquired during evolution after the divergence of *M. viride*.

M. viride PSII possesses LHCP (Koziol et al., 2007), the main peripheral antenna protein in the model prasinophyte alga, *O. tauri* (Six et al., 2005; Swingley et al., 2010), whereas LHCP has not been found in land plants. The peaks in the migration profile of LHCP corresponded with the PSI-PSII megacomplex and the PSII-LHCII (Fig. 3C). These data suggest that LHCP binds to PSII-LHCII in *M. viride*, although further biochemical studies will be required to confirm this possibility. It should be noted that detection of the LHCP peak at the position of the PSI-PSII megacomplex could be explained by assuming co-migration of the PSI-PSII megacomplex and PSII-LHCII at the top part of CN-PAGE (Fig. 3C). Such co-migration of these complexes was actually observed in *A. thaliana* (Yokono et al. 2019).

A previous report revealed that LHCP, which is conserved among algae in the Mamiellophyceae, including *O. tauri* and *M. pusilla,* is a major LHC antenna protein in PSI-LHCI in the model prasinophyte algae, *O. tauri* (Swingley et al., 2010). It has also been suggested that LHCP functions as the peripheral antenna of PSII-LHCII, although no direct evidence has been provided (Six et al., 2005; Swingley et al., 2010). Our data support the idea that LHCP can function as the peripheral antenna of PSII-LHCII (Fig. 3). Collectively, the data suggest that ancestral green algae possessed LHCP proteins as their peripheral antenna because the earlier-branched streptophyte *M. viride* and the prasinophyte *O. tauri* possess LHCP proteins in their photosystems. The LHCP-containing photosystem in *M. viride* was likely inherited from a common ancestor of chlorophyte and streptophyte algae that has been lost during the evolution of streptophyte algae to land plants. Considering that the loss of LHCP and the acquisition of the LHCB6 seemed to occur concomitantly according to our phylogenetic analysis, it is possible to hypothesize that LHCP in streptophyte PSII-LHCII have been replaced by LHCB6 in land plants. Further investigation will be required, however, to confirm this hypothesis.

In regards to the peripheral antenna of PSI, two PSI-LHCI-LHCII supercomplexes were stably observed after their separation by CN-PAGE. LHCII was more abundant in the higher molecular weight band than it was in the lower molecular weight band, suggesting that the higher molecular PSI-LHCI-LHCII possesses at least two LHCII trimers. It is not presently known if at least two PSI-bound LHCII trimers change their binding to PSI in response to different light environments, i.e., if they are involved in state transitions. Since the binding of LHCII trimer to PSI was involved in the state transition according to previous studies in green algae and land plants (Drop et al., 2014; Pesaresi et al., 2009; Pinnola et al., 2018, 2018; Pribil et al., 2010; Takahashi et al., 2006), we hypothesize that the additional *M. viride* LHCII trimers are also involved in the state transition. We do not exclude the possibility, however, that the smaller form of PSI-LHCI-LHCII represents an artifact caused by the dissociation of LHCII from the larger form. Further studies are required to confirm this supposition.

In addition, PSI-LHCI in *M. viride* possesses algae-type LHCA2 and LHCA9 proteins, which are conserved among chlorophytes including *C. reinhardtii* (Su et al., 2019; Suga et al., 2019) and *Bryposis corticulans* (Qin et al., 2019), but not conserved among streptophytes (Fig. 2). The loss of these algae-specific LHCs must have occurred after the *M. viride* lineage branched off from the other streptophyte algae, although further studies will also be required to confirm this possibility.

We constructed a hypothetical model of the antenna structure of PSI-LHCI-LHCII in *M. viride* using the reported structure of PSI-LHCI in chlorophyte algae and land plants as references (Fig. 6). Based on their sequence similarities, the binding manner of LHCA1, two LHCA2s, and LHCA3 proteins to the PSI core in *M. viride* may be similar to the binding of LHCA1, LHCA2, LHCA3, and LHCA4 proteins to the PSI core in vascular plants. Likewise, the binding manner of algae-type LHCA2 and LHCA2 to PSI-LHCI may be similar to the binding mechanism that occurs in PSI-LHCI in *C. reinhardtii* (Su et al., 2019; Suga et al., 2019) and *B. corticulans* (Qin et al., 2019). The binding site of the second LHCII trimer to PSI-LHCI, however, is unknown, as a PSI-LHCI with two LHCII trimers has not, to our knowledge, been reported. Similarly, the binding site of LHCP to PSII-LHCII is also unknown as no structural studies on LHCP-bound PSI-LHCI have been reported. Given the phylogenetic position of *M. viride* and its unique antenna system in PSI and PSII, further structural analyses of *M. viride* PSI-LHCI-LHCII are warranted and would significantly contribute to our understanding of the changes that occurred in photosystems during the evolution of a common ancestor of chlorophytes and streptophytes to land plants.

[Materials and Methods]

Algal strain and culture conditions

M. viride strain, NIES-296, was obtained from the National Institute for Environmental Studies (NIES) (Ibaraki, Japan). The strain was cultured in 500 mL of C medium in a 1L Erlenmeyer flasks at 22°C under a 14-h photoperiod of 20 μmol photons m^{-2} s^{-1}, as previously described by (Kunugi et al., 2016).

RNA extraction

One liter of a *M. viride* culture was centrifuged at 10,000 x *g* for 10 min at 4 °C. The pellet was resuspended in 0.5 mL RLT buffer (RNeasy Plant Mini Kit, Qiagen) supplemented with 1% 2-mercaptoethanol. The suspension was then transferred to a 2-ml vial containing 500 mg of glass beads (0.5 mm diameter). The vial was subjected to 10s disruption treatments using a Mini-Bead Beater (Merck, Germany) and the suspension was then centrifuged at 21,600×g at 4 °C for 5 min. Supernatants from the vials were used for total RNA isolation with a RNeasy Plant Mini Kit (Qiagen), according to the manufacturer's instructions.

Full-length isoform sequencing (Iso-seq) using PacBio data

The integrity of extracted total RNA was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc). The Iso-seq library was prepared according to the protocol described by Pacific Biosciences (PN 101-070-200) using a SMARTer PCR cDNA Synthesis Kit (Clontech) and SMRTbell Template Prep Kit 1.0 SPv3 (Pacific Biosciences). Sequencing was performed on a PacBio Sequel platform.

Raw sequence data were processed using SMRT Link v6.0 (Pacific Biosciences) software and then further processed using IsoSeq3 (version 3.1) software tools (https://github.com/PacificBiosciences/IsoSeq) with default parameters. Circular consistency sequence (CCS) reads were generated from subread sequences. Full-length cDNA reads were then selected by finding the 5’ and 3’ primers or polyA tail using the lima tool in IsoSeq3. After trimming of the polyA tail and removal of the concatemer sequences, 84,545 full-length cDNA reads were obtained. Finally, 7,209 high-quality consensus full-length cDNA sequences were obtained after isoform clustering and polishing.

Estimation of full-length amino acid sequences using the Iso-seq data

TransDecoder (https://github.com/TransDecoder/TransDecoder/) software was used to identify full-length protein sequences in the 7,209 full-length cDNAs. Candidate open reading frames (ORFs) with a minimum length of 100 amino acids were identified using the TransDecoder.LongOrfs module. Candidate ORFs were validated by BLASTP queries using an e-value cutoff of 10^{-5} against data protein database comprising proteins of *Arabidopsis thaliana*, *Chara braunii*, *Chlamydomonas reinhardtii*, *Coccomyxa subellipsoidea, Cyanidioschyzon merolae, Cyanophora paradoxa*, *Klebsormidium flaccidum*, *Oryza sativa*, *Ostreococcus lucimarinus, Physcomitrella patens,* and *Selaginella moellendorffii* within the Phytozome database (v12.1)(Goodstein et al., 2012). After the blast queries, 5,829 protein sequences with significant blast hits were obtained using the TransDecoder.Predict module. CD-HIT (Fu et al., 2012) software was employed to remove redundant protein sequences using a cutoff value of 0.9. A total of 3,198 protein sequences were obtained using CD-HIT software. These protein sequences were used as a *M. viride* protein database for MS queries.

Phylogenetic analysis of *M. viride* LHC proteins

M. viride LHC proteins were identified by NCBI-BLASTP homology queries against the *M. viride* protein sequence database described above. *A. thaliana* and *C. reinhardtii* LHC proteins were used as query sequences in the BLASTP searches. The LHC sequences of *Arabidopsis thaliana*, *Marchantia polymorpha*, *Klebsormidium flaccidum*, *Ostreococcus lucimarinus*, and *Chlamydomonas reinhardtii* were obtained from the Phytozome database (https://phytozome.jgi.doe.gov/). The LHCs of *Cyanidioschyzon merolae* and *Pyropia yezoensis* were used as the outgroups in the phylogenetic tree analysis. Amino acid sequences of LHCs were aligned using the MAFFT algorithm (Katoh and Standley, 2013). Trimming of the alignment was done by a ClipKIT program (Steenwyk et al. 2020) with default parameters. A maximum likelihood (ML) phylogenetic tree including *M. viride* LHCs was constructed using W-IQ-TREE (Trifinopoulos et al., 2016) software under the best-fitting model (LG+F+I+G4). Ultrafast bootstrap values (1,000 replicates) are shown below the branches.

Thylakoid membrane preparation

A pellet obtained from the centrifugation (10,000 x *g* for 10 min at 4 °C) of 1 L of a *M. viride* culture was suspended in 2 mL BN- solubilization buffer (50 mM imidazole/HCl (pH 7.0), 20% glycerol) with 1% Protease Inhibitor Cocktail for plant cell lysate

(Merck, Germany). The suspension was transferred to a 2-ml vial containing 500 mg of glass beads (0.5 mm diameter). The vial was first immersed in liquid nitrogen and the cells were subsequently subjected to three 10s disruption treatments using a Mini-Bead Beater (Merck, Germany). The resulting sample was centrifuged at 21,600×g at 4 °C for 5 min, and the supernatant was suspended in BN- solubilization buffer. The suspension was centrifuged at 200×g at 4 °C for 1 min, and the supernatant was centrifuged at 21,600×g at 4 °C for 5 min to obtain the thylakoid membranes.

Clear-Native (CN)-PAGE

CN-PAGE was performed as previously described (Furukawa et al. 2019). A linear 4–13% gradient polyacrylamide gel was used as the separation gel, and a 3.5% polyacrylamide gel was used as the sample gel. Thylakoid membranes were solubilized in 1% α-DDM (α-dodecyl maltoside) at 4 °C for 1 min. After centrifugation at 21,600×g at 4 °C for 5 min, the supernatant (roughly 5 μg chlorophyll equivalent) was supplemented with 1% Amphipol A8-35 and loaded onto CN-PAGE gels. An anode buffer (25mM imidazole/HCl (pH 7.0)) and cathode buffer (50mM Tricine, 7.5mM imidazole) were used for the electrophoresis.

Two-dimensional (2D)-CN/SDS-PAGE

CN-gel strips were soaked in denaturation buffer (1% SDS, 50mM DTT) for 30 min, and the 2D-SDS-PAGE was performed using a 14% polyacrylamide gel with 4M urea. The proteins on the 2D-CN/SDS-gel were visualized by silver-staining using a Pierce Silver Stain kit (ThermoFisher Scientific, USA), according to the manufacturer’s instructions. Alternatively, the obtained 2D-gels were used in the immunoblot analyses.

Immunoblot analysis

Proteins from the 2D-CN/SDS-gel were transferred to a polyvinylidene fluoride membrane (PolyScreen PVDF transfer membrane, PerkinElmer Life Sciences, MA, USA). Protein detection using specific antibodies was performed using Western Lightning Plus-ECL (PerkinElmer Life science, MA, USA) reagent. All antibodies, except anti-PsbB (AS04 038)), were purchased from Agrisera (Vännäs, Sweden).

Sucrose density gradient

M. viride thylakoid membranes were solubilized in 1% α-DDM and then loaded onto a continuous sucrose density gradient (0.3M sucrose to 1.3M sucrose in a buffer containing 25mM MES-KOH (pH 6.5) and 0.1% GDN (GDN101, Anatrace)). The samples were then subjected to ultracentrifugation at 72,000 x *g* using a S-65T rotor (Hitachi-koki, Japan) for 15 h at 4 °C.

LC-MS/MS analysis

CN-PAGE gel strips were cut horizontally as shown in Fig.1A. All gel pieces were subjected to in-gel digestion with trypsin and analyzed to identify their peptides by LC-MS/MS using Orbitrap Elite mass spectrometry (Thermo Fisher Scientific, Waltham, MA, USA) coupled with Thermo Easy-nLC (Thermo Fisher Scientific, Waltham, MA, USA). Each sample was loaded onto a C18-reversed phase EASY-Column (0.1 mm × 20 mm, 5 μm particle size, 120 Å pore size), before separation on a C18 Tip column (75 μm × 120 mm; Nikkyo Technos, Tokyo, Japan). The samples were separated by a gradient formed by solvent A (0.1% formic acid) and solvent B (acetonitrile in 0.1% formic acid) at a flow rate of 300 nL/min. The gradient separation setting was as follows: 0-1min, 0–5% B; 1-12 min, 5–35% B; 12–25 min, 35%–90% B; 90% B; 25–45 min.

Proteins in each sample were identified using SearchGUI (version 3.3.15) software and an andromeda search engine (Barsnes and Vaudel, 2018). The 3,198 *M. viride* protein sequences identified in the Iso-seq analysis were used for protein identification. PeptideShaker (version 1.16.40) software with default settings was used to validate the protein identifications (Vaudel et al., 2015).

The normalized spectral abundance factor (NASF) (Zybailov et al., 2006) calculated by the PeptideShaker program was used as a label-free quantification method based on the LC-MS/MS data to estimate the abundances of the identified proteins in each gel slice. As previously described, protein migration profiles were then generated by plotting the NASF values on the y-axis, and plotting the gel slice number on the x-axis (Takabayashi et al., 2017, 2013).

[Acknowledgement]

Funding

This work was supported by JSPS (Japan Society for the Promotion of Science) KAKENHI Grant numbers 16H06554 to R. Tanaka.

Data availability

Raw data of the Iso-seq analysis have been deposited to the DDBJ Sequence Read Archive (DRA) (https://www.ddbj.nig.ac.jp/dra/index.html) under the accession number PRJDB10366. In addition, the assembled full-length cDNA sequences of PSI, PSII, and

LHC sequences have been deposited to DDBJ/EMBL/GenBank as Transcriptome Shotgun Assembly (TSA) data under the accession numbers from ICQU01000001 to ICQU01000034.

Disclosures

Conflicts of interest: The authors declare no conflicts of interest.

[References]

Barsnes, H., and Vaudel, M. (2018) SearchGUI: A Highly Adaptable Common Interface for Proteomics Search and de Novo Engines. *J Proteome Res*. 17: 2552–2555.

Carpenter, E.J., Matasci, N., Ayyampalayam, S., Wu, S., Sun, J., Yu, J., et al. (2019) Access to RNA-sequencing data from 1,173 plant species: The 1000 Plant transcriptomes initiative (1KP). *Gigascience*. 8.

Correa-Galvis, V., Redekop, P., Guan, K., Griess, A., Truong, T.B., Wakao, S., et al. (2016) Photosystem II Subunit PsbS Is Involved in the Induction of LHCSR Protein-dependent Energy Dissipation in *Chlamydomonas reinhardtii*. *J Biological Chem*. 291: 17478–87.

Croce, R., and Amerongen, V.H. (2014) Natural strategies for photosynthetic light harvesting.

Drop, B., K.N., S.Y., Boekema, E.J., and Croce, R. (2014) Consequences of state transitions on the structural and functional organization of Photosystem I in the green alga *Chlamydomonas reinhardtii*. *Plant J*. 78: 181–191.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*. 28: 3150–3152.

Furukawa, R., Aso, M., Fujita, T., Akimoto, S., Tanaka, R., Tanaka, A., et al. (2019) Formation of a PSI–PSII megacomplex containing LHCSR and PsbS in the moss *Physcomitrella patens*. *J Plant Res*. 132: 867–880.

Giovagnetti, V., and Ruban, A.V. (2018) The evolution of the photoprotective antenna proteins in oxygenic photosynthetic eukaryotes. *Biochem Soc T*. 46: 1263–1277.

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., et al. (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 40: D1178–D1186.

Goss, R., and Lepetit, B. (2015) Biodiversity of NPQ. *Journal of plant physiology*. 172: 13–32.

Grebe, S., Trotta, A., Bajwa, A.A., Suorsa, M., Gollan, P.J., Jansson, S., et al. (2019) The unique photosynthetic apparatus of Pinaceae: analysis of photosynthetic complexes in *Picea abies*. *Journal of experimental botany*. 70: 3211–3225.

Green, B.R., and Durnford, D.G. (1996) The chlorophyll-carotenoid proteins of oxygenic photosynthesis. *Annu Rev Plant Phys*. 47: 685–714.

Helbig, A.O., Groot, M.J.L. de, Gestel, R.A. van, Mohammed, S., Hulster, E.A.F. de, Luttik, M.A.H., et al. (2009) A three-way proteomics strategy allows differential analysis of yeast mitochondrial membrane protein complexes under anaerobic and aerobic conditions. *Proteomics*. 9: 4787–98.

Iwai, M., Grob, P., Iavarone, A.T., Nogales, E., and Niyogi, K.K. (2018) A unique supramolecular organization of photosystem I in the moss *Physcomitrella patens*. *Nature plants*. 4: 904–909.

Iwai, M., Yokono, M., Kono, M., Noguchi, K., Akimoto, S., and Nakano, A. (2015) Light-harvesting complex Lhcb9 confers a green alga-type photosystem I supercomplex to the moss *Physcomitrella patens*. *Nature plants*. 1: 14008.

Järvi, S., Suorsa, M., Paakkarinen, V., and Aro, E.-M. (2011) Optimized native gel systems for separation of thylakoid protein complexes: novel super- and mega-complexes. *Biochemical Journal*. 439: 207–214.

Kouřil, R., Nosek, L., Bartoš, J., Boekema, E.J., and Ilík, P. (2016) Evolutionary loss of light-harvesting proteins Lhcb6 and Lhcb3 in major land plant groups – break-up of current dogma. *New Phytol*. 210: 808–814.

Koziol, A.G., Borza, T., Ishida, K.-I., Keeling, P., Lee, R.W., and Durnford, D.G. (2007) Tracing the evolution of the light-harvesting antennae in chlorophyll *a*/*b*-containing organisms. *Plant Physiol*. 143: 1802–1816.

Kunugi, M., Satoh, S., Ihara, K., Shibata, K., Yamagishi, Y., Kogame, K., et al. (2016) Evolution of green plants accompanied changes in light-harvesting systems. *Plant Cell Physiology*. 57: 1231–43.

Leebens-Mack, J.H., Barker, M.S., Carpenter, E.J., Deyholos, M.K., Gitzendanner, M.A., Graham, S.W., et al. (2019) One thousand plant transcriptomes and the phylogenomics of green plants. *Nature*. 574: 679–685.

Lemieux, C., Otis, C., and Turmel, M. (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature*. 403: 649–652.

Lemieux, C., Otis, C., and Turmel, M. (2007) A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *Bmc Biol*. 5: 2.

Li, L., Wang, S., Wang, H., Sahu, S.K., Marin, B., Li, H., et al. (2020) The genome of *Prasinoderma coloniale* unveils the existence of a third phylum within green plants. *Nat Ecol Evol*. 1–12.

Mazor, Y., Borovikova, A., Caspy, I., and Nelson, N. (2017) Structure of the plant photosystem I supercomplex at 2.6 Å resolution. *Nat Plants*. 3: 17014.

Müller, C.S., Bildl, W., Haupt, A., Ellenrieder, L., Becker, T., Hunte, C., et al. (2016) Cryo-slicing blue bative-mass spectrometry (csBN-MS), a novel technology for high resolution complexome profiling. *Mol Cell Proteomics*. 15: 669–681.

Neilson, J.A.D., and Durnford, D.G. (2010) Structural and functional diversification of the light-harvesting complexes in photosynthetic eukaryotes. *Photosynth Res*. 106: 57–71.

Niyogi, K.K., and Truong, T.B. (2013) Evolution of flexible non-photochemical quenching mechanisms that regulate light harvesting in oxygenic photosynthesis. *Current opinion in plant biology*. 16: 307–14.

Pan, X., Cao, P., Su, X., Liu, Z., and Li, M. (2020) Structural analysis and comparison of light-harvesting complexes I and II. *Biochimica Et Biophysica Acta Bba - Bioenergetics*. 1861: 148038.

Pan, X., Ma, J., Su, X., Cao, P., Chang, W., Liu, Z., et al. (2018) Structure of the maize photosystem I supercomplex with light-harvesting complexes I and II. *Science*. 360: 1109–1113.

Pesaresi, P., Hertle, A., Pribil, M., Kleine, T., Wagner, R., Strissel, H., et al. (2009) Arabidopsis STN7 kinase provides a link between short- and long-term photosynthetic acclimation. *Plant Cell Online*. 21: 2402–2423.

Perez-Boerema, A., Klaiman, D., Caspy, I., Netzer-EI, S., Amunts, A., and Nelson, N. (2020) Structure of a minimal photosystem I from the green alga *Dunaliella salina*. *Nature plants*. 6: 321-327.

Pinnola, A. (2019) The rise and fall of light-harvesting complex stress-related proteins as photoprotection agents during evolution. *J Exp Bot*. 70: 5527–5535.

Pinnola, A., Alboresi, A., Nosek, L., Semchonok, D., Rameez, A., Trotta, A., et al. (2018) A LHCB9-dependent photosystem I megacomplex induced under low light in *Physcomitrella patens*. *Nature plants*. 4: 910–919.

Pribil, M., Pesaresi, P., Hertle, A., Barbato, R., and Leister, D. (2010) Role of plastid protein phosphatase TAP38 in LHCII dephosphorylation and thylakoid electron flow. *Plos Biol*. 8: e1000288.

Qin, X., Pi, X., Wang, W., Han, G., Zhu, L., Liu, M., et al. (2019) Structure of a green algal photosystem I in complex with a large number of light-harvesting complex I subunits. *Nat Plants*. 5: 263–272.

Qin, X., Suga, M., Kuang, T., and Shen, J.-R. (2015) Structural basis for energy transfer pathways in the plant PSI-LHCI supercomplex. *Science*. 348: 989–995.

Remmerie, N., Vijlder, T.D., Valkenborg, D., Laukens, K., Smets, K., Vreeken, J., et al. (2011) Unraveling tobacco BY-2 protein complexes with BN PAGE/LC–MS/MS and clustering methods. *J Proteomics*. 74: 1201–1217.

Ruban, A.V. (2018) Light harvesting control in plants. *Febs Lett*. 592: 3030–3039.

Six, C., Worden, A.Z., Rodríguez, F., Moreau, H., and Partensky, F. (2005) New insights into the nature and phylogeny of prasinophyte antenna proteins: *Ostreococcus tauri*, a Case Study. *Mol Biol Evol*. 22: 2217–2230.

Steenwyk, J.L., Buida 3rd, T.J., Li, Y., Shen, XX., Rokas, A. (2020) ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* 18: e3001007.

Stomp, M., Huisman, J., Stal, L.J., and Matthijs, H.C.P. (2007) Colorful niches of phototrophic microorganisms shaped by vibrations of the water molecule. *Isme J*. 1: 271–282.

Su, X., Ma, J., Pan, X., Zhao, X., Chang, W., Liu, Z., et al. (2019) Antenna arrangement and energy transfer pathways of a green algal photosystem-I–LHCI supercomplex. *Nature Plants*. 5: 273–281.

Suga, M., Ozawa, S.-I., Yoshida-Motomura, K., Akita, F., Miyazaki, N., and Takahashi, Y. (2019) Structure of the green algal photosystem I supercomplex with a decameric light-harvesting complex I. *Nature Plants*. 5: 626–636.

Suga, M., and Shen, J.-R. (2020) Structural variations of photosystem I-antenna supercomplex in response to adaptations to different light environments. *Current Opinion in Structural Biology*. 63: 10–17.

Swingley, W.D., Iwai, M., Chen, Y., Ozawa, S., Takizawa, K., Takahashi, Y., et al. (2010) Characterization of photosystem I antenna proteins in the prasinophyte *Ostreococcus tauri*. *Biochimica Et Biophysica Acta Bba - Bioenergetics*. 1797: 1458–1464.

Takabayashi, A., Kadoya, R., Kuwano, M., Kurihara, K., Ito, H., Tanaka, R., et al. (2013) Protein co-migration database (PCoM -DB) for Arabidopsis thylakoids and Synechocystis cells. *Springerplus*. 2: 148.

Takabayashi, A., Kurihara, K., Kuwano, M., Kasahara, Y., Tanaka, R., and Tanaka, A. (2011) The oligomeric states of the photosystems and the light-harvesting complexes in the Chl *b*-less mutant. *Plant Cell Physiol*. 52: 2103–2114.

Takabayashi, A., Takabayashi, S., Takahashi, K., Watanabe, M., Uchida, H., Murakami, A., et al. (2017) PCoM-DB update: a protein co-migration database for photosynthetic organisms. *Plant Cell Physiol*. 58: e10–e10.

Takahashi, H., Iwai, M., Takahashi, Y., and Minagawa, J. (2006) Identification of the mobile light-harvesting complex II polypeptides for state transitions in *Chlamydomonas reinhardtii*. *Proc National Acad Sci*. 103: 477–482.

Tibiletti, T., Auroy, P., Peltier, G., and Caffarri, S. (2016) *Chlamydomonas reinhardtii* PsbS protein is functional and accumulates rapidly and transiently under high light. *Plant Physiol*. 171: 2717–2730.

Vaudel, M., Burkhart, J.M., Zahedi, R.P., Oveland, E., Berven, F.S., Sickmann, A., et al. (2015) PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol*. 33: 22–24.

Wang, S., Li, L., Li, H., Sahu, S.K., Wang, H., Xu, Y., et al. (2020) Genomes of early-diverging streptophyte algae shed light on plant terrestrialization. *Nat Plants*. 6: 95–106.

Wessels, H.J.C.T., Vogel, R.O., Heuvel, L. van den, Smeitink, J.A., Rodenburg, R.J., Nijtmans, L.G., et al. (2009) LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. *Proteomics*. 9: 4221–8.

Wobbe, L., Bassi, R., and Kruse, O. (2016) Multi-level light capture control in plants and green algae. *Trends Plant Sci*. 21: 55–68.

Yokono, M., Takabayashi, A., Kishimoto, J., Fujita, T., Iwai, M., Murakami, A., et al. (2019) The PSI–PSII megacomplex in green plants. *Plant Cell Physiol*. 60: 1098–1108.

Zhao, L., Zhang, H., Kohnen, M.V., Prasad, K.V.S.K., Gu, L., and Reddy, A.S.N. (2019) Analysis of transcriptome and epitranscriptome in plants using PacBio Iso-Seq and Nanopore-based direct RNA sequencing. *Frontiers Genetics*. 10: 253.

Zybailov, B., Mosley, A.L., Sardiu, M.E., Coleman, M.K., Florens, L., and Washburn, M.P. (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J Proteome Res*. 5: 2339–2347.

[Figure legends]

Figure 1. Separation of photosynthetic complexes by CN-PAGE

Separation of *M. viride* thylakoid membrane protein solubilized in 1% α-DDM by amphipol A8-35-based CN-PAGE (A). The number in parentheses represents the number of the corresponding gel slice which was further subjected to mass spectrometry analysis (see Fig. 1A). Immunoblot analysis of PsaB (PSI), and PsbB (PSII) proteins after their separation by 2D-CN/SDS-PAGE (B). Anti-PsaB and anti-PsbB antibodies were purchased from Agrisera. 2D- CN/SDS-PAGE of *M. viride* thylakoid membrane protein complexes visualized by silver-staining (C).

Figure 2. A phylogenetic tree of *M. viride* LHC proteins

A maximum likelihood phylogenetic tree of *M. viride* LHC proteins. Sequences highlighted in yellow are clades containing the LHC from *M. viride*. The *M. viride* LHC genes that are surrounded by a box represent the LHC genes identified in the Iso-seq analysis within the present study. LHCs that were classified differently between the previous report (Koziol et al., 2007) and our present study are marked with an asterisk.

Figure 3. Comparison of LHC migration profiles in PSI and PSII

M. viride thylakoid membranes were solubilized in α-DDM and separated by CN-PAGE. After separation, gel regions containing the resolved photosystems were horizontally cut into approximately 1 mm slices. The number shown in parentheses in Fig. 1A represents the number of the corresponding gel slice. (A) Migration profiles of PSI and PSII. The X-axis indicates the number of the gel slice as indicated on the CN-PAGE gel (Fig. 1A). The Y-axis indicates the relative protein abundance estimated using a label-free quantification method on the data obtained in the MS analysis. The PSII peak shown in gel slices 8 and 9 likely correspond to the PSII-LHCII band on the CN-PAGE gel, while the PSI peaks shown in gel slices 11 to 15 likely correspond to the PSI-LHCI-LHCII band and the PSI-LHCI band. (B) Migration profiles of LHCA1, LHCA2, LHCA3, algae-type LHCA2, and LHCA9 proteins. Their highest peaks correspond to the PSI-LHCI-LHCII and PSI-LHCI bands. The protein data of the two algae-type LHCA2 protein were combined to construct the migration profile of the algae-type LHCA2, as the two LHCA2 protein sequences were similar. (C) Migration profiles of LHCB4, LHCB5, and LHCBM proteins. The highest peaks in the LHCB4 and the LHCB5 migration profiles correspond with the PSII-LHCII band, while the highest peak in the LHCBM migration profile corresponds to the LHCII trimer. All of the MS data obtained for the LHCBM proteins were combined due their sequence similarities.

Figure 4. Separation of PSI-LHCI-LHCII supercomplexes using sucrose density gradient centrifugation followed by CN-PAGE.

(A) Separation of *M. viride* thylakoid membrane protein complexes solubilized in 1% α-DDM by sucrose density gradient centrifugation (0.3M-1.3M). The middle band (B2) contained the PSI-LHCI and the PSI-LHCI-LHCII, while the bottom band (B3) primarily contained PSII-LHCII and the upper band (B3) contained LHCII trimer. (B) Further separation of the PSI-enriched fraction (B2) in the sucrose gradient by CN-PAGE.

Figure 5. Separation of two PSI-LHCI-LHCII bands and a PSI-LHCI band by 2D-SDS-PAGE

Two PSI-LHCI-LHCII bands and a PSI-LHCI band were separated by CN-PAGE, followed by sucrose density gradient centrifugation (Fig. 4), and were then subjected to the 2D-SDS-PAGE. Protein bands were visualized by silver-staining.

Figure 6. A hypothetical model of evolutionary changes in the PSI peripheral antenna system of green plants.

During evolution, green plants have diverged into two groups: streptophytes and chlorophytes. Land plants are thought to have diverged from one group of freshwater streptophytes, while core chlorophytes, including *C. reinhardtii*, diverged from seawater chlorophytes (prasinophytes). The structure of PSI and its peripheral antennas have been previously reported in vascular plants (Qin et al. 2015; Mazor et al. 2017; Pan et al., 2018), moss plants (Iwai et al., 2018; Pinnola et al., 2018), and core chlorophytes, including *C. reinhardtii* (Su et al., 2019; Suga et al., 2019) and *D. salina* (Perez-Boerema et al. 2020). Notably, the “minimal” PSI-LHCI in *D. salina* lacks several core subunits in addition to the second LHCI belt. The PSI structure of streptophyte algae and prasinophyte algae, however, has not been reported. Based on the data obtained in the present study, *M. viride* PSI possesses LHCA1, two LHCA2s, LHCA3, algae-type LHCA2, and LHCA9. Since *M. viride* LHCA1, two LHCA2s, and LHCA3 exhibit considerable sequence similarity with LHCA proteins in vascular plants, their binding site may be similar to the binding site of the LHCIs in vascular plants. Based on sequence similarities, the binding site of algae-type LHCA2 and LHA9 may be similar to the binding site of PSI-LHCI in chlorophytes. The binding site of LHCP to PSII-LHCII is unclear. *M. viride* PSI can bind at least two LHCII trimers, however, the binding site of the additional LHCII trimer is unknown.

Fig. S1 A multiple sequence alignment of LHC sequences. Below is the multiple sequence alignment used to construct a phylogenetic tree in Fig. 2. A MAFFT program (https://mafft.cbrc.jp/alignment/software/) was used to align LHC sequences and a ClipKIT program was used to do trimming of the alignment.

Figure S2. 2D-SDS-PAGE of the high-molecular-weight PSI-LHCI-LHCII, PSI-LHCI, and LHCII trimer.

The high-molecular-weight PSI-LHCI-LHCII, PSI-LHCI, and LHCII trimer bands separated by CN-PAGE after sucrose density gradient centrifugation were subjected to 2D-SDS-PAGE. Protein bands were visualized by silver-staining.

Supplementary Table Legends:

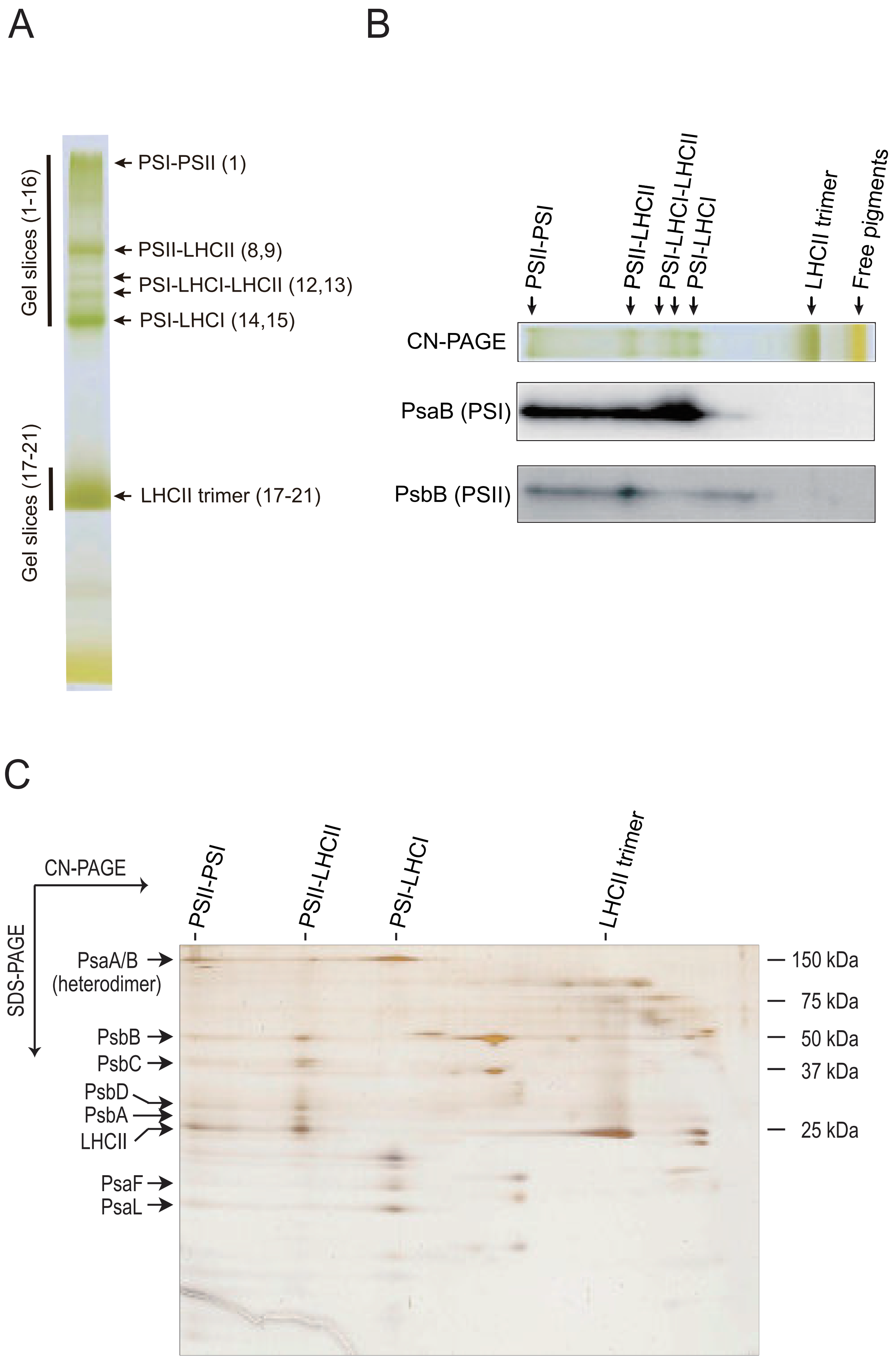
Table S1. List of *M. viride* LHC proteins predicted in the Iso-seq analysis.

Table S2. Proteins identified in the *M. viride* PSI-PSII band by MS.

Table S3. Proteins identified in the *M. viride* PSII-LHCII band by MS.

Table S4. Proteins identified in the *M. viride* PSI-LHCI band by MS.

Fig. 1



Aso et al.

Fig. 2

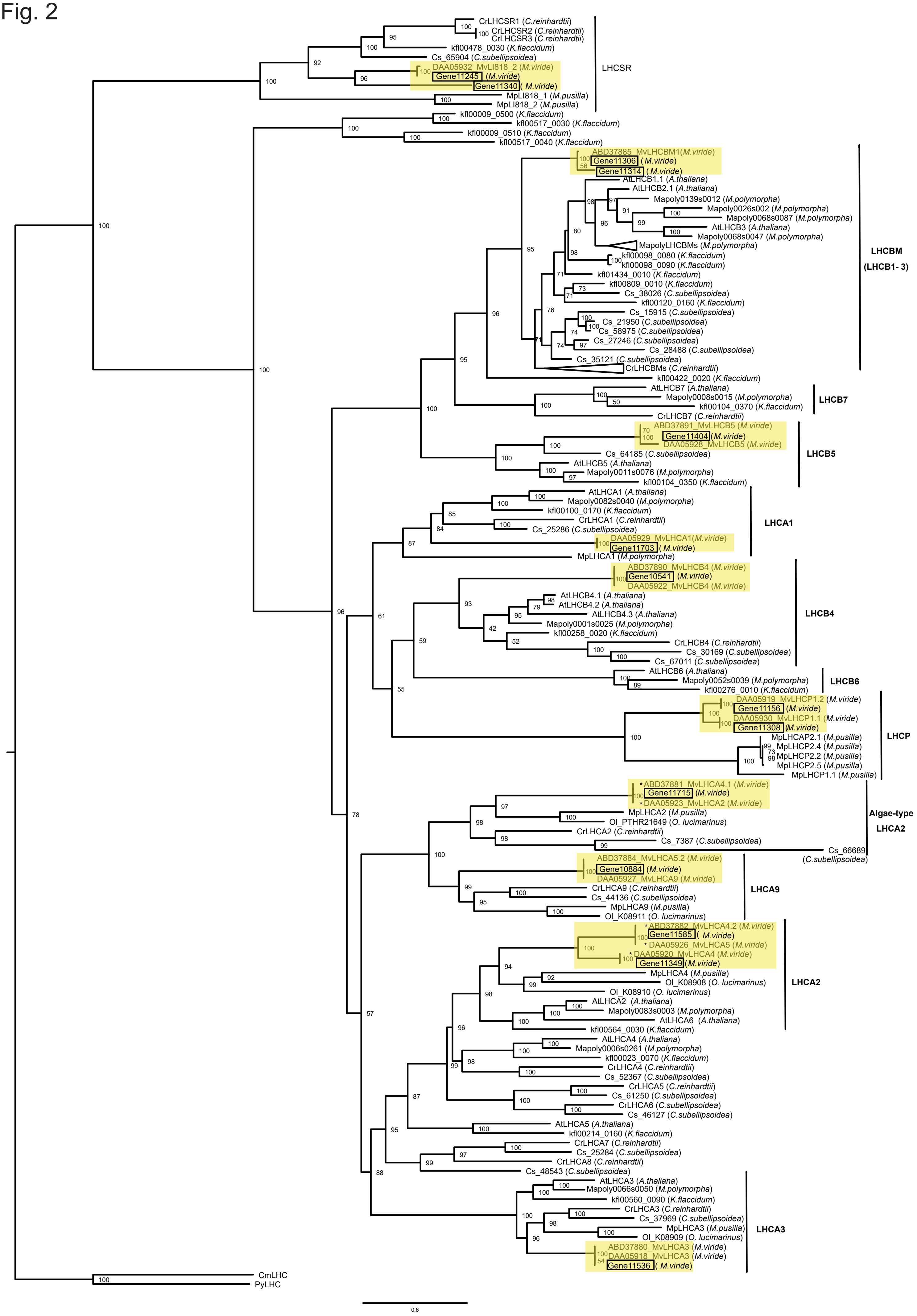

CrLHCSR1 (C.reinhardtii)
CrLHCSR2 (C.reinhardtii)
CrLHCSR3 (C.reinhardtii)
kfl00478_0030 (K.flaccidum)
Cs_65904 (C.subellipsoidea)
DAA05932_MvLI818_2 (M.viride)
Gene11245 (M.viride)
Gene11340 (M.viride)
MpLI818_1 (M.pusilla)
MpLI818_2 (M.pusilla)
LHCSR
kfl00009_0500 (K.flaccidum)
kfl00517_0030 (K.flaccidum)
kfl00009_0510 (K.flaccidum)
kfl00517_0040 (K.flaccidum)
ABD37885_MvLHCBM1(M.viride)
Gene11306 (M.viride)
Gene11314 (M.viride)
AtLHCB1.1 (A.thaliana)
AtLHCB2.1 (A.thaliana)
Mapoly0139s0012 (M.polymorpha)
Mapoly0026s002 (M.polymorpha)
Mapoly0068s0087 (M.polymorpha)
AtLHCB3 (A.thaliana)
Mapoly0068s0047 (M.polymorpha)
MapolyLHCBMs (M.polymorpha)
kfl00098_0080 (K.flaccidum)
kfl00098_0090 (K.flaccidum)
kfl01434_0010 (K.flaccidum)
kfl00809_0010 (K.flaccidum)
Cs_38026 (C.subellipsoidea)
kfl00120_0160 (K.flaccidum)
Cs_15915 (C.subellipsoidea)
Cs_21950 (C.subellipsoidea)
Cs_58975 (C.subellipsoidea)
Cs_27246 (C.subellipsoidea)
Cs_28488 (C.subellipsoidea)
Cs_35121 (C.subellipsoidea)
CrLHCBMs (C.reinhardtii)
kfl00422_0020 (K.flaccidum)
LHCBM
(LHCB1- 3)
AtLHCB7 (A.thaliana)
Mapoly0008s0015 (M.polymorpha)
kfl00104_0370 (K.flaccidum)
CrLHCB7 (C.reinhardtii)
LHCB7
ABD37891_MvLHCB5 (M.viride)
Gene11404 (M.viride)
DAA05928_MvLHCB5 (M.viride)
Cs_64185 (C.subellipsoidea)
AtLHCB5 (A.thaliana)
Mapoly0011s0076 (M.polymorpha)
kfl00104_0350 (K.flaccidum)
LHCB5
AtLHCA1 (A.thaliana)
Mapoly0082s0040 (M.polymorpha)
kfl00100_0170 (K.flaccidum)
CrLHCA1 (C.reinhardtii)
Cs_25286 (C.subellipsoidea)
DAA05929_MvLHCA1(M.viride)
Gene11703 (M.viride)
MpLHCA1 (M.polymorpha)
LHCA1
ABD37890_MvLHCB4 (M.viride)
Gene10541 (M.viride)
DAA05922_MvLHCB4 (M.viride)
AtLHCB4.1 (A.thaliana)
AtLHCB4.2 (A.thaliana)
AtLHCB4.3 (A.thaliana)
Mapoly0001s0025 (M.polymorpha)
kfl00258_0020 (K.flaccidum)
CrLHCB4 (C.reinhardtii)
Cs_30169 (C.subellipsoidea)
Cs_67011 (C.subellipsoidea)
LHCB4
AtLHCB6 (A.thaliana)
Mapoly0052s0039 (M.polymorpha)
kfl00276_0010 (K.flaccidum)
LHCB6
DAA05919_MvLHCP1.2 (M.viride)
Gene11156 (M.viride)
DAA05930_MvLHCP1.1 (M.viride)
Gene11308 (M.viride)
MpLHCAP2.1 (M.pusilla)
MpLHCP2.4 (M.pusilla)
MpLHCP2.2 (M.pusilla)
MpLHCP2.5 (M.pusilla)
MpLHCP1.1 (M.pusilla)
LHCP
*ABD37881_MvLHCA4.1 (M.viride)
Gene11715 (M.viride)
*DAA05923_MvLHCA2 (M.viride)
MpLHCA2 (M.pusilla)
Ol_PTHR21649 (O. lucimarinus)
CrLHCA2 (C.reinhardtii)
Cs_7387 (C.subellipsoidea)
Cs_66689 (C.subellipsoidea)
Algae-type
LHCA2
ABD37884_MvLHCA5.2 (M.viride)
Gene10884 (M.viride)
DAA05927_MvLHCA9 (M.viride)
CrLHCA9 (C.reinhardtii)
Cs_44136 (C.subellipsoidea)
MpLHCA9 (M.pusilla)
Ol_K08911 (O. lucimarinus)
LHCA9
*ABD37882_MvLHCA4.2 (M.viride)
Gene11585 (M.viride)
*DAA05926_MvLHCA5 (M.viride)
*DAA05920_MvLHCA4 (M.viride)
Gene11349 (M.viride)
MpLHCA4 (M.pusilla)
Ol_K08908 (O. lucimarinus)
Ol_K08910 (O. lucimarinus)
AtLHCA2 (A.thaliana)
Mapoly0083s0003 (M.polymorpha)
AtLHCA6 (A.thaliana)
kfl00564_0030 (K.flaccidum)
LHCA2
AtLHCA4 (A.thaliana)
Mapoly0006s0261 (M.polymorpha)
kfl00023_0070 (K.flaccidum)
CrLHCA4 (C.reinhardtii)
Cs_52367 (C.subellipsoidea)
CrLHCA5 (C.reinhardtii)
Cs_61250 (C.subellipsoidea)
CrLHCA6 (C.subellipsoidea)
Cs_46127 (C.subellipsoidea)
AtLHCA5 (A.thaliana)
kfl00214_0160 (K.flaccidum)
CrLHCA7 (C.reinhardtii)
Cs_25284 (C.subellipsoidea)
CrLHCA8 (C.reinhardtii)
Cs_48543 (C.subellipsoidea)
AtLHCA3 (A.thaliana)
Mapoly0066s0050 (M.polymorpha)
kfl00560_0090 (K.flaccidum)
CrLHCA3 (C.reinhardtii)
Cs_37969 (C.subellipsoidea)
MpLHCA3 (M.pusilla)
Ol_K08909 (O. lucimarinus)
ABD37880_MvLHCA3 (M.viride)
DAA05918_MvLHCA3 (M.viride)
Gene11536 (M.viride)
LHCA3
CmLHC
PyLHC
0.6

Fig. 3

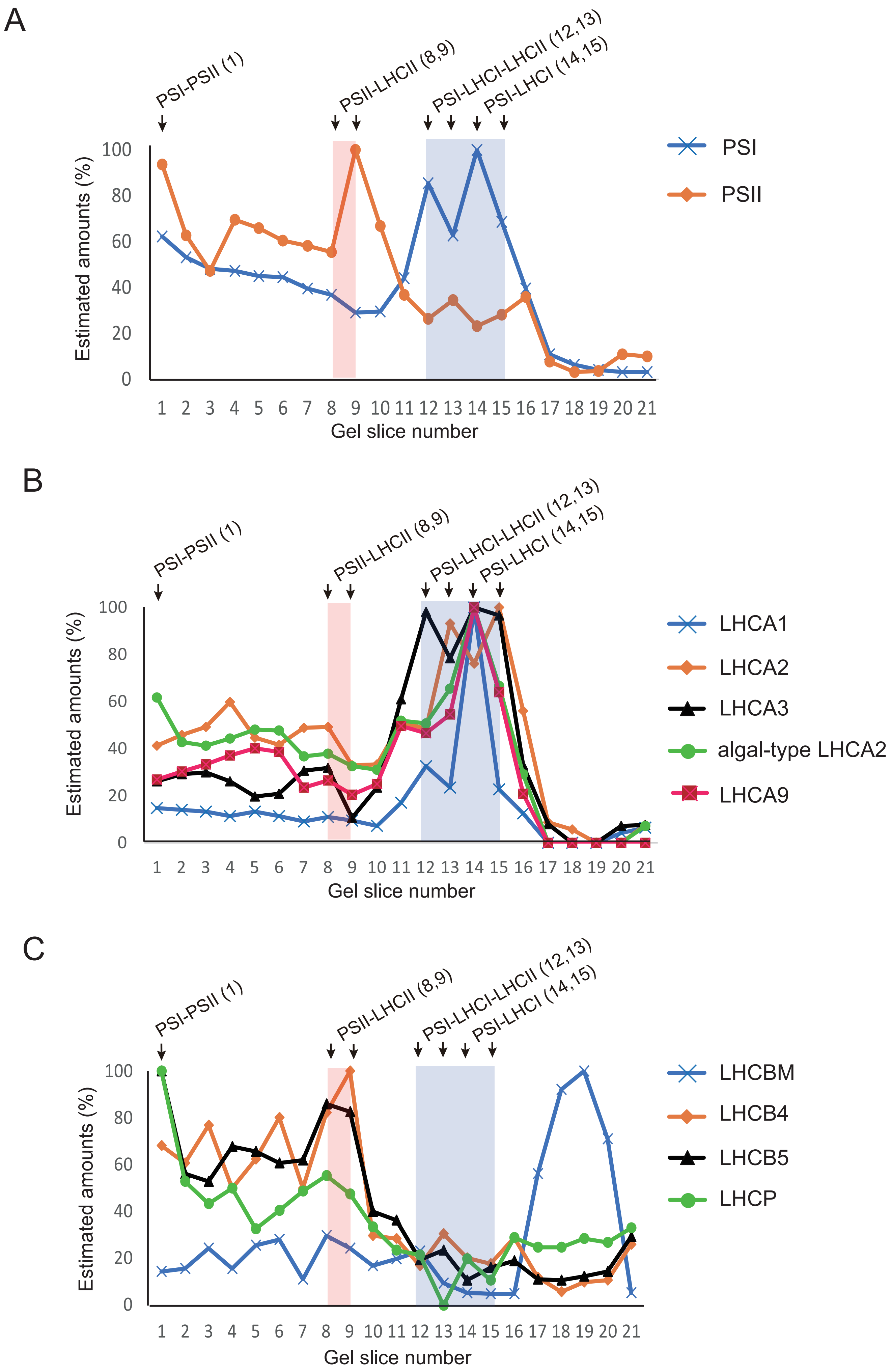

A
PSI-PSII (1)
PSII-LHCII (8,9)
PSI-LHCI-LHCII (12,13)
PSI-LHCI (14,15)
Estimated amounts (%)
0
20
40
60
80
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Gel slice number
PSI
PSII
B
PSI-PSII (1)
PSII-LHCII (8,9)
PSI-LHCI-LHCII (12,13)
PSI-LHCI (14,15)
Estimated amounts (%)
0
20
40
60
80
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Gel slice number
LHCA1
LHCA2
LHCA3
algal-type LHCA2
LHCA9
C
PSI-PSII (1)
PSII-LHCII (8,9)
PSI-LHCI-LHCII (12,13)
PSI-LHCI (14,15)
Estimated amounts (%)
0
20
40
60
80
100
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Gel slice number
LHCBM
LHCB4
LHCB5
LHCP

Fig. 4

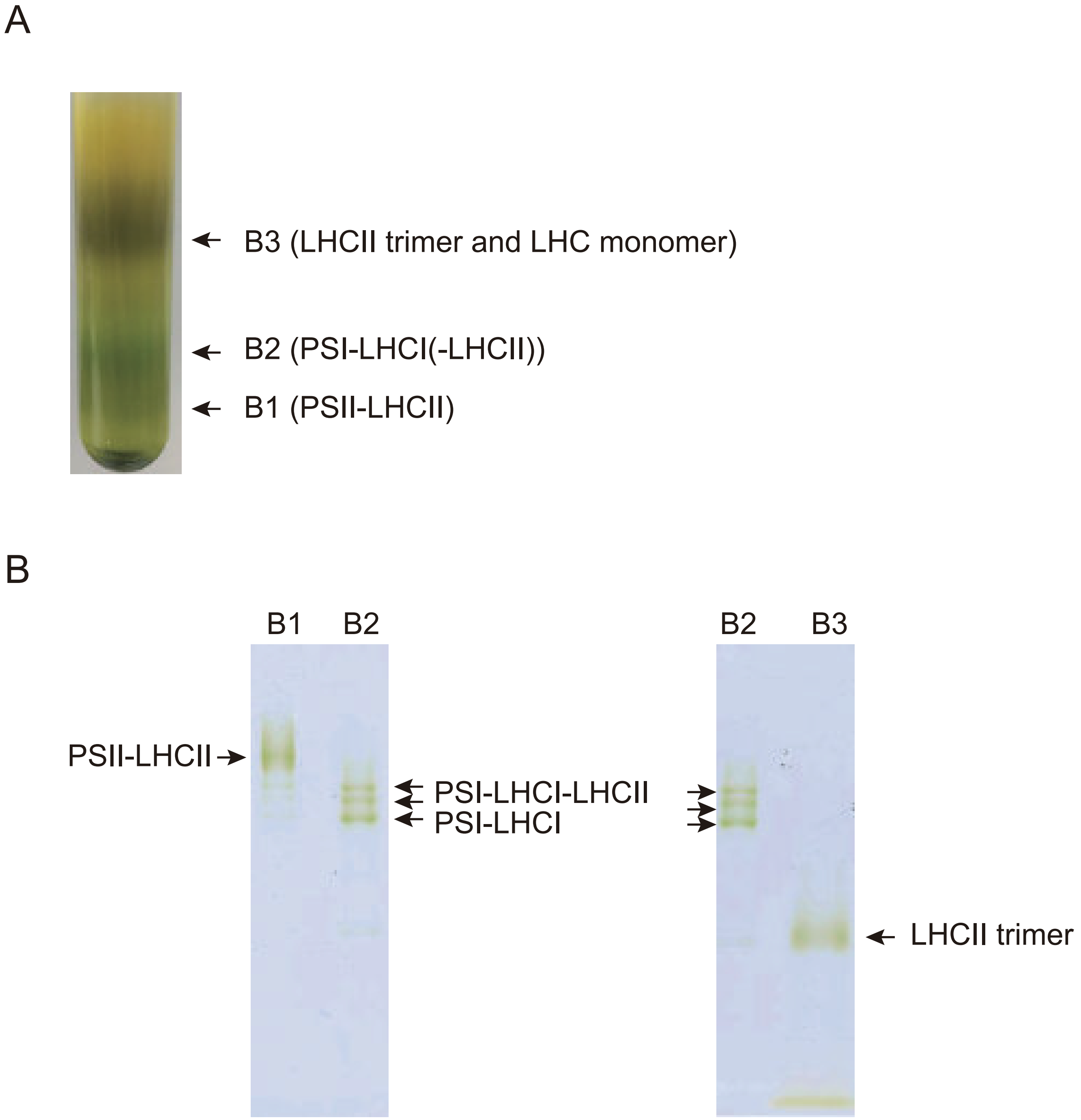
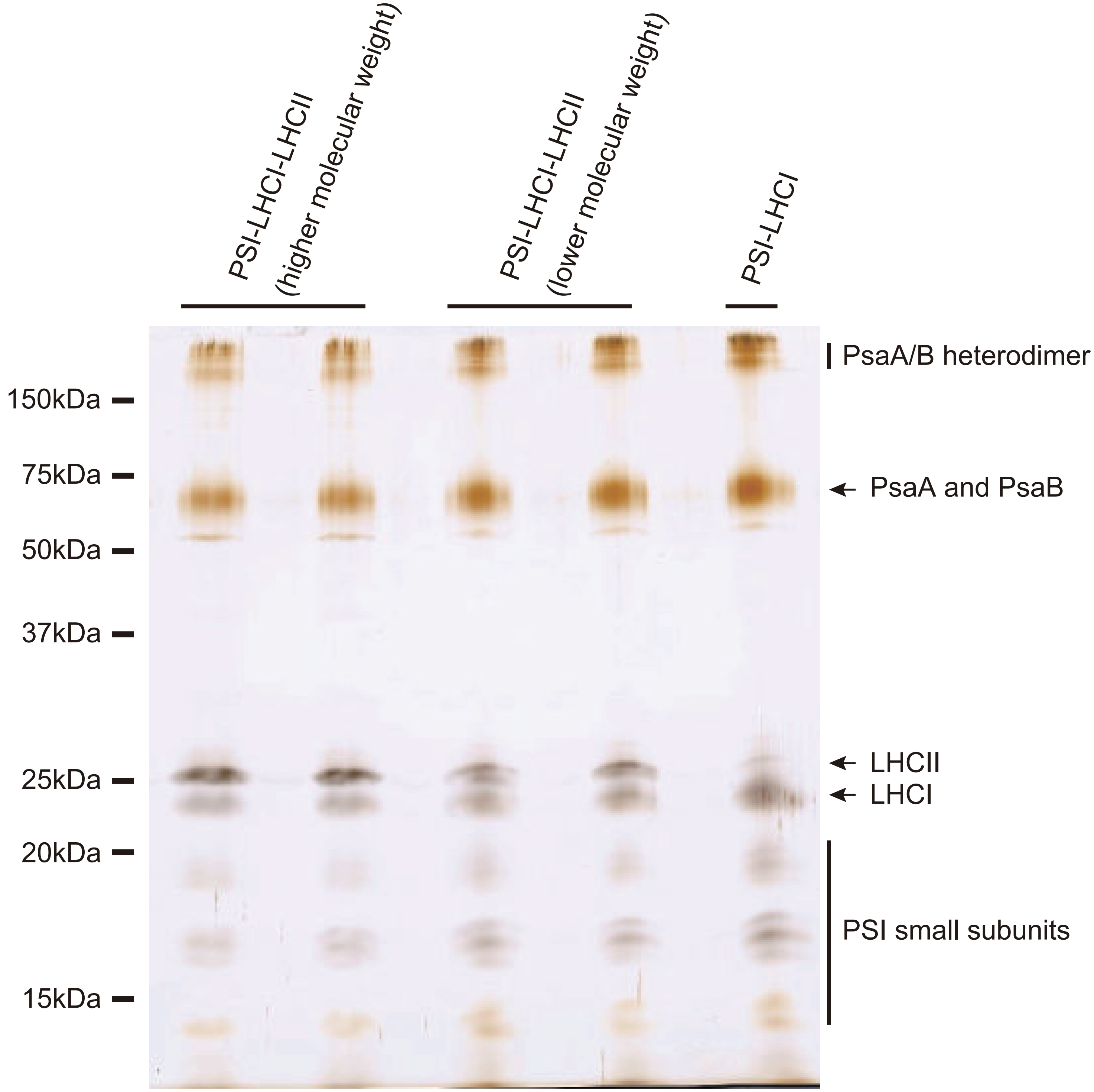
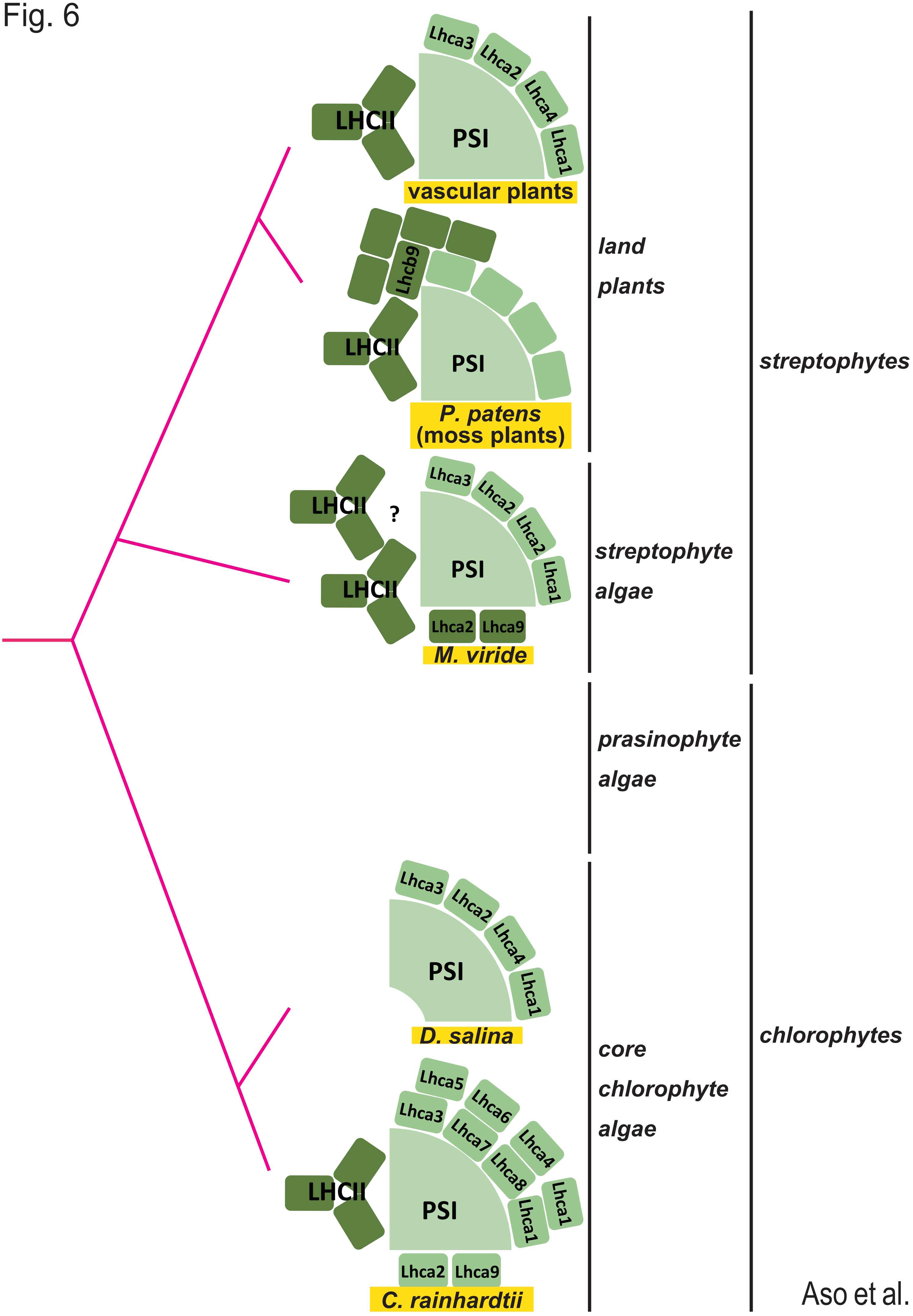

A
B3 (LHCII trimer and LHC monomer)
B2 (PSI-LHCI(-LHCII))
B1 (PSII-LHCII)
B
B1
B2
PSII-LHCII
PSI-LHCI-LHCII
PSI-LHCI
B2
B3
LHCII trimer

Fig. 5



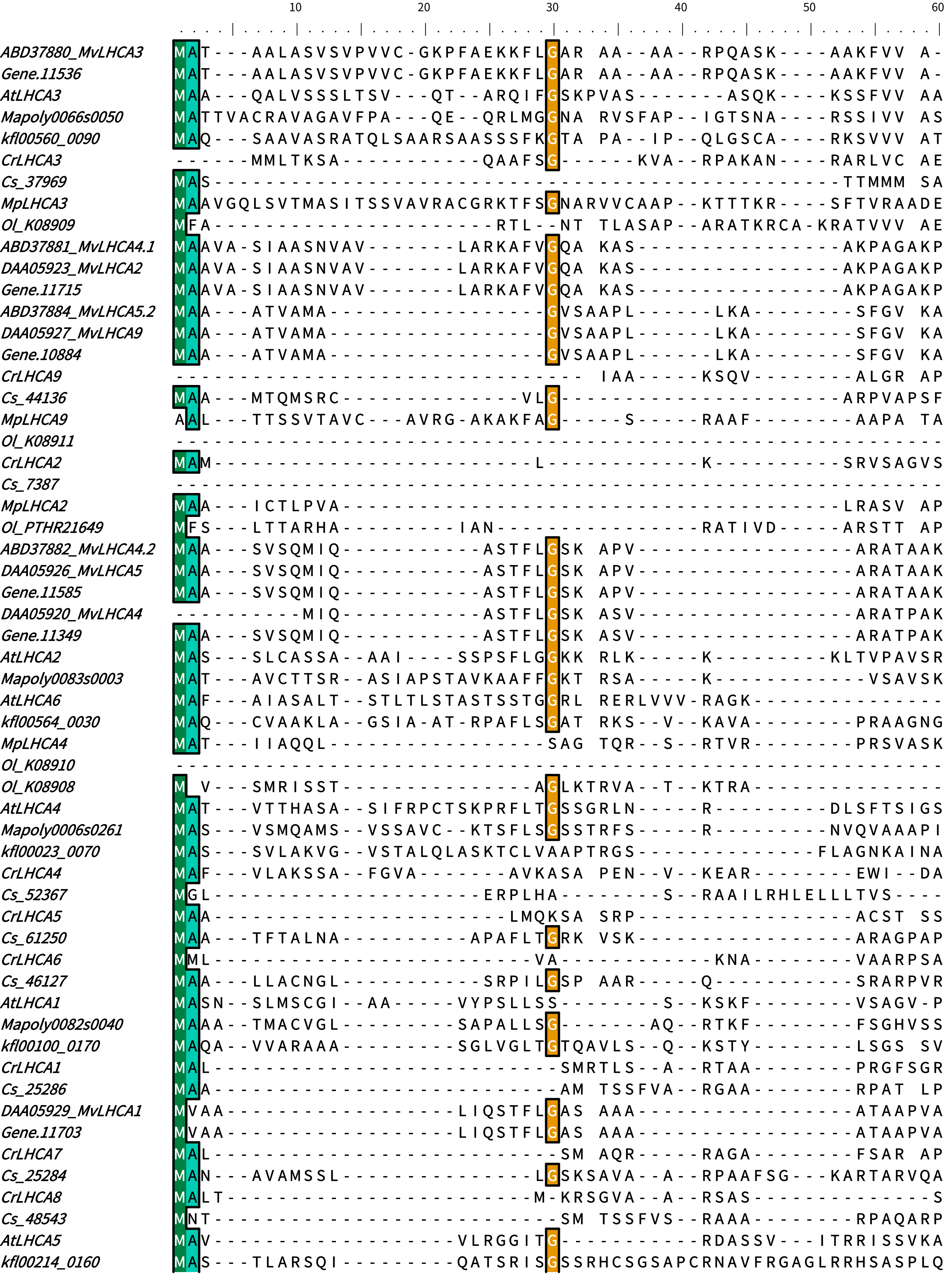
Aso et al.

Fig. 6

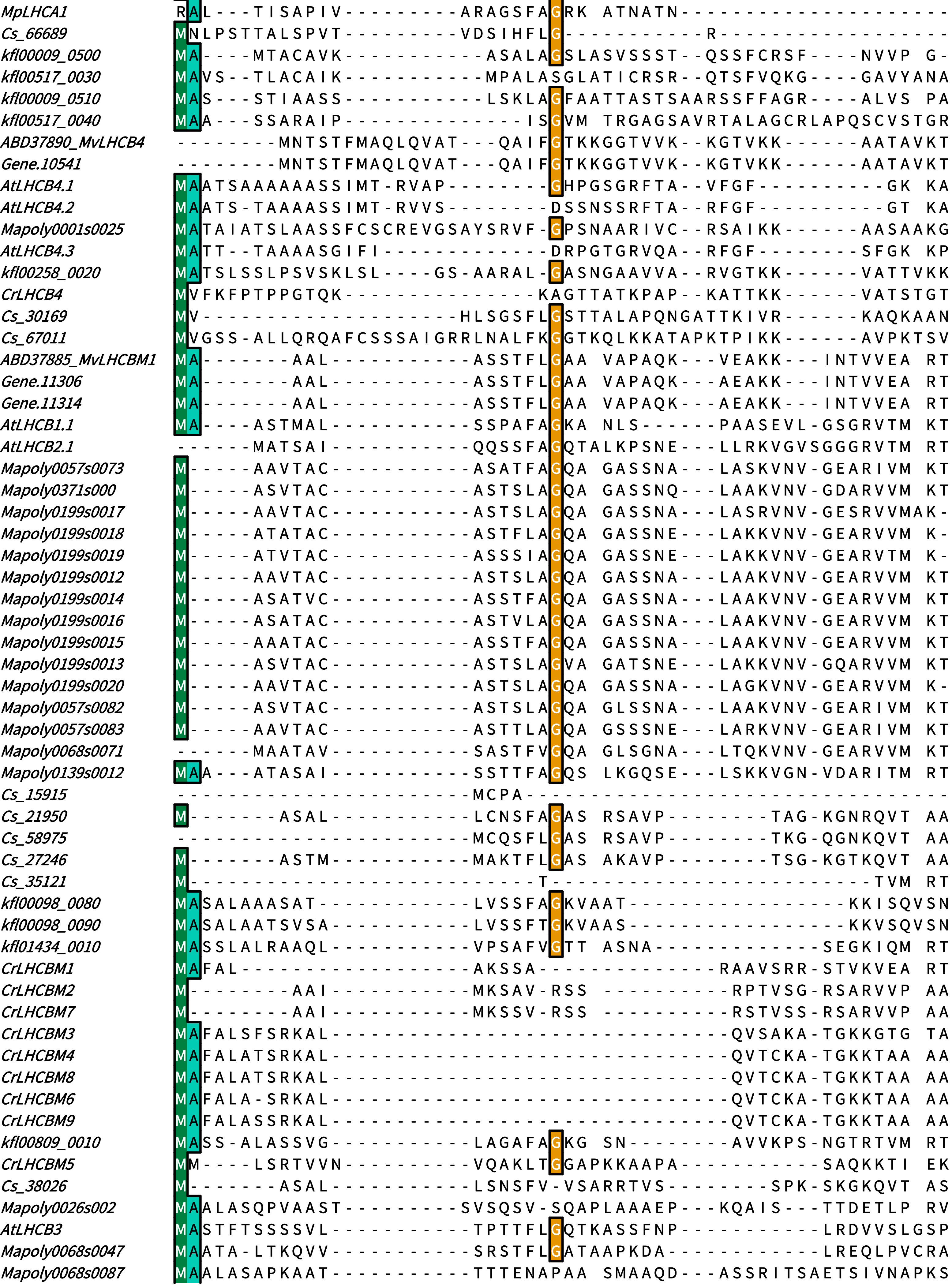


Aso et al.

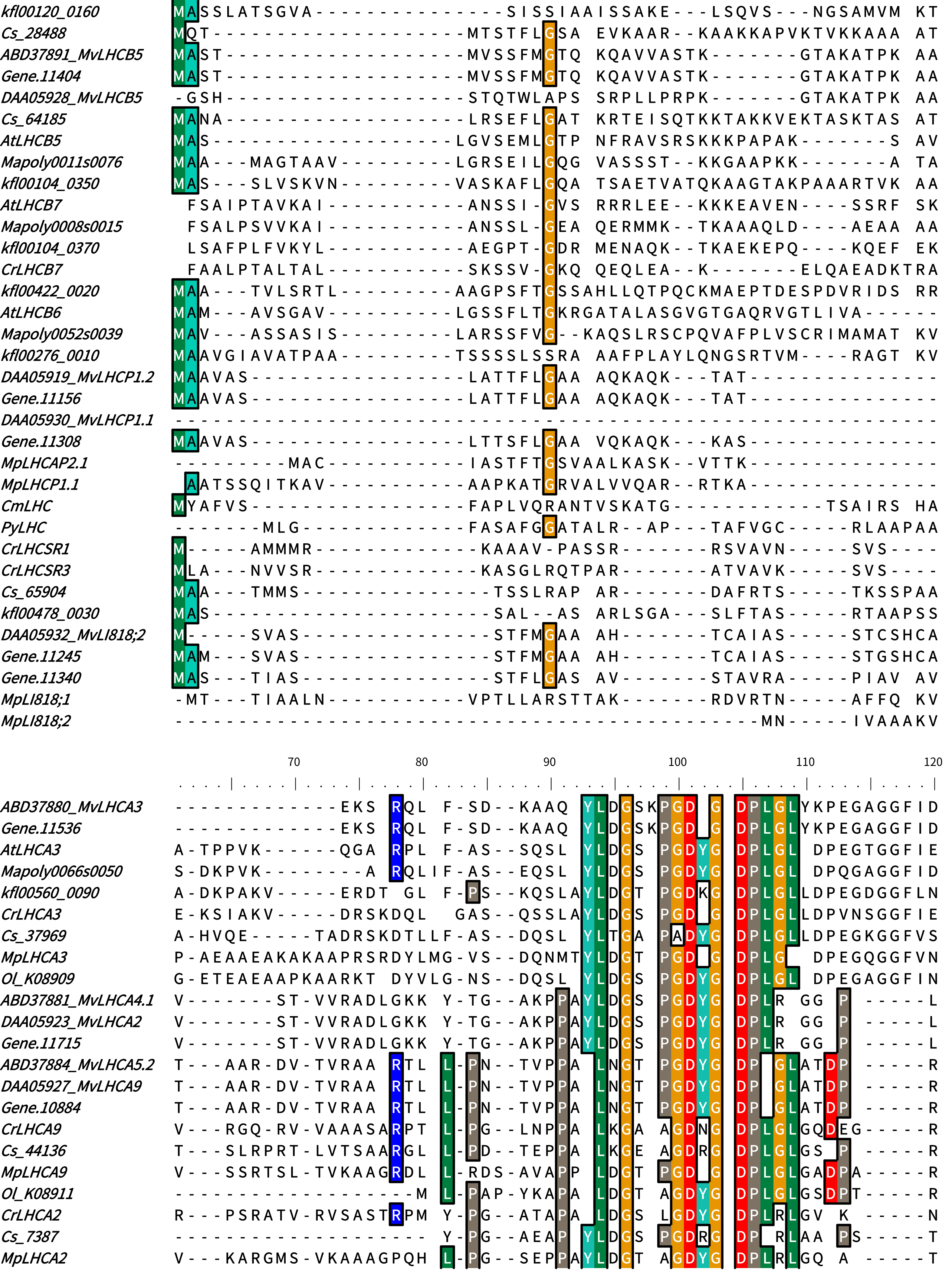
Fig. S1 A multiple sequence alignment of LHC sequences. Below is the multiple sequence alignment used to construct a phylogenetic tree in Fig. 2. A MAFFT program (https://mafft.cbrc.jp/alignment/software/) was used to align LHC sequences and a ClipKIT program was used to do trimming of the alignment.



```
                             10        20        30        40        50        60
ABD37880_MvLHCA3     MAT---AALASVSVPVVC-GKPFAEKKFLGAR AA--AA--RPQASK----AAKFVV A-
Gene.11536           MAT---AALASVSVPVVC-GKPFAEKKFLGAR AA--AA--RPQASK----AAKFVV A-
AtLHCA3              MAA---QALVSSSLTSV---QT--ARQIFGSKPVAS-------ASQK----KSSFVV AA
Mapoly0066s0050      MATTVACRAVAGAVFPA---QE--QRLMGGNA RVSFAP--IGTSNA----RSSIVV AS
kfl00560_0090        MAQ---SAAVASRATQLSAARSAASSSFKGTA PA--IP--QLGSCA----RKSVVV AT
CrLHCA3              ------MMLTKSA-----------QAAFSG------KVA--RPAKAN----RARLVC AE
Cs_37969             MAS-------------------------------------------------TTMMM SA
MpLHCA3              MAAVGQLSVTMASITSSVAVRACGRKTFSGNARVVCAAP--KTTTKR----SFTVRAADE
Ol_K08909            MFA----------------------RTL--NT TLASAP--ARATKRCA-KRATVVV AE
ABD37881_MvLHCA4.1   MAAVA-SIAASNVAV-------LARKAFVGQA KAS----------------AKPAGAKP
DAA05923_MvLHCA2     MAAVA-SIAASNVAV-------LARKAFVGQA KAS----------------AKPAGAKP
Gene.11715           MAAVA-SIAASNVAV-------LARKAFVGQA KAS----------------AKPAGAKP
ABD37884_MvLHCA5.2   MAA---ATVAMA-----------------GVSAAPL------LKA--------SFGV KA
DAA05927_MvLHCA9     MAA---ATVAMA-----------------GVSAAPL------LKA--------SFGV KA
Gene.10884           MAA---ATVAMA-----------------GVSAAPL------LKA--------SFGV KA
CrLHCA9              -------------------------------- IAA-----KSQV--------ALGR AP
Cs_44136             MAA---MTQMSRC--------------VLG----------------------ARPVAPSF
MpLHCA9              AAL---TTSSVTAVC---AVRG-AKAKFAG-----S-----RAAF--------AAPA TA
Ol_K08911            ------------------------------------------------------------
CrLHCA2              MAM-------------------------L------------K----------SRVSAGVS
Cs_7387              ------------------------------------------------------------
MpLHCA2              MAA---ICTLPVA---------------------------------------LRASV AP
Ol_PTHR21649         MFS---LTTARHA---------IAN----------------RATIVD-----ARSTT AP
ABD37882_MvLHCA4.2   MAA---SVSQMIQ-----------ASTFLGSK APV-----------------ARATAAK
DAA05926_MvLHCA5     MAA---SVSQMIQ-----------ASTFLGSK APV-----------------ARATAAK
Gene.11585           MAA---SVSQMIQ-----------ASTFLGSK APV-----------------ARATAAK
DAA05920_MvLHCA4     ----------MIQ-----------ASTFLGSK ASV-----------------ARATPAK
Gene.11349           MAA---SVSQMIQ-----------ASTFLGSK ASV-----------------ARATPAK
AtLHCA2              MAS---SLCASSA--AAI----SSPSFLGGKK RLK-----K---------KLTVPAVSR
Mapoly0083s0003      MAT---AVCTTSR--ASIAPSTAVKAAFFGKT RSA-----K------------VSAVSK
AtLHCA6              MAF---AIASALT--STLTLSTASTSSTGGRL RERLVVV-RAGK---------------
kfl00564_0030        MAQ---CVAAKLA--GSIA-AT-RPAFLSGAT RKS--V--KAVA--------PRAAGNG
MpLHCA4              MAT---IIAQQL-----------------SAG TQR--S--RTVR--------PRSVASK
Ol_K08910            ------------------------------------------------------------
Ol_K08908            M V---SMRISST---------------AGLKTRVA--T--KTRA---------------
AtLHCA4              MAT---VTTHASA--SIFRPCTSKPRFLTGSSGRLN-----R---------DLSFTSIGS
Mapoly0006s0261      MAS---VSMQAMS--VSSAVC--KTSFLSGSSTRFS-----R---------NVQVAAAPI
kfl00023_0070        MAS---SVLAKVG--VSTALQLASKTCLVAAPTRGS--------------FLAGNKAINA
CrLHCA4              MAF---VLAKSSA--FGVA-------AVKASA PEN--V--KEAR--------EWI--DA
Cs_52367             MGL---------------------ERPLHA--------S--RAAILRHLELLLTVS----
CrLHCA5              MAA-----------------------LMQKSA SRP-----------------ACST SS
Cs_61250             MAA---TFTALNA----------APAFLTGRK VSK-----------------ARAGPAP
CrLHCA6              MML-------------------------VA------------KNA--------VAARPSA
Cs_46127             MAA---LLACNGL-----------SRPILGSP AAR-----Q-----------SRARPVR
AtLHCA1              MASN--SLMSCGI--AA-----VYPSLLSS--------S--KSKF--------VSAGV-P
Mapoly0082s0040      MAAA--TMACVGL---------SAPALLSG-------AQ--RTKF--------FSGHVSS
kfl00100_0170        MAQA--VVARAAA---------SGLVGLTGTQAVLS--Q--KSTY--------LSGS SV
CrLHCA1              MAL---------------------------SMRTLS--A--RTAA--------PRGFSGR
Cs_25286             MAA---------------------------AM TSSFVA--RGAA--------RPAT LP
DAA05929_MvLHCA1     MVAA------------------LIQSTFLGAS AAA-----------------ATAAPVA
Gene.11703           MVAA------------------LIQSTFLGAS AAA-----------------ATAAPVA
CrLHCA7              MAL---------------------------SM AQR-----RAGA--------FSAR AP
Cs_25284             MAN---AVAMSSL---------------LGSKSAVA--A--RPAAFSG---KARTARVQA
CrLHCA8              MALT------------------------M-KRSGVA--A--RSAS--------------S
Cs_48543             MNT---------------------------SM TSSFVS--RAAA--------RPAQARP
AtLHCA5              MAV-------------------VLRGGITG-----------RDASSV---ITRRISSVKA
kfl00214_0160        MAS---TLARSQI---------QATSRISGSSRHCSGSAPCRNAVFRGAGLRRHSASPLQ
```

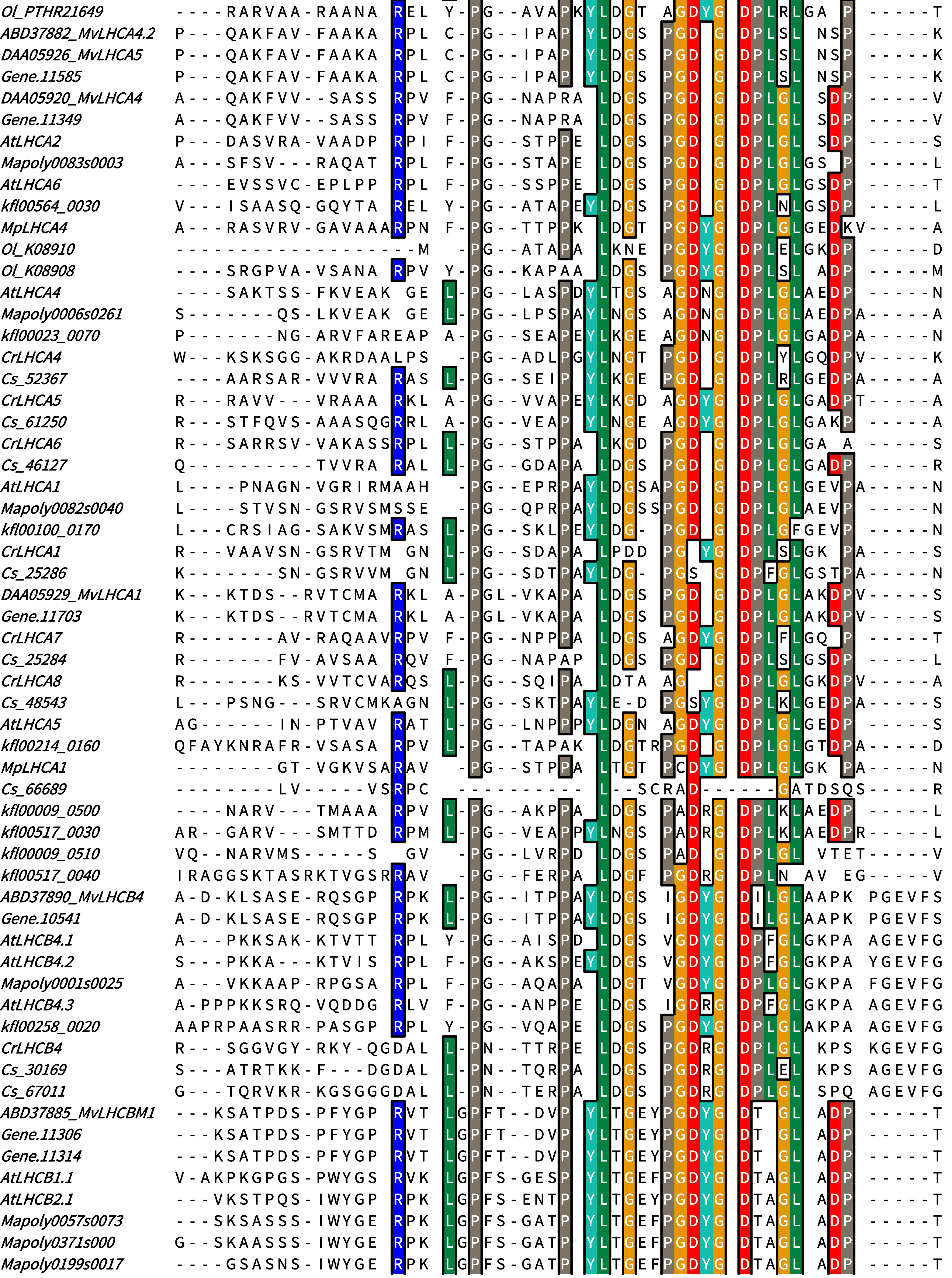


```
MpLHCA1            RAL---TISAPIV---------ARAGSFAGRK ATNATN---------------------
Cs_66689           MNLPSTTALSPVT---------VDSIHFLG-----------R------------------
kfl00009_0500      MA----MTACAVK-----------ASALAGSLASVSSST--QSSFCRSF----NVVP G-
kfl00517_0030      MAVS--TLACAIK-----------MPALASGLATICRSR--QTSFVQKG----GAVYANA
kfl00009_0510      MAS---STIAASS-----------LSKLAGFAATTASTSAARSSFFAGR----ALVS PA
kfl00517_0040      MAA---SSARAIP--------------ISGVM TRGAGSAVRTALAGCRLAPQSCVSTGR
ABD37890_MvLHCB4   --------MNTSTFMAQLQVAT---QAIFGTKKGGTVVK--KGTVKK------AATAVKT
Gene.10541         --------MNTSTFMAQLQVAT---QAIFGTKKGGTVVK--KGTVKK------AATAVKT
AtLHCB4.1          MAATSAAAAAASSIMT-RVAP--------GHPGSGRFTA--VFGF----------GK KA
AtLHCB4.2          MAATS-TAAAASSIMT-RVVS--------DSSNSSRFTA--RFGF----------GT KA
Mapoly0001s0025    MATAIATSLAASSFCSCREVGSAYSRVF-GPSNAARIVC--RSAIKK------AASAAKG
AtLHCB4.3          MATT--TAAAASGIFI-------------DRPGTGRVQA--RFGF--------SFGK KP
kfl00258_0020      MATSLSSLPSVSKLSL----GS-AARAL-GASNGAAVVA--RVGTKK------VATTVKK
CrLHCB4            MVFKFPTPPGTQK---------------KAGTTATKPAP--KATTKK------VATSTGT
Cs_30169           MV--------------------HLSGSFLGSTTALAPQNGATTKIVR------KAQKAAN
Cs_67011           MVGSS-ALLQRQAFCSSSAIGRRLNALFKGGTKQLKKATAPKTPIKK------AVPKTSV
ABD37885_MvLHCBM1  MA-------AAL-----------ASSTFLGAA VAPAQK---VEAKK---INTVVEA RT
Gene.11306         MA-------AAL-----------ASSTFLGAA VAPAQK---AEAKK---INTVVEA RT
Gene.11314         MA-------AAL-----------ASSTFLGAA VAPAQK---AEAKK---INTVVEA RT
AtLHCB1.1          MA----ASTMAL-----------SSPAFAGKA NLS------PAASEVL-GSGRVTM KT
AtLHCB2.1          ------MATSAI-----------QQSSFAGQTALKPSNE---LLRKVGVSGGGRVTM RT
Mapoly0057s0073    M-----AAVTAC-----------ASATFAGQA GASSNA---LASKVNV-GEARIVM KT
Mapoly0371s000     M-----ASVTAC-----------ASTSLAGQA GASSNQ---LAAKVNV-GDARVVM KT
Mapoly0199s0017    M-----AAVTAC-----------ASTSLAGQA GASSNA---LASRVNV-GESRVVMAK-
Mapoly0199s0018    M-----ATATAC-----------ASTFLAGQA GASSNE---LAKKVNV-GEARVVM K-
Mapoly0199s0019    M-----ATVTAC-----------ASSSIAGQA GASSNE---LAKKVNV-GEARVTM K-
Mapoly0199s0012    M-----AAVTAC-----------ASTSLAGQA GASSNA---LAAKVNV-GEARVVM KT
Mapoly0199s0014    M-----ASATVC-----------ASSTFAGQA GASSNA---LAAKVNV-GEARVVM KT
Mapoly0199s0016    M-----ASATAC-----------ASTVLAGQA GASSNA---LAAKVNV-GEARVVM KT
Mapoly0199s0015    M-----AAATAC-----------ASSTFAGQA GASSNA---LAAKVNV-GEARVVM KT
Mapoly0199s0013    M-----ASVTAC-----------ASTSLAGVA GATSNE---LAKKVNV-GQARVVM KT
Mapoly0199s0020    M-----AAVTAC-----------ASTSLAGQA GASSNA---LAGKVNV-GEARVVM K-
Mapoly0057s0082    M-----ASVTAC-----------ASTSLAGQA GLSSNA---LAAKVNV-GEARIVM KT
Mapoly0057s0083    M-----AAVTAC-----------ASTTLAGQA GSSSNE---LARKVNV-GEARIVM KT
Mapoly0068s0071    ------MAATAV-----------SASTFVGQA GLSGNA---LTQKVNV-GEARVVM KT
Mapoly0139s0012    MAA---ATASAI-----------SSTTFAGQS LKGQSE---LSKKVGN-VDARITM RT
Cs_15915           -----------------------MCPA---------------------------------
Cs_21950           M-------ASAL-----------LCNSFAGAS RSAVP--------TAG-KGNRQVT AA
Cs_58975           -----------------------MCQSFLGAS RSAVP--------TKG-QGNKQVT AA
Cs_27246           M-------ASTM-----------MAKTFLGAS AKAVP--------TSG-KGTKQVT AA
Cs_35121           M---------------------------T-------------------------TVM RT
kfl00098_0080      MASALAAASAT------------LVSSFAGKVAAT-----------------KKISQVSN
kfl00098_0090      MASALAATSVSA-----------LVSSFTGKVAAS-----------------KKVSQVSN
kfl01434_0010      MASSLALRAAQL-----------VPSAFVGTT ASNA-------------SEGKIQM RT
CrLHCBM1           MAFAL------------------AKSSA--------------RAAVSRR-STVKVEA RT
CrLHCBM2           M--------AAI-----------MKSAV-RSS ----------RPTVSG-RSARVVP AA
CrLHCBM7           M--------AAI-----------MKSSV-RSS ----------RSTVSS-RSARVVP AA
CrLHCBM3           MAFALSFSRKAL-------------------------------QVSAKA-TGKKGTG TA
CrLHCBM4           MAFALATSRKAL-------------------------------QVTCKA-TGKKTAA AA
CrLHCBM8           MAFALATSRKAL-------------------------------QVTCKA-TGKKTAA AA
CrLHCBM6           MAFALA-SRKAL-------------------------------QVTCKA-TGKKTAA AA
CrLHCBM9           MAFALASSRKAL-------------------------------QVTCKA-TGKKTAA AA
kfl00809_0010      MASS-ALASSVG-----------LAGAFAGKG SN--------AVVKPS-NGTRTVM RT
CrLHCBM5           MM----LSRTVVN----------VQAKLTGGAPKKAAPA-----------SAQKKTI EK
Cs_38026           M-------ASAL-----------LSNSFV-VSARRTVS--------SPK-SKGKQVT AS
Mapoly0026s002     MAALASQPVAAST---------SVSQSV-SQAPLAAAEP---KQAIS---TTDETLP RV
AtLHCB3            MASTFTSSSSVL-----------TPTTFLGQTKASSFNP-----------LRDVVSLGSP
Mapoly0068s0047    MAATA-LTKQVV-----------SRSTFLGATAAPKDA------------LREQLPVCRA
Mapoly0068s0087    MAALASAPKAAT-----------TTTENAPAA SMAAQD---ASSRITSAETSIVNAPKS
```

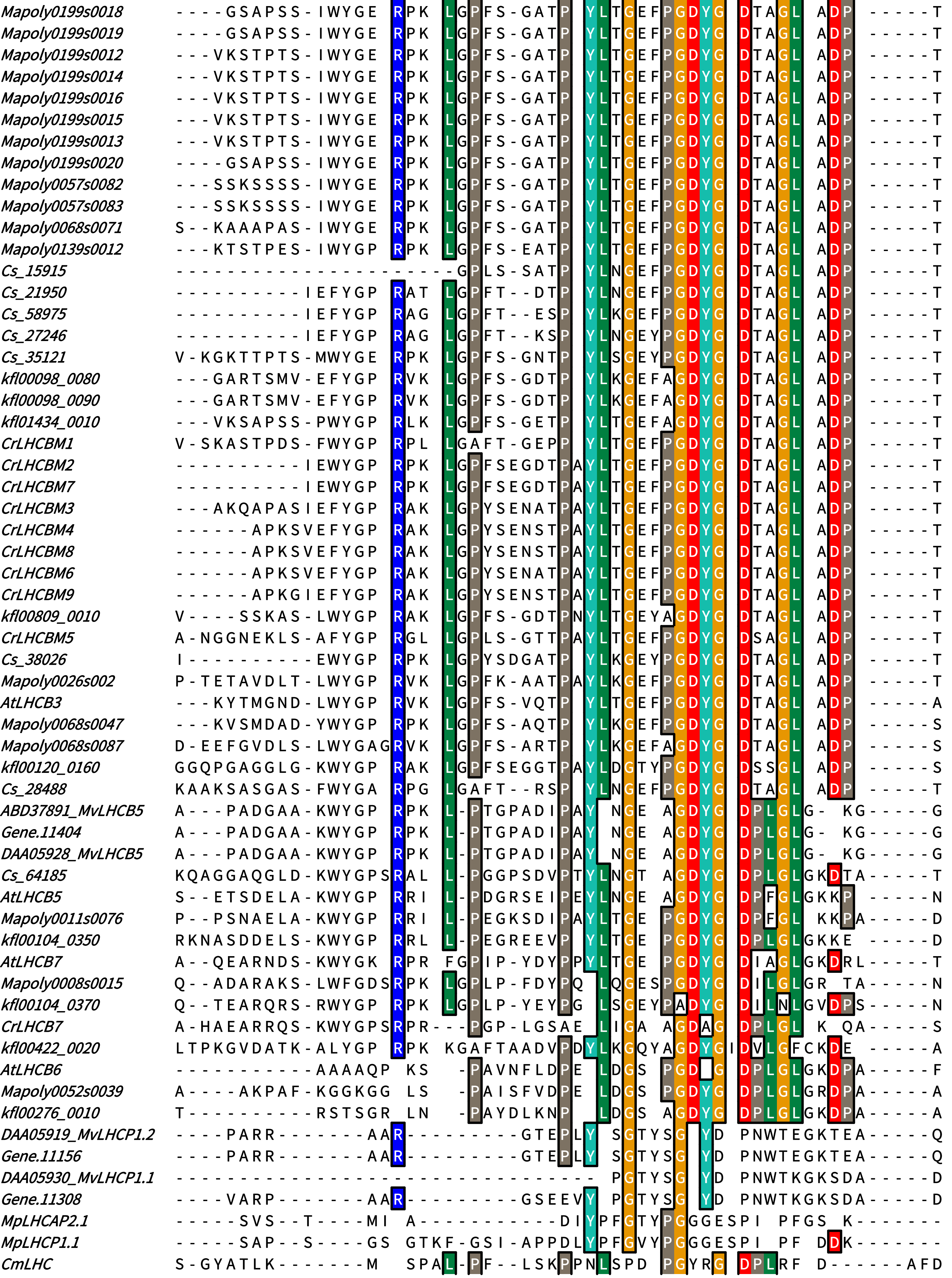


```
kfl00120_0160       MASSLATSGVA---------------SISSIAAISSAKE---LSQVS---NGSAMVM KT
Cs_28488            MQT--------------------MTSTFLGSA EVKAAR--KAAKKAPVKTVKKAAA AT
ABD37891_MvLHCB5    MAST-------------------MVSSFMGTQ KQAVVASTK-------GTAKATPK AA
Gene.11404          MAST-------------------MVSSFMGTQ KQAVVASTK-------GTAKATPK AA
DAA05928_MvLHCB5    -GSH-------------------STQTWLAPS SRPLLPRPK-------GTAKATPK AA
Cs_64185            MANA-------------------LRSEFLGAT KRTEISQTKKTAKKVEKTASKTAS AT
AtLHCB5             MAS-------------------LGVSEMLGTP NFRAVSRSKKKPAPAK-------S AV
Mapoly0011s0076     MAA---MAGTAAV---------LGRSEILGQG VASSST--KKGAAPKK-------A TA
kfl00104_0350       MAS---SLVSKVN---------VASKAFLGQA TSAETVATQKAAGTAKPAAARTVK AA
AtLHCB7              FSAIPTAVKAI-----------ANSSI-GVS RRRLEE--KKKEAVEN----SSRF SK
Mapoly0008s0015      FSALPSVVKAI-----------ANSSL-GEA QERMMK--TKAAAQLD----AEAA AA
kfl00104_0370        LSAFPLFVKYL-----------AEGPT-GDR MENAQK--TKAEKEPQ----KQEF EK
CrLHCB7              FAALPTALTAL-----------SKSSV-GKQ QEQLEA--K-------ELQAEADKTRA
kfl00422_0020       MAA---TVLSRTL---------AAGPSFTGSSAHLLQTPQCKMAEPTDESPDVRIDS RR
AtLHCB6             MAM---AVSGAV----------LGSSFLTGKRGATALASGVGTGAQRVGTLIVA------
Mapoly0052s0039     MAV---ASSASIS---------LARSSFVG--KAQSLRSCPQVAFPLVSCRIMAMAT KV
kfl00276_0010       MAAVGIAVATPAA---------TSSSSLSSRA AAFPLAYLQNGSRTVM----RAGT KV
DAA05919_MvLHCP1.2  MAAVAS-----------------LATTFLGAA AQKAQK---TAT---------------
Gene.11156          MAAVAS-----------------LATTFLGAA AQKAQK---TAT---------------
DAA05930_MvLHCP1.1  ------------------------------------------------------------
Gene.11308          MAAVAS-----------------LTTSFLGAA VQKAQK---KAS---------------
MpLHCAP2.1          ---------MAC-----------IASTFTGSVAALKASK--VTTK---------------
MpLHCP1.1            AATSSQITKAV-----------AAPKATGRVALVVQAR--RTKA---------------
CmLHC               MYAFVS-----------------FAPLVQRANTVSKATG------------TSAIRS HA
PyLHC               -------MLG-------------FASAFGGATALR--AP---TAFVGC-----RLAAPAA
CrLHCSR1            M-----AMMMR-------------KAAAV-PASSR-------RSVAVN-----SVS----
CrLHCSR3            MLA---NVVSR-------------KASGLRQTPAR-------ATVAVK-----SVS----
Cs_65904            MAA---TMMS---------------TSSLRAP AR-------DAFRTS-----TKSSPAA
kfl00478_0030       MAS----------------------SAL--AS ARLSGA---SLFTAS-----RTAAPSS
DAA05932_MvLI818;2  M-----SVAS---------------STFMGAA AH-------TCAIAS-----STCSHCA
Gene.11245          MAM---SVAS---------------STFMGAA AH-------TCAIAS-----STGSHCA
Gene.11340          MAS---TIAS---------------STFLGAS AV-------STAVRA-----PIAV AV
MpLI818;1           -MT---TIAALN-----------VPTLLARSTTAK-------RDVRTN-----AFFQ KV
MpLI818;2           ----------------------------------------------MN-----IVAAAKV
```

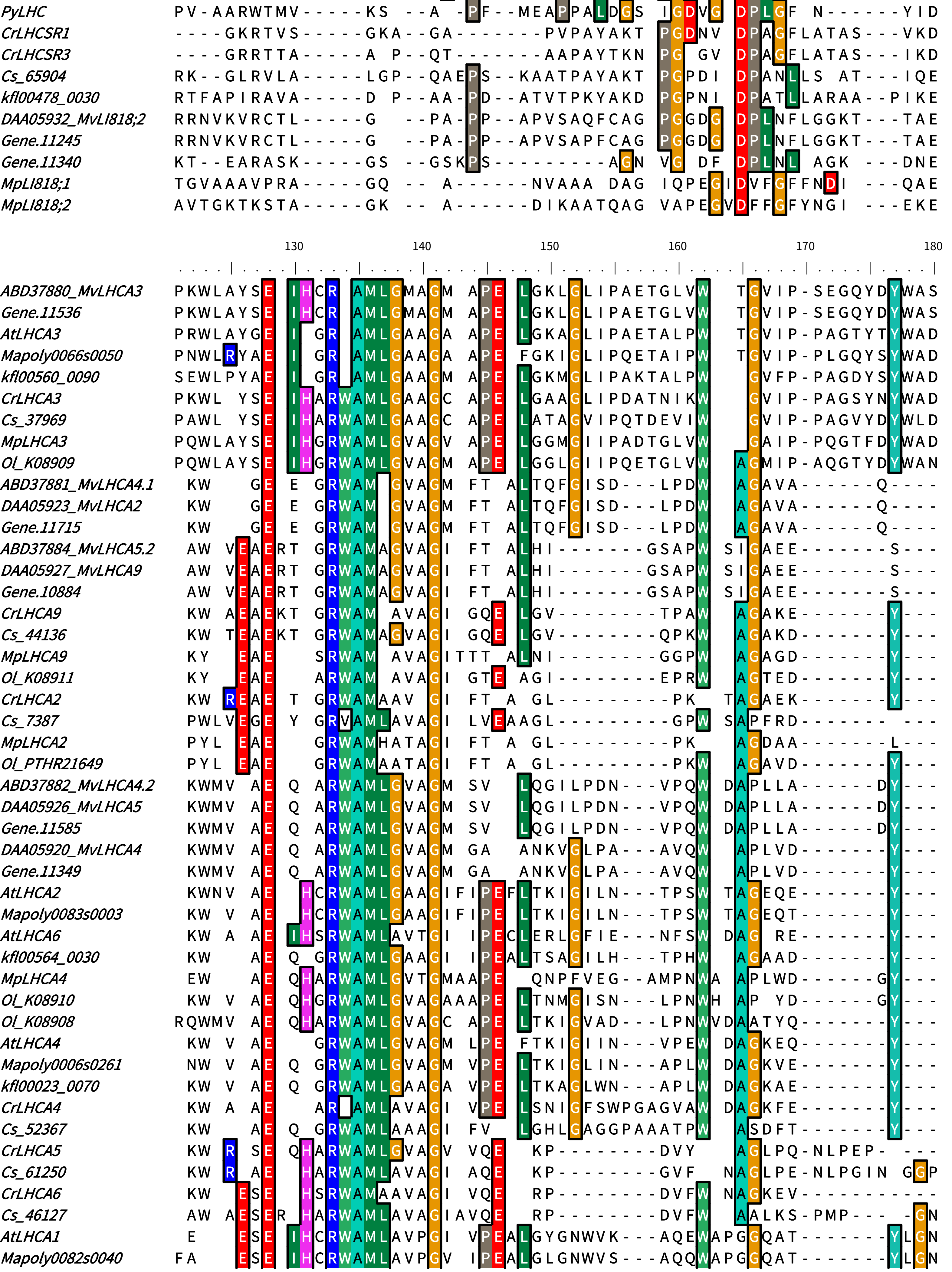
```
                            70        80        90       100       110       120
                    ....|....|....|....|....|....|....|....|....|....|....|....|
ABD37880_MvLHCA3    -------------EKS RQL F-SD--KAAQ YLDGSKPGD G DPLGLYKPEGAGGFID
Gene.11536          -------------EKS RQL F-SD--KAAQ YLDGSKPGD G DPLGLYKPEGAGGFID
AtLHCA3             A-TPPVK------QGA RPL F-AS--SQSL YLDGS PGDYG DPLGL DPEGTGGFIE
Mapoly0066s0050     S-DKPVK--------A RQLIF-AS--EQSL YLDGS PGDYG DPLGL DPQGAGGFID
kfl00560_0090       A-DKPAKV-----ERDT GL F-PS--KQSLAYLDGT PGDKG DPLGLLDPEGDGGFLN
CrLHCA3             E-KSIAKV-----DRSKDQL  GAS--QSSLAYLDGS PGD G DPLGLLDPVNSGGFIE
Cs_37969            A-HVQE-----TADRSKDTLLF-AS--DQSL YLTGA PADYG DPLGLLDPEGKGGFVS
MpLHCA3             P-AEAAEAKAKAAPRSRDYLMG-VS--DQNMTYLDGT PGD G DPLG  DPEGQGGFVN
Ol_K08909           G-ETEAEAAPKAARKT DYVLG-NS--DQSL YLDGS PGDYG DPLGL DPEGAGGFIN
ABD37881_MvLHCA4.1  V-------ST-VVRADLGKK Y-TG--AKPPAYLDGS PGDYG DPLR GG P -----L
DAA05923_MvLHCA2    V-------ST-VVRADLGKK Y-TG--AKPPAYLDGS PGDYG DPLR GG P -----L
Gene.11715          V-------ST-VVRADLGKK Y-TG--AKPPAYLDGS PGDYG DPLR GG P -----L
ABD37884_MvLHCA5.2  T---AAR-DV-TVRAA RTL L-PN--TVPPA LNGT PGDYG DP GLATDP -----R
DAA05927_MvLHCA9    T---AAR-DV-TVRAA RTL L-PN--TVPPA LNGT PGDYG DP GLATDP -----R
Gene.10884          T---AAR-DV-TVRAA RTL L-PN--TVPPA LNGT PGDYG DP GLATDP -----R
CrLHCA9             V---RGQ-RV-VAAASARPT L-PG--LNPPA LKGA AGDNG DPLGLGQDEG-----R
Cs_44136            T---SLRPRT-LVTSAARGL L-PD--TEPPA LKGE AGDRG DPLGLGS P -----R
MpLHCA9             V---SSRTSL-TVKAAGRDL L-RDS-AVAPP LDGT PGD G DPLGLGADPA-----R
Ol_K08911           -------------------M L-PAP-YKAPA LDGT AGDYG DPLGLGSDPT-----R
CrLHCA2             R---PSRATV-RVSASTRPM Y-PG--ATAPA LDGS LGDYG DPLRLGV K -----N
Cs_7387             -------------------- Y-PG--AEAP YLDGS PGDRG DP RLAA PS-----T
MpLHCA2             V---KARGMS-VKAAAGPQH L-PG--SEPPAYLDGT AGDYG DPLRLGQ A -----T
```



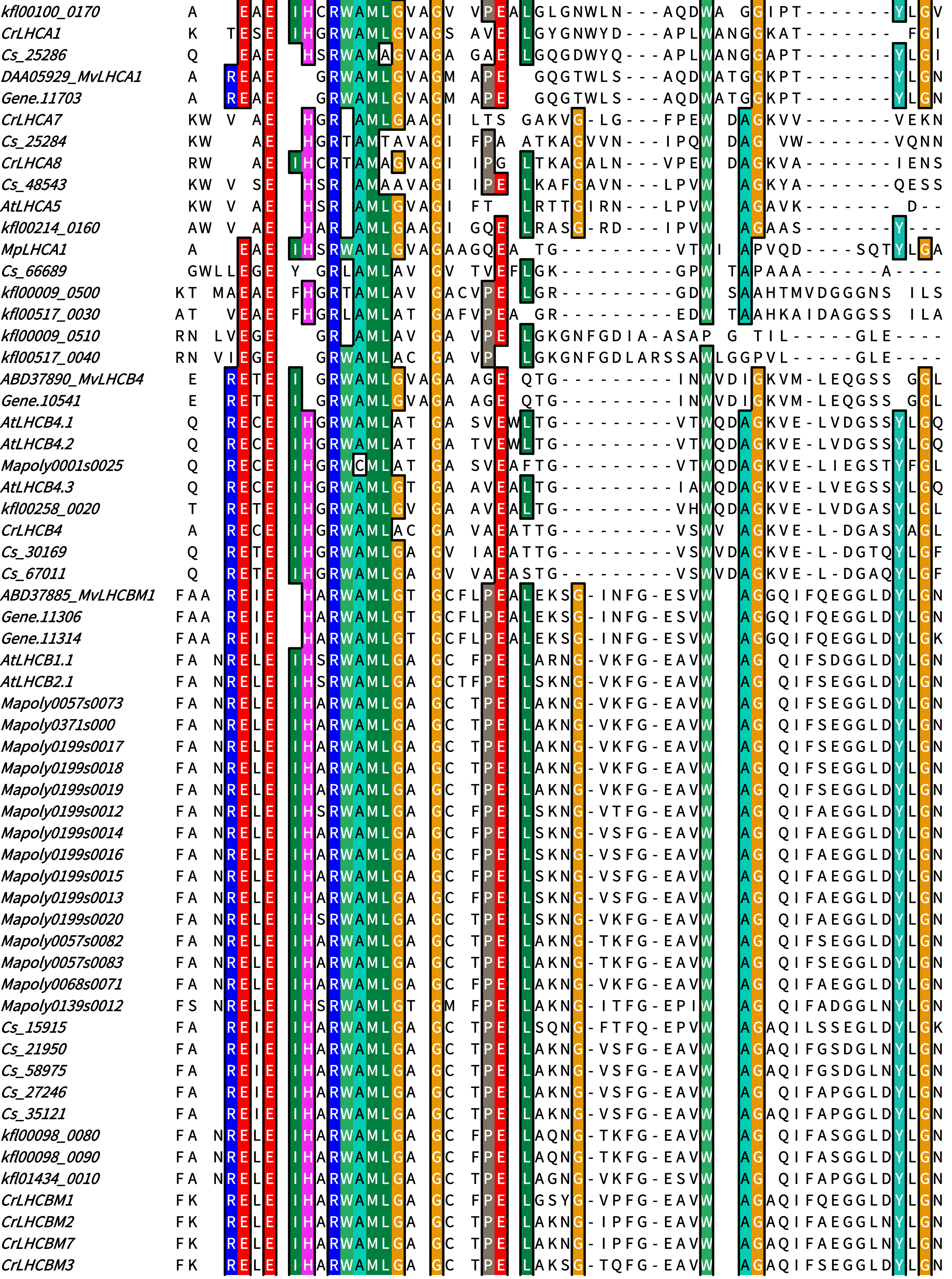
```
Ol_PTHR21649        ----RARVAA-RAANA REL Y-PG--AVAPKYLDGT AGDYG DPLRLGA P -----T
ABD37882_MvLHCA4.2  P---QAKFAV-FAAKA RPL C-PG--IPAP YLDGS PGD G DPLSL NSP -----K
DAA05926_MvLHCA5    P---QAKFAV-FAAKA RPL C-PG--IPAP YLDGS PGD G DPLSL NSP -----K
Gene.11585          P---QAKFAV-FAAKA RPL C-PG--IPAP YLDGS PGD G DPLSL NSP -----K
DAA05920_MvLHCA4    A---QAKFVV--SASS RPV F-PG--NAPRA LDGS PGD G DPLGL SDP -----V
Gene.11349          A---QAKFVV--SASS RPV F-PG--NAPRA LDGS PGD G DPLGL SDP -----V
AtLHCA2             P---DASVRA-VAADP RPI F-PG--STPPE LDGS PGD G DPLGL SDP -----S
Mapoly0083s0003     A---SFSV---RAQAT RPL F-PG--STAPE LDGS PGD G DPLGLGS P -----L
AtLHCA6             ----EVSSVC-EPLPP RPL F-PG--SSPPE LDGS PGD G DPLGLGSDP -----T
kfl00564_0030       V---ISAASQ-GQYTA REL Y-PG--ATAPEYLDGS PGD G DPLNLGSDP -----L
MpLHCA4             A---RASVRV-GAVAAARPN F-PG--TTPPK LDGT PGDYG DPLGLGEDKV-----A
Ol_K08910           -------------------M  -PG--ATAPA LKNE PGDYG DPLELGKDP -----D
Ol_K08908           ----SRGPVA-VSANA RPV Y-PG--KAPAA LDGS PGDYG DPLSL ADP -----M
AtLHCA4             ----SAKTSS-FKVEAK GE L-PG--LASPDYLTGS AGDNG DPLGLAEDP -----N
Mapoly0006s0261     S-------QS-LKVEAK GE L-PG--LPSPAYLNGS AGDNG DPLGLAEDPA-----A
kfl00023_0070       P-------NG-ARVFAREAP A-PG--SEAPEYLKGE AGDNG DPLGLGADPA-----N
CrLHCA4             W---KSKSGG-AKRDAALPS  -PG--ADLPGYLNGT PGD G DPLYLGQDPV-----K
Cs_52367            ----AARSAR-VVVRA RAS L-PG--SEIP YLKGE PGD G DPLRLGEDPA-----A
CrLHCA5             R---RAVV---VRAAA RKL A-PG--VVAPEYLKGD AGDYG DPLGLGADPT-----A
Cs_61250            R---STFQVS-AAASQGRRL A-PG--VEAP YLNGE AGDYG DPLGLGAKP -----A
CrLHCA6             R---SARRSV-VAKASSRPL L-PG--STPPA LKGD PGD G DPLGLGA A -----S
Cs_46127            Q----------TVVRA RAL L-PG--GDAPA LDGS PGD G DPLGLGADP -----R
AtLHCA1             L----PNAGN-VGRIRMAAH  -PG--EPRPAYLDGSAPGD G DPLGLGEVPA-----N
Mapoly0082s0040     L----STVSN-GSRVSMSSE  -PG--QPRPAYLDGSSPGD G DPLGLAEVP -----N
kfl00100_0170       L---CRSIAG-SAKVSMRAS L-PG--SKLPEYLDG- PGD G DPLGFGEVP -----N
CrLHCA1             R---VAAVSN-GSRVTM GN L-PG--SDAPA LPDD PG YG DPLSLGK PA-----S
Cs_25286            K-------SN-GSRVVM GN L-PG--SDTPAYLDG- PGS G DPFGLGSTPA-----N
DAA05929_MvLHCA1    K---KTDS--RVTCMA RKL A-PGL-VKAPA LDGS PGD G DPLGLAKDPV-----S
Gene.11703          K---KTDS--RVTCMA RKL A-PGL-VKAPA LDGS PGD G DPLGLAKDPV-----S
CrLHCA7             R-------AV-RAQAAVRPV F-PG--NPPPA LDGS AGDYG DPLFLGQ P -----T
Cs_25284            R-------FV-AVSAA RQV F-PG--NAPAP LDGS PGD G DPLSLGSDP -----L
CrLHCA8             R-------KS-VVTCVARQS L-PG--SQIPA LDTA AG  G DPLGLGKDPV-----A
Cs_48543            L---PSNG---SRVCMKAGN L-PG--SKTPAYLE-D PGSYG DPLKLGEDPA-----S
AtLHCA5             AG------IN-PTVAV RAT L-PG--LNPPPYLDGN AGDYG DPLGLGEDP -----S
kfl00214_0160       QFAYKNRAFR-VSASA RPV L-PG--TAPAK LDGTRPGD G DPLGLGTDPA-----D
MpLHCA1             --------GT-VGKVSARAV  -PG--STPPA LTGT PCDYG DPLGLGK PA-----N
Cs_66689            --------LV-----VSRPC  ---------- L--SCRAD------GATDSQS-----R
kfl00009_0500       ----NARV---TMAAA RPV L-PG--AKPPA LDGS PADRG DPLKLAEDP -----L
kfl00517_0030       AR--GARV---SMTTD RPM L-PG--VEAPPYLNGS PADRG DPLKLAEDPR-----L
kfl00009_0510       VQ--NARVMS-----S  GV  -PG--LVRPD LDGS PAD G DPLGL VTET-----V
kfl00517_0040       IRAGGSKTASRKTVGSRRAV  -PG--FERPA LDGF PGDRG DPLN AV EG-----V
ABD37890_MvLHCB4    A-D-KLSASE-RQSGP RPK L-PG--ITPPAYLDGS IGDYG DILGLAAPK PGEVFS
Gene.10541          A-D-KLSASE-RQSGP RPK L-PG--ITPPAYLDGS IGDYG DILGLAAPK PGEVFS
AtLHCB4.1           A---PKKSAK-KTVTT RPL Y-PG--AISPD LDGS VGDYG DPFGLGKPA AGEVFG
AtLHCB4.2           S---PKKA---KTVIS RPL F-PG--AKSPEYLDGS VGDYG DPFGLGKPA YGEVFG
Mapoly0001s0025     A---VKKAAP-RPGSA RPL F-PG--AQAPA LDGT VGDYG DPLGLGKPA FGEVFG
AtLHCB4.3           A-PPPKKSRQ-VQDDG RLV F-PG--ANPPE LDGS IGDRG DPFGLGKPA AGEVFG
kfl00258_0020       AAPRPAASRR-PASGP RPL Y-PG--VQAPE LDGS PGDYG DPLGLAKPA AGEVFG
CrLHCB4             R---SGGVGY-RKY-QGDAL L-PN--TTRPE LDGS PGDRG DPLGL KPS KGEVFG
Cs_30169            S---ATRTKK-F---DGDAL L-PN--TQRPA LDGS PGDRG DPLEL KPS AGEVFG
Cs_67011            G---TQRVKR-KGSGGGDAL L-PN--TERPA LDGS PGDRG DPLGL SPQ AGEVFG
ABD37885_MvLHCBM1   ---KSATPDS-PFYGP RVT LGPFT--DVP YLTGEYPGDYG DT GL ADP -----T
Gene.11306          ---KSATPDS-PFYGP RVT LGPFT--DVP YLTGEYPGDYG DT GL ADP -----T
Gene.11314          ---KSATPDS-PFYGP RVT LGPFT--DVP YLTGEYPGDYG DT GL ADP -----T
AtLHCB1.1           V-AKPKGPGS-PWYGS RVK LGPFS-GESP YLTGEFPGDYG DTAGL ADP -----T
AtLHCB2.1           ---VKSTPQS-IWYGP RPK LGPFS-ENTP YLTGEYPGDYG DTAGL ADP -----T
Mapoly0057s0073     ---SKSASSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0371s000      G--SKAASSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0017     ----GSASNS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
```



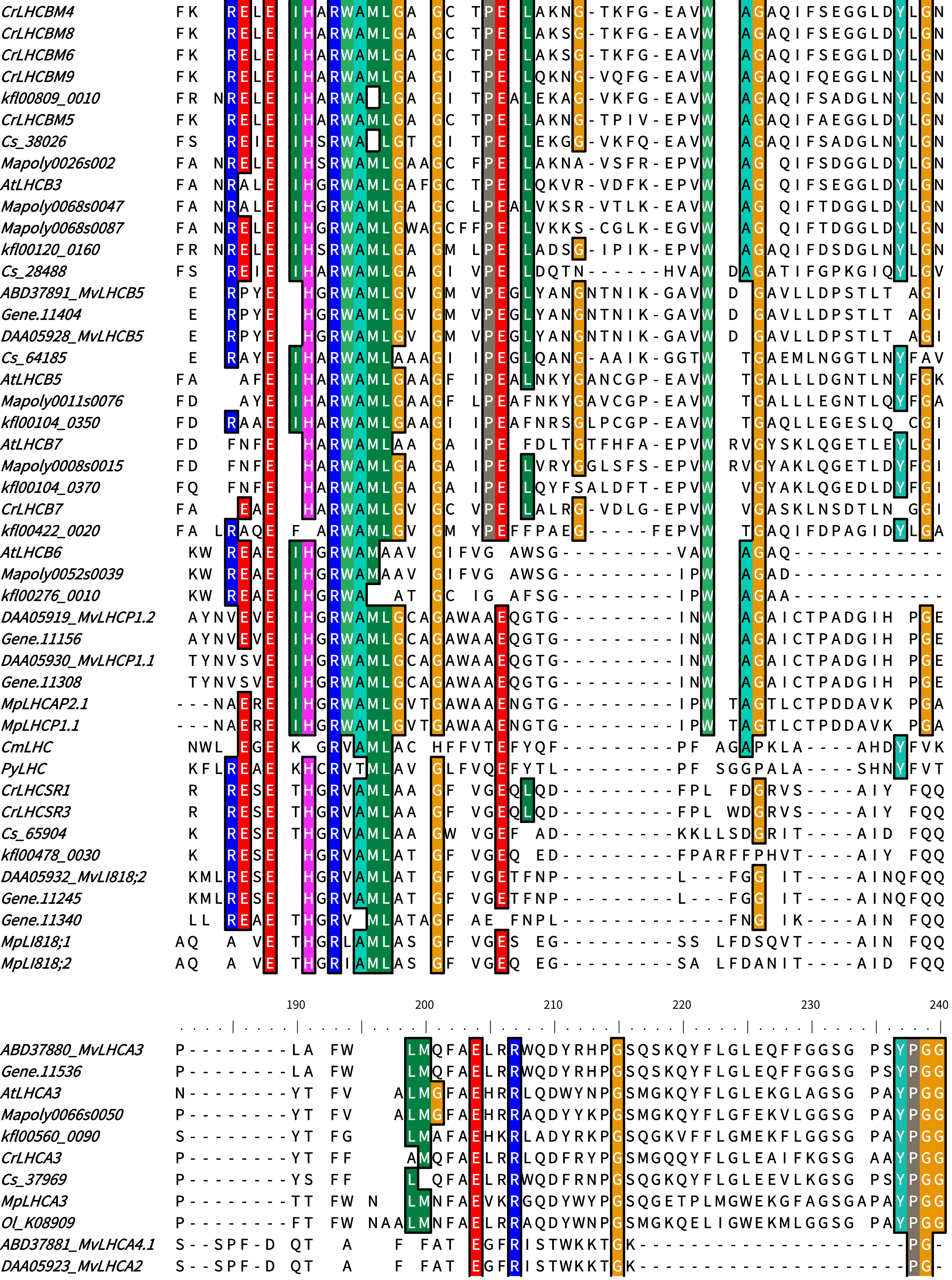
```
Mapoly0199s0018      ----GSAPSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0019      ----GSAPSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0012      ---VKSTPTS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0014      ---VKSTPTS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0016      ---VKSTPTS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0015      ---VKSTPTS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0013      ---VKSTPTS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0199s0020      ----GSAPSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0057s0082      ---SSKSSSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0057s0083      ---SSKSSSS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0068s0071      S--KAAAPAS-IWYGE RPK LGPFS-GATP YLTGEFPGDYG DTAGL ADP -----T
Mapoly0139s0012      ---KTSTPES-IWYGP RPK LGPFS-EATP YLTGEFPGDYG DTAGL ADP -----T
Cs_15915             ----------------------GPLS-SATP YLNGEFPGDYG DTAGL ADP -----T
Cs_21950             ----------IEFYGP RAT LGPFT--DTP YLNGEFPGDYG DTAGL ADP -----T
Cs_58975             ----------IEFYGP RAG LGPFT--ESP YLKGEFPGDYG DTAGL ADP -----T
Cs_27246             ----------IEFYGP RAG LGPFT--KSP YLNGEYPGDYG DTAGL ADP -----T
Cs_35121             V-KGKTTPTS-MWYGE RPK LGPFS-GNTP YLSGEYPGDYG DTAGL ADP -----T
kfl00098_0080        ---GARTSMV-EFYGP RVK LGPFS-GDTP YLKGEFAGDYG DTAGL ADP -----T
kfl00098_0090        ---GARTSMV-EFYGP RVK LGPFS-GDTP YLKGEFAGDYG DTAGL ADP -----T
kfl01434_0010        ---VKSAPSS-PWYGP RLK LGPFS-GETP YLTGEFAGDYG DTAGL ADP -----T
CrLHCBM1             V-SKASTPDS-FWYGP RPL LGAFT-GEPP YLTGEFPGDYG DTAGL ADP -----T
CrLHCBM2             ----------IEWYGP RPK LGPFSEGDTPAYLTGEFPGDYG DTAGL ADP -----T
CrLHCBM7             ----------IEWYGP RPK LGPFSEGDTPAYLTGEFPGDYG DTAGL ADP -----T
CrLHCBM3             ---AKQAPASIEFYGP RAK LGPYSENATPAYLTGEFPGDYG DTAGL ADP -----T
CrLHCBM4             ------APKSVEFYGP RAK LGPYSENSTPAYLTGEFPGDYG DTAGL ADP -----T
CrLHCBM8             ------APKSVEFYGP RAK LGPYSENSTPAYLTGEFPGDYG DTAGL ADP -----T
CrLHCBM6             ------APKSVEFYGP RAK LGPYSENATPAYLTGEFPGDYG DTAGL ADP -----T
CrLHCBM9             ------APKGIEFYGP RAK LGPYSENSTPAYLTGEFPGDYG DTAGL ADP -----T
kfl00809_0010        V----SSKAS-LWYGP RAK LGPFS-GDTPNYLTGEYAGDYG DTAGL ADP -----T
CrLHCBM5             A-NGGNEKLS-AFYGP RGL LGPLS-GTTPAYLTGEFPGDYG DSAGL ADP -----T
Cs_38026             I----------EWYGP RPK LGPYSDGATP YLKGEYPGDYG DTAGL ADP -----T
Mapoly0026s002       P-TETAVDLT-LWYGP RVK LGPFK-AATPAYLKGEFPGDYG DTAGL ADP -----T
AtLHCB3              ---KYTMGND-LWYGP RVK LGPFS-VQTP YLTGEFPGDYG DTAGL ADP -----A
Mapoly0068s0047      ---KVSMDAD-YWYGP RPK LGPFS-AQTP YLKGEFPGDYG DTAGL ADP -----S
Mapoly0068s0087      D-EEFGVDLS-LWYGAGRVK LGPFS-ARTP YLKGEFAGDYG DTAGL ADP -----S
kfl00120_0160        GGQPGAGGLG-KWYGP RAK LGPFSEGGTPAYLDGTYPGDYG DSSGL ADP -----S
Cs_28488             KAAKSASGAS-FWYGA RPG LGAFT--RSP YLNGEFPGDYG DTAGL ADP -----T
ABD37891_MvLHCB5     A---PADGAA-KWYGP RPK L-PTGPADIPAY NGE AGDYG DPLGLG- KG-----G
Gene.11404           A---PADGAA-KWYGP RPK L-PTGPADIPAY NGE AGDYG DPLGLG- KG-----G
DAA05928_MvLHCB5     A---PADGAA-KWYGP RPK L-PTGPADIPAY NGE AGDYG DPLGLG- KG-----G
Cs_64185             KQAGGAQGLD-KWYGPSRAL L-PGGPSDVPTYLNGT AGDYG DPLGLGKDTA-----T
AtLHCB5              S--ETSDELA-KWYGP RRI L-PDGRSEIPEYLNGE AGDYG DPFGLGKKP -----N
Mapoly0011s0076      P--PSNAELA-KWYGP RRI L-PEGKSDIPAYLTGE PGDYG DPFGL KKPA-----D
kfl00104_0350        RKNASDDELS-KWYGP RRL L-PEGREEVP YLTGE PGDYG DPLGLGKKE -----D
AtLHCB7              A--QEARNDS-KWYGK RPR FGPIP-YDYPPYLTGE PGDYG DIAGLGKDRL-----T
Mapoly0008s0015      Q--ADARAKS-LWFGDSRPK LGPLP-FDYPQ LQGESPGDYG DILGLGR TA-----N
kfl00104_0370        Q--TEARQRS-RWYGP RPK LGPLP-YEYPG LSGEYPADYG DILNLGVDPS-----N
CrLHCB7              A-HAEARRQS-KWYGPSRPR---PGP-LGSAE LIGA AGDAG DPLGL K QA-----S
kfl00422_0020        LTPKGVDATK-ALYGP RPK KGAFTAADVPDYLKGQYAGDYGIDVLGFCKDE -----A
AtLHCB6              -----------AAAAQP KS  -PAVNFLDPE LDGS PGD G DPLGLGKDPA-----F
Mapoly0052s0039      A----AKPAF-KGGKGG LS  -PAISFVDPE LDGS PGDYG DPLGLGRDPA-----A
kfl00276_0010        T----------RSTSGR LN  -PAYDLKNP  LDGS AGDYG DPLGLGKDPA-----A
DAA05919_MvLHCP1.2   ----PARR-------AAR---------GTEPLY SGTYSG YD PNWTEGKTEA-----Q
Gene.11156           ----PARR-------AAR---------GTEPLY SGTYSG YD PNWTEGKTEA-----Q
DAA05930_MvLHCP1.1   --------------------------------- PGTYSG YD PNWTKGKSDA-----D
Gene.11308           ----VARP-------AAR---------GSEEVY PGTYSG YD PNWTKGKSDA-----D
MpLHCAP2.1           -----SVS--T----MI A-----------DIYPFGTYPGGGESPI PFGS K-------
MpLHCP1.1            -----SAP--S----GS GTKF-GSI-APPDLYPFGVYPGGGESPI PF DDK-------
CmLHC                S-GYATLK-------M  SPAL-PF--LSKPPNLSPD PGYRG DPLRF D------AFD
```



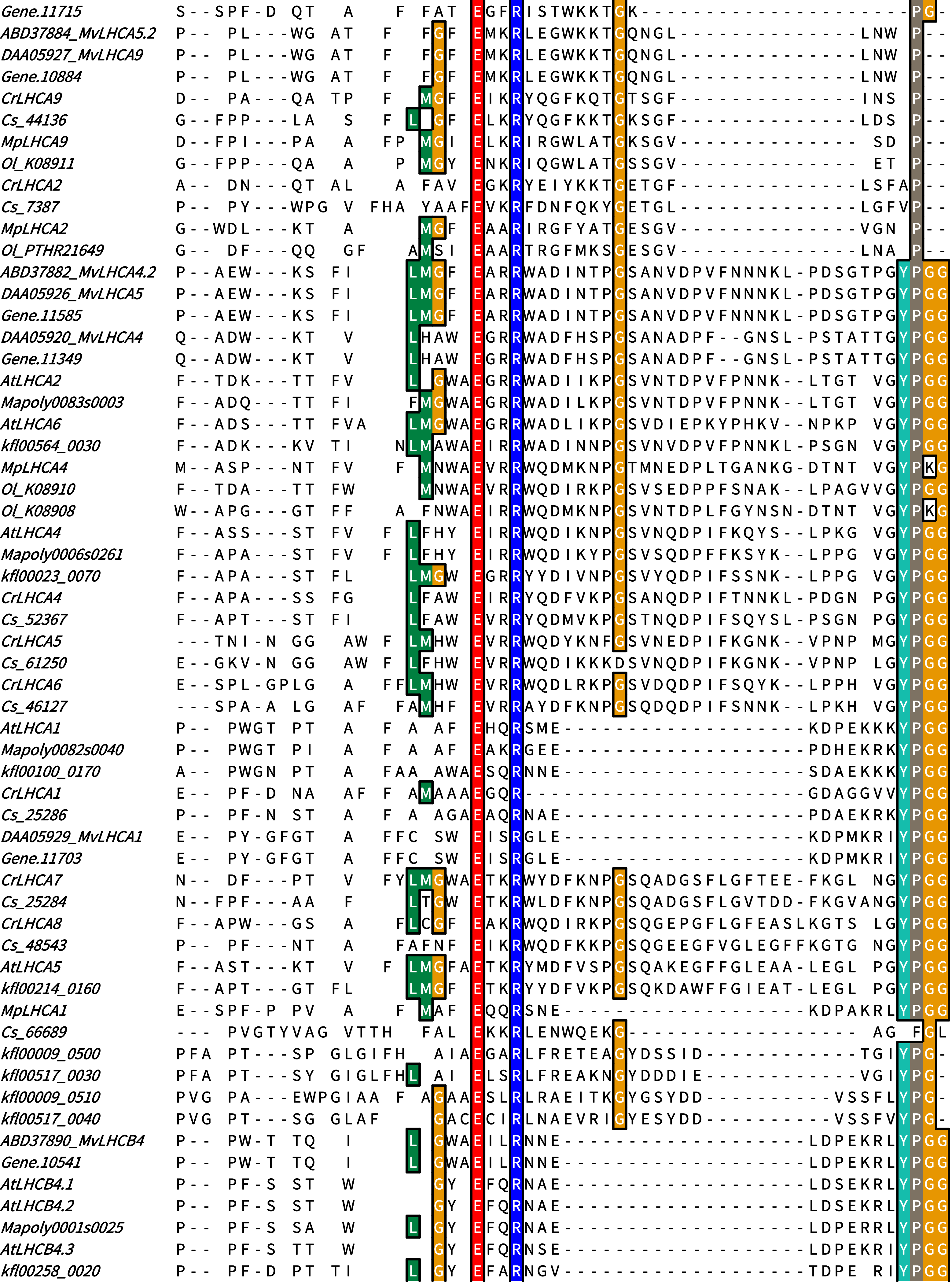
```
PyLHC              PV-AARWTMV-----KS --A -PF--MEAPPALDGS IGDVG DPLGF N------YID
CrLHCSR1           ----GKRTVS-----GKA--GA-------PVPAYAKT PGDNV DPAGFLATAS---VKD
CrLHCSR3           ----GRRTTA-----A P--QT-------AAPAYTKN PG GV DPAGFLATAS---IKD
Cs_65904           RK--GLRVLA-----LGP--QAEPS--KAATPAYAKT PGPDI DPANLLS AT---IQE
kfl00478_0030      RTFAPIRAVA-----D P--AA-PD--ATVTPKYAKD PGPNI DPATLLARAA--PIKE
DAA05932_MvLI818;2 RRNVKVRCTL-----G --PA-PP--APVSAQFCAG PGGDG DPLNFLGGKT---TAE
Gene.11245         RRNVKVRCTL-----G --PA-PP--APVSAPFCAG PGGDG DPLNFLGGKT---TAE
Gene.11340         KT--EARASK-----GS --GSKPS---------AGN VG DF DPLNL AGK ---DNE
MpLI818;1          TGVAAAVPRA-----GQ -- A------NVAAA DAG IQPEGIDVFGFFNDI ---QAE
MpLI818;2          AVTGKTKSTA-----GK -- A------DIKAATQAG VAPEGVDFFGFYNGI ---EKE

                            130       140       150       160       170       180
                   ....|....|....|....|....|....|....|....|....|....|....|....|
ABD37880_MvLHCA3   PKWLAYSE IHCR AMLGMAGM APE LGKLGLIPAETGLVW TGVIP-SEGQYDYWAS
Gene.11536         PKWLAYSE IHCR AMLGMAGM APE LGKLGLIPAETGLVW TGVIP-SEGQYDYWAS
AtLHCA3            PRWLAYGE I GR AMLGAAGA APE LGKAGLIPAETALPW TGVIP-PAGTYTYWAD
Mapoly0066s0050    PNWLRYAE I GR AMLGAAGA APE FGKIGLIPQETAIPW TGVIP-PLGQYSYWAD
kfl00560_0090      SEWLPYAE I GR AMLGAAGM APE LGKMGLIPAKTALPW  GVFP-PAGDYSYWAD
CrLHCA3            PKWL YSE IHARWAMLGAAGC APE LGAAGLIPDATNIKW  GVIP-PAGSYNYWAD
Cs_37969           PAWL YSE IHARWAMLGAAGC APE LATAGVIPQTDEVIW  GVIP-PAGVYDYWLD
MpLHCA3            PQWLAYSE IHGRWAMLGVAGV APE LGGMGIIPADTGLVW  GAIP-PQGTFDYWAD
Ol_K08909          PQWLAYSE IHGRWAMLGVAGM APE LGGLGIIPQETGLVW AGMIP-AQGTYDYWAN
ABD37881_MvLHCA4.1  KW   GE E GRWAM GVAGM FT ALTQFGISD---LPDW AGAVA------Q----
DAA05923_MvLHCA2    KW   GE E GRWAM GVAGM FT ALTQFGISD---LPDW AGAVA------Q----
Gene.11715          KW   GE E GRWAM GVAGM FT ALTQFGISD---LPDW AGAVA------Q----
ABD37884_MvLHCA5.2  AW VEAERT GRWAMAGVAGI FT ALHI-------GSAPW SIGAEE-------S---
DAA05927_MvLHCA9    AW VEAERT GRWAMAGVAGI FT ALHI-------GSAPW SIGAEE-------S---
Gene.10884          AW VEAERT GRWAMAGVAGI FT ALHI-------GSAPW SIGAEE-------S---
CrLHCA9             KW AEAEKT GRWAM AVAGI GQE LGV--------TPAW AGAKE-------Y---
Cs_44136            KW TEAEKT GRWAMAGVAGI GQE LGV--------QPKW AGAKD-------Y---
MpLHCA9             KY  EAE  SRWAM AVAGITTT ALNI--------GGPW AGAGD-------Y---
Ol_K08911           KY  EAE  ARWAM AVAGI GTE AGI--------EPRW AGTED-------Y---
CrLHCA2             KW REAE T GRWAMAAV GI FT A GL---------PK TAGAEK-------Y---
Cs_7387             PWLVEGE Y GRVAMLAVAGI LVEAAGL----------GPW SAPFRD-------   ---
MpLHCA2             PYL EAE  GRWAMHATAGI FT A GL---------PK AGDAA-------L---
Ol_PTHR21649        PYL EAE  GRWAMAATAGI FT A GL---------PKW AGAVD-------Y---
ABD37882_MvLHCA4.2  KWMV AE Q ARWAMLGVAGM SV LQGILPDN---VPQW DAPLLA------DY---
DAA05926_MvLHCA5    KWMV AE Q ARWAMLGVAGM SV LQGILPDN---VPQW DAPLLA------DY---
Gene.11585          KWMV AE Q ARWAMLGVAGM SV LQGILPDN---VPQW DAPLLA------DY---
DAA05920_MvLHCA4    KWMV AE Q ARWAMLGVAGM GA ANKVGLPA---AVQW APLVD-------Y---
Gene.11349          KWMV AE Q ARWAMLGVAGM GA ANKVGLPA---AVQW APLVD-------Y---
AtLHCA2             KWNV AE  HCRWAMLGAAGIFIPEFLTKIGILN---TPSW TAGEQE-------Y---
Mapoly0083s0003     KW V AE  HCRWAMLGAAGIFIPE LTKIGILN---TPSW TAGEQT-------Y---
AtLHCA6             KW A AE IHSRWAMLAVTGI IPECLERLGFIE---NFSW DAG RE-------Y---
kfl00564_0030       KW   AE Q GRWAMLGAAGI IPEALTSAGILH---TPHW AGAAD-------Y---
MpLHCA4             EW   AE QHARWAMLGVTGMAAPE QNPFVEG--AMPNWA APLWD------GY---
Ol_K08910           KW V AE QHGRWAMLGVAGAAAPE LTNMGISN---LPNWH AP YD------GY---
Ol_K08908          RQWMV AE QHARWAMLGVAGC APE LTKIGVAD---LPNWVDAATYQ-------Y---
AtLHCA4             KW V AE   GRWAMLGVAGM LPE FTKIGIIN---VPEW DAGKEQ-------Y---
Mapoly0006s0261     NW V AE Q GRWAMLGVAGM VPE LTKIGLIN---APLW DAGKVE-------Y---
kfl00023_0070       KW V AE Q GRWAMLGAAGA VPE LTKAGLWN---APLW DAGKAE-------Y---
CrLHCA4             KW A AE   AR AMLAVAGI VPE LSNIGFSWPGAGVAW DAGKFE-------Y---
Cs_52367            KW   AE Q GRWAMLAAAGI FV LGHLGAGGPAAATPW ASDFT-------Y---
CrLHCA5             KW R SE QHARWAMLGVAGV VQE KP--------DVY AGLPQ-NLPEP- ---
Cs_61250            KW R AE  HARWAMLAVAGI AQE KP--------GVF NAGLPE-NLPGIN GGP
CrLHCA6             KW  ESE  HSRWAMAAVAGI VQE RP--------DVFW NAGKEV-----------
Cs_46127            AW AESER HARWAMLAVAGIAVQE RP--------DVFW AALKS-PMP--- -GN
AtLHCA1             E   ESE IHCRWAMLAVPGI VPEALGYGNWVK---AQEWAPGGQAT-------YLGN
Mapoly0082s0040    FA   ESE IHCRWAMLAVPGV IPEALGLGNWVS---AQQWAPGGQAT-------YLGN
```



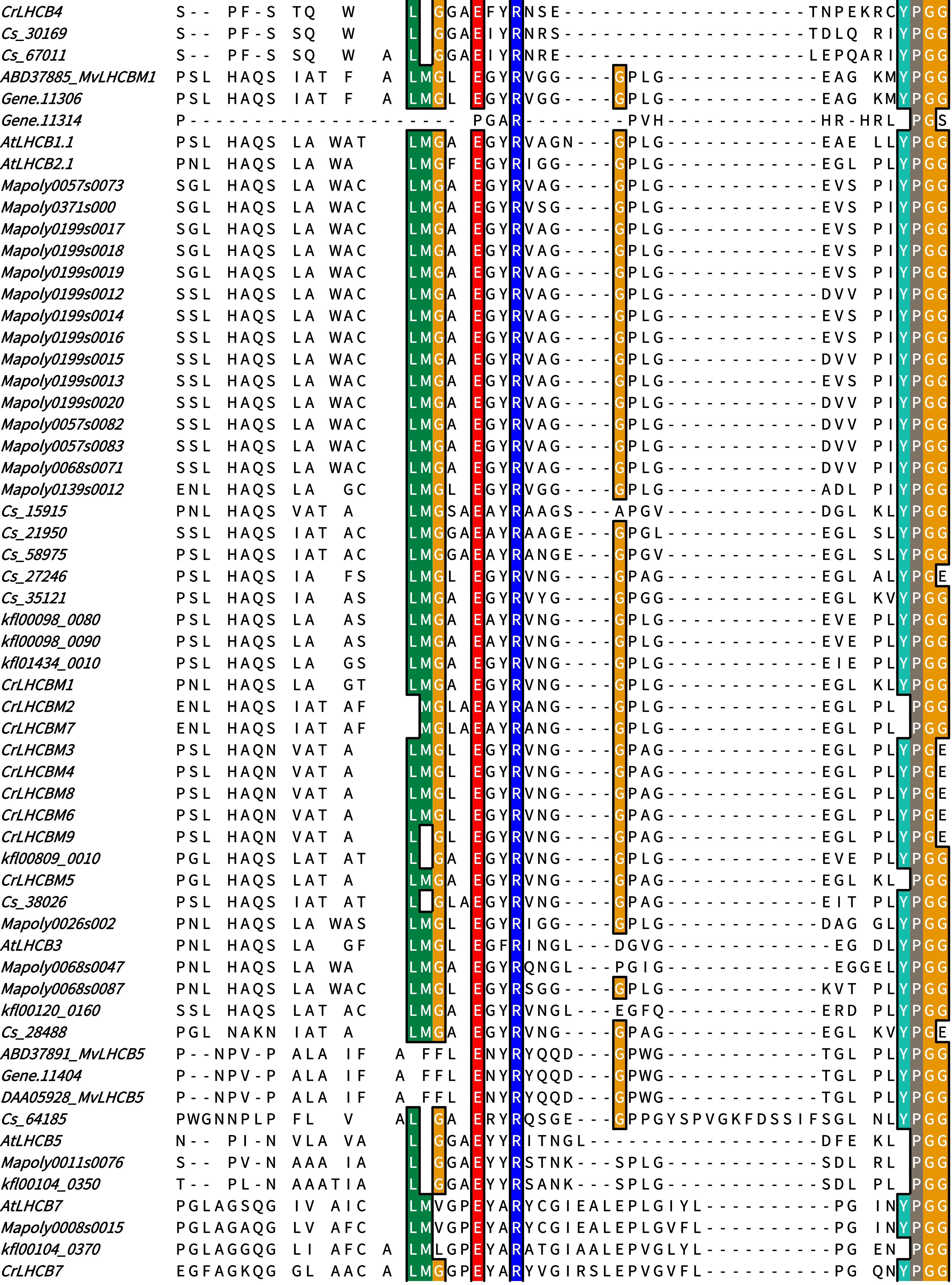
```
kfl00100_0170      A   EAE IHCRWAMLGVAGV VPEALGLGNWLN---AQDWA GGIPT-------YLGV
CrLHCA1            K  TESE IHGRWAMLGVAGS AVE LGYGNWYD---APLWANGGKAT------- FGI
Cs_25286           Q   EAE  HSRWAMAGVAGA GAE LGQGDWYQ---APLWANGGAPT-------YLGI
DAA05929_MvLHCA1   A  REAE   GRWAMLGVAGM APE  GQGTWLS---AQDWATGGKPT-------YLGN
Gene.11703         A  REAE   GRWAMLGVAGM APE  GQGTWLS---AQDWATGGKPT-------YLGN
CrLHCA7            KW V AE  HGR AMLGAAGI LTS GAKVG-LG---FPEW DAGKVV-------VEKN
Cs_25284           KW   AE  HGRTAMTAVAGI FPA ATKAGVVN---IPQW DAG VW-------VQNN
CrLHCA8            RW   AE IHCRTAMAGVAGI IPG LTKAGALN---VPEW DAGKVA-------IENS
Cs_48543           KW V SE  HSR AMAAVAGI IPE LKAFGAVN---LPVW  AGKYA-------QESS
AtLHCA5            KW V AE  HSR AMLGVAGI FT  LRTTGIRN---LPVW  AGAVK------- D--
kfl00214_0160      AW V AE  HAR AMLGAAGI GQE LRASG-RD---IPVW  AGAAS-------Y---
MpLHCA1            A   EAE IHSRWAMLGVAGAAGQEA TG---------VTWI APVQD----SQTYLGA
Cs_66689           GWLLEGE Y GRLAMLAV GV TVEFLGK---------GPW TAPAAA------A----
kfl00009_0500      KT MAEAE FHGRTAMLAV GACVPE LGR---------GDW SAAHTMVDGGGNS ILS
kfl00517_0030      AT  VEAE FHGRLAMLAT GAFVPEA GR---------EDW TAAHKAIDAGGSS ILA
kfl00009_0510      RN LVEGE   GR AMLAV GA VPE LGKGNFGDIA-ASAP G TIL-----GLE----
kfl00517_0040      RN VIEGE   GRWAMLAC GA VP LGKGNFGDLARSSAWLGGPVL-----GLE----
ABD37890_MvLHCB4   E  RETE I GRWAMLGVAGA AGE QTG---------INWVDIGKVM-LEQGSS GGL
Gene.10541         E  RETE I GRWAMLGVAGA AGE QTG---------INWVDIGKVM-LEQGSS GGL
AtLHCB4.1          Q  RECE IHGRWAMLAT GA SVEWLTG---------VTWQDAGKVE-LVDGSSYLGQ
AtLHCB4.2          Q  RECE IHGRWAMLAT GA TVEWLTG---------VTWQDAGKVE-LVDGSSYLGQ
Mapoly0001s0025    Q  RECE IHGRWCMLAT GA SVEAFTG---------VTWQDAGKVE-LIEGSTYFGL
AtLHCB4.3          Q  RECE IHGRWAMLGT GA AVEALTG---------IAWQDAGKVE-LVEGSSYLGQ
kfl00258_0020      T  RETE IHGRWAMLGV GA AVEALTG---------VHWQDAGKVE-LVDGASYLGL
CrLHCB4            A  RECE IHGRWAMLAC GA VAEATTG---------VSWV AGKVE-L-DGASYAGL
Cs_30169           Q  RETE IHGRWAMLGA GV IAEATTG---------VSWVDAGKVE-L-DGTQYLGF
Cs_67011           Q  RETE IHGRWAMLGA GV VAEASTG---------VSWVDAGKVE-L-DGAQYLGF
ABD37885_MvLHCBM1  FAA REIE  HARWAMLGT GCFLPEALEKSG-INFG-ESVW AGGQIFQEGGLDYLGN
Gene.11306         FAA REIE  HARWAMLGT GCFLPEALEKSG-INFG-ESVW AGGQIFQEGGLDYLGN
Gene.11314         FAA REIE  HARWAMLGT GCFLPEALEKSG-INFG-ESVW AGGQIFQEGGLDYLGK
AtLHCB1.1          FA NRELE IHSRWAMLGA GC FPE LARNG-VKFG-EAVW AG QIFSDGGLDYLGN
AtLHCB2.1          FA NRELE IHSRWAMLGA GCTFPE LSKNG-VKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0057s0073    FA NRELE IHARWAMLGA GC TPE LAKNG-VKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0371s000     FA NRELE IHARWAMLGA GC TPE LAKNG-VKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0199s0017    FA NRELE IHARWAMLGA GC TPE LAKNG-VKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0199s0018    FA NRELE IHARWAMLGA GC TPE LAKNG-VKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0199s0019    FA NRELE IHARWAMLGA GC TPE LAKNG-VKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0199s0012    FA NRELE IHSRWAMLGA GC FPE LSKNG-VTFG-EAVW AG QIFAEGGLDYLGN
Mapoly0199s0014    FA NRELE IHARWAMLGA GC FPE LSKNG-VSFG-EAVW AG QIFAEGGLDYLGN
Mapoly0199s0016    FA NRELE IHARWAMLGA GC FPE LSKNG-VSFG-EAVW AG QIFAEGGLDYLGN
Mapoly0199s0015    FA NRELE IHARWAMLGA GC FPE LSKNG-VSFG-EAVW AG QIFAEGGLDYLGN
Mapoly0199s0013    FA NRELE IHARWAMLGA GC FPE LSKNG-VSFG-EAVW AG QIFAEGGLDYLGN
Mapoly0199s0020    FA NRELE IHSRWAMLGA GC TPE LSKNG-VKFG-EAVW AG QIFAEGGLDYLGN
Mapoly0057s0082    FA NRELE IHARWAMLGA GC TPE LAKNG-TKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0057s0083    FA NRELE IHARWAMLGA GC TPE LAKNG-TKFG-EAVW AG QIFSEGGLDYLGN
Mapoly0068s0071    FA NRELE IHARWAMLGA GC FPE LAKNG-VKFG-EAVW AG QIFAEGGLDYLGN
Mapoly0139s0012    FS NRELE IHSRWAMLGT GM FPE LAKNG-ITFG-EPIW AG QIFADGGLNYLGN
Cs_15915           FA  REIE IHARWAMLGA GC TPE LSQNG-FTFQ-EPVW AGAQILSSEGLDYLGK
Cs_21950           FA  REIE IHARWAMLGA GC TPE LAKNG-VSFG-EAVW AGAQIFGSDGLNYLGN
Cs_58975           FA  REIE IHARWAMLGA GC TPE LAKNG-VSFG-EAVW AGAQIFGSDGLNYLGN
Cs_27246           FA  REIE IHARWAMLGA GC TPE LAKNG-VSFG-EAVW AG QIFAPGGLDYLGN
Cs_35121           FA  REIE IHARWAMLGA GC TPE LAKNG-VSFG-EAVW AGAQIFAPGGLDYLGN
kfl00098_0080      FA NRELE IHARWAMLGA GC FPE LAQNG-TKFG-EAVW AG QIFASGGLDYLGN
kfl00098_0090      FA NRELE IHARWAMLGA GC FPE LAQNG-TKFG-EAVW AG QIFASGGLDYLGN
kfl01434_0010      FA NRELE IHARWAMLGA GC TPE LAGNG-VKFG-ESVW AG QIFAPGGLDYLGN
CrLHCBM1           FK  RELE IHARWAMLGA GC FPE LGSYG-VPFG-EAVW AGAQIFQEGGLDYLGN
CrLHCBM2           FK  RELE IHARWAMLGA GC TPE LAKNG-IPFG-EAVW AGAQIFAEGGLNYLGN
CrLHCBM7           FK  RELE IHARWAMLGA GC TPE LAKNG-IPFG-EAVW AGAQIFAEGGLNYLGN
CrLHCBM3           FK  RELE IHARWAMLGA GC TPE LAKSG-TQFG-EAVW AGAQIFSEGGLDYLGN
```



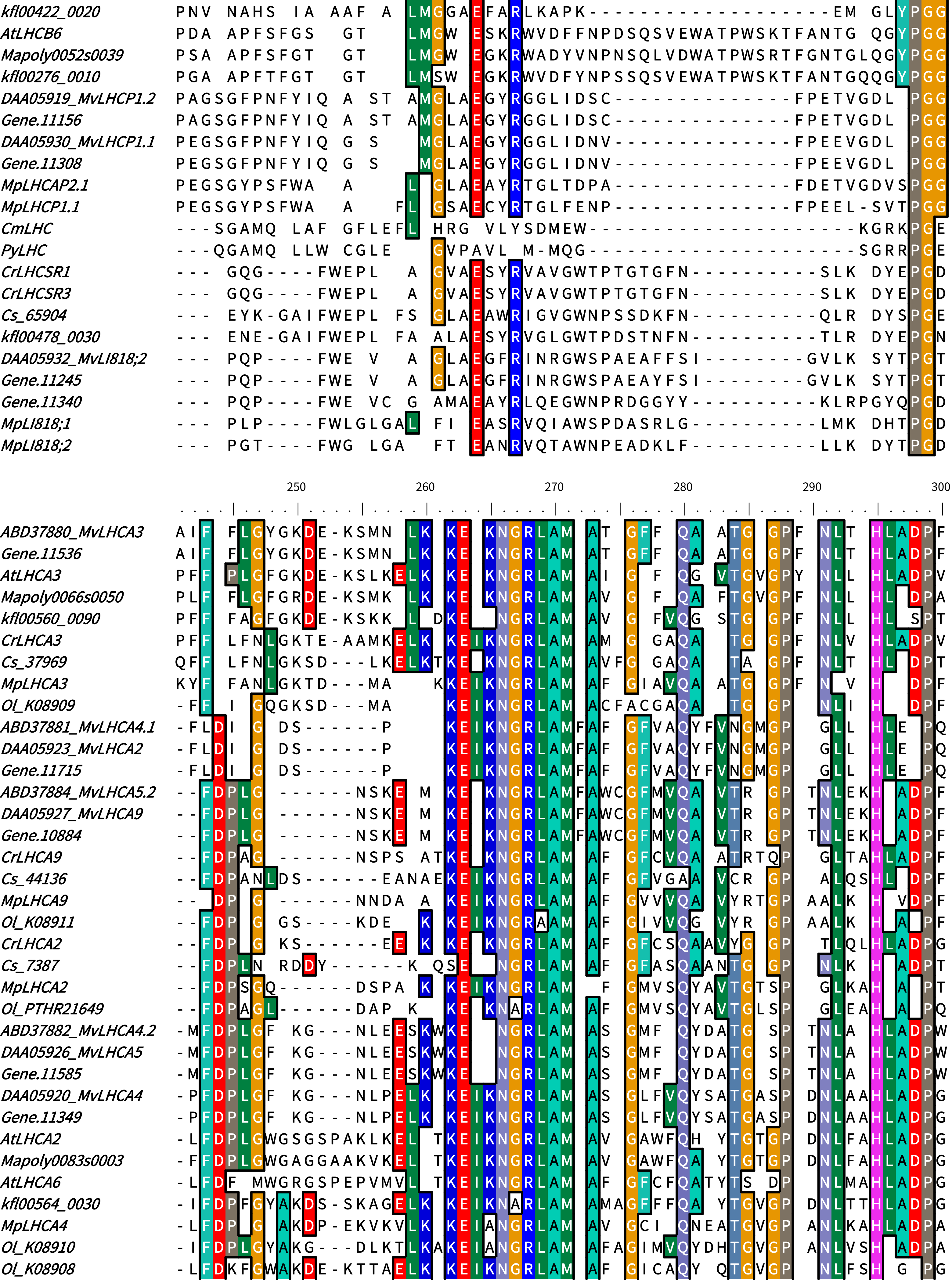
```
CrLHCBM4            FK  RELE IHARWAMLGA GC TPE LAKNG-TKFG-EAVW  AGAQIFSEGGLDYLGN
CrLHCBM8            FK  RELE IHARWAMLGA GC TPE LAKSG-TKFG-EAVW  AGAQIFSEGGLDYLGN
CrLHCBM6            FK  RELE IHARWAMLGA GC TPE LAKSG-TKFG-EAVW  AGAQIFSEGGLDYLGN
CrLHCBM9            FK  RELE IHARWAMLGA GI TPE LQKNG-VQFG-EAVW  AGAQIFQEGGLNYLGN
kfl00809_0010       FR NRELE IHARWA LGA GI TPEALEKAG-VKFG-EAVW  AGAQIFSADGLNYLGN
CrLHCBM5            FK  RELE IHARWAMLGA GC TPE LAKNG-TPIV-EPVW  AGAQIFAEGGLDYLGN
Cs_38026            FS  REIE IHSRWA LGT GI TPE LEKGG-VKFQ-EAVW  AGAQIFSADGLNYLGN
Mapoly0026s002      FA NRELE IHSRWAMLGAAGC FPE LAKNA-VSFR-EPVW  AG QIFSDGGLDYLGN
AtLHCB3             FA NRALE IHGRWAMLGAFGC TPE LQKVR-VDFK-EPVW  AG QIFSEGGLDYLGN
Mapoly0068s0047     FA NRALE IHGRWAMLGA GC LPEALVKSR-VTLK-EAVW  AG QIFTDGGLDYLGN
Mapoly0068s0087     FA NRELE IHGRWAMLGWAGCFFPE LVKKS-CGLK-EGVW  AG QIFTDGGLDYLGN
kfl00120_0160       FR NRELE IHSRWAMLGA GM LPE LADSG-IPIK-EPVW  AGAQIFDSDGLNYLGN
Cs_28488            FS  REIE IHARWAMLGA GI VPE LDQTN------HVAW DAGATIFGPKGIQYLGV
ABD37891_MvLHCB5     E  RPYE  HGRWAMLGV GM VPEGLYANGNTNIK-GAVW D GAVLLDPSTLT AGI
Gene.11404           E  RPYE  HGRWAMLGV GM VPEGLYANGNTNIK-GAVW D GAVLLDPSTLT AGI
DAA05928_MvLHCB5     E  RPYE  HGRWAMLGV GM VPEGLYANGNTNIK-GAVW D GAVLLDPSTLT AGI
Cs_64185             E  RAYE IHARWAMLAAAGI IPEGLQANG-AAIK-GGTW  TGAEMLNGGTLNYFAV
AtLHCB5             FA   AFE IHARWAMLGAAGF IPEALNKYGANCGP-EAVW  TGALLLDGNTLNYFGK
Mapoly0011s0076     FD   AYE IHARWAMLGAAGF LPEAFNKYGAVCGP-EAVW  TGALLLEGNTLQYFGA
kfl00104_0350       FD  RAAE IHARWAMLGAAGI IPEAFNRSGLPCGP-EAVW  TGAQLLEGESLQ CGI
AtLHCB7             FD  FNFE  HARWAMLAA GA IPE FDLTGTFHFA-EPVW RVGYSKLQGETLEYLGI
Mapoly0008s0015     FD  FNFE  HARWAMLGA GA IPE LVRYGGLSFS-EPVW RVGYAKLQGETLDYFGI
kfl00104_0370       FQ  FNFE  HARWAMLGA GA IPE LQYFSALDFT-EPVW  VGYAKLQGEDLDYFGI
CrLHCB7             FA   EAE  HARWAMLGV GC VPE LALRG-VDLG-EPVW  VGASKLNSDTLN GGI
kfl00422_0020       FA LRAQE F ARWAMLGV GM YPEFFPAEG-----FEPVW  TGAQIFDPAGIDYLGA
AtLHCB6              KW REAE IHGRWAMAAV GIFVG AWSG---------VAW  AGAQ------------
Mapoly0052s0039      KW REAE IHGRWAMAAV GIFVG AWSG---------IPW  AGAD------------
kfl00276_0010        KW REAE IHGRWA  AT GC IG AFSG---------IPW  AGAA------------
DAA05919_MvLHCP1.2   AYNVEVE IHGRWAMLGCAGAWAAEQGTG---------INW  AGAICTPADGIH PGE
Gene.11156           AYNVEVE IHGRWAMLGCAGAWAAEQGTG---------INW  AGAICTPADGIH PGE
DAA05930_MvLHCP1.1   TYNVSVE IHGRWAMLGCAGAWAAEQGTG---------INW  AGAICTPADGIH PGE
Gene.11308           TYNVSVE IHGRWAMLGCAGAWAAEQGTG---------INW  AGAICTPADGIH PGE
MpLHCAP2.1          ---NAERE IHGRWAMLGVTGAWAAENGTG---------IPW TAGTLCTPDDAVK PGA
MpLHCP1.1           ---NAERE IHGRWAMLGVTGAWAAENGTG---------IPW TAGTLCTPDDAVK PGA
CmLHC                NWL EGE K GRVAMLAC HFFVTEFYQF---------PF AGAPKLA----AHDYFVK
PyLHC                KFLREAE KHCRVTMLAV GLFVQEFYTL---------PF SGGPALA----SHNYFVT
CrLHCSR1             R  RESE THGRVAMLAA GF VGEQLQD---------FPL FDGRVS----AIY FQQ
CrLHCSR3             R  RESE THGRVAMLAA GF VGEQLQD---------FPL WDGRVS----AIY FQQ
Cs_65904             K  RESE THGRVAMLAA GW VGEF AD---------KKLLSDGRIT----AID FQQ
kfl00478_0030        K  RESE  HGRVAMLAT GF VGEQ ED---------FPARFFPHVT----AIY FQQ
DAA05932_MvLI818;2   KMLRESE  HGRVAMLAT GF VGETFNP---------L---FGG IT----AINQFQQ
Gene.11245           KMLRESE  HGRVAMLAT GF VGETFNP---------L---FGG IT----AINQFQQ
Gene.11340           LL REAE THGRV MLATAGF AE FNPL-------------FNG IK----AIN FQQ
MpLI818;1           AQ  A VE THGRLAMLAS GF VGES EG---------SS LFDSQVT----AIN FQQ
MpLI818;2           AQ  A VE THGRIAMLAS GF VGEQ EG---------SA LFDANIT----AID FQQ

                            190       200       210       220       230       240
                    ....|....|....|....|....|....|....|....|....|....|....|....|
ABD37880_MvLHCA3    P--------LA FW    LMQFAELRRWQDYRHPGSQSKQYFLGLEQFFGGSG PSYPGG
Gene.11536          P--------LA FW    LMQFAELRRWQDYRHPGSQSKQYFLGLEQFFGGSG PSYPGG
AtLHCA3             N--------YT FV   ALMGFAEHRRLQDWYNPGSMGKQYFLGLEKGLAGSG PAYPGG
Mapoly0066s0050     P--------YT FV   ALMGFAEHRRAQDYYKPGSMGKQYFLGFEKVLGGSG PAYPGG
kfl00560_0090       S--------YT FG    LMAFAEHKRLADYRKPGSQGKVFFLGMEKFLGGSG PAYPGG
CrLHCA3             P--------YT FF    AMQFAELRRLQDFRYPGSMGQQYFLGLEAIFKGSG AAYPGG
Cs_37969            P--------YS FF    L QFAELRRWQDFRNPGSQGKQYFLGLEEVLKGSG PSYPGG
MpLHCA3             P--------TT FW N  LMNFAEVKRGQDYWYPGSQGETPLMGWEKGFAGSGAPAYPGG
Ol_K08909           P--------FT FW NAALMNFAELRRAQDYWNPGSMGKQELIGWEKMLGGSG PAYPGG
ABD37881_MvLHCA4.1  S--SPF-D QT  A   F FAT EGFRISTWKKTGK---------------------PG-
DAA05923_MvLHCA2    S--SPF-D QT  A   F FAT EGFRISTWKKTGK---------------------PG-
```



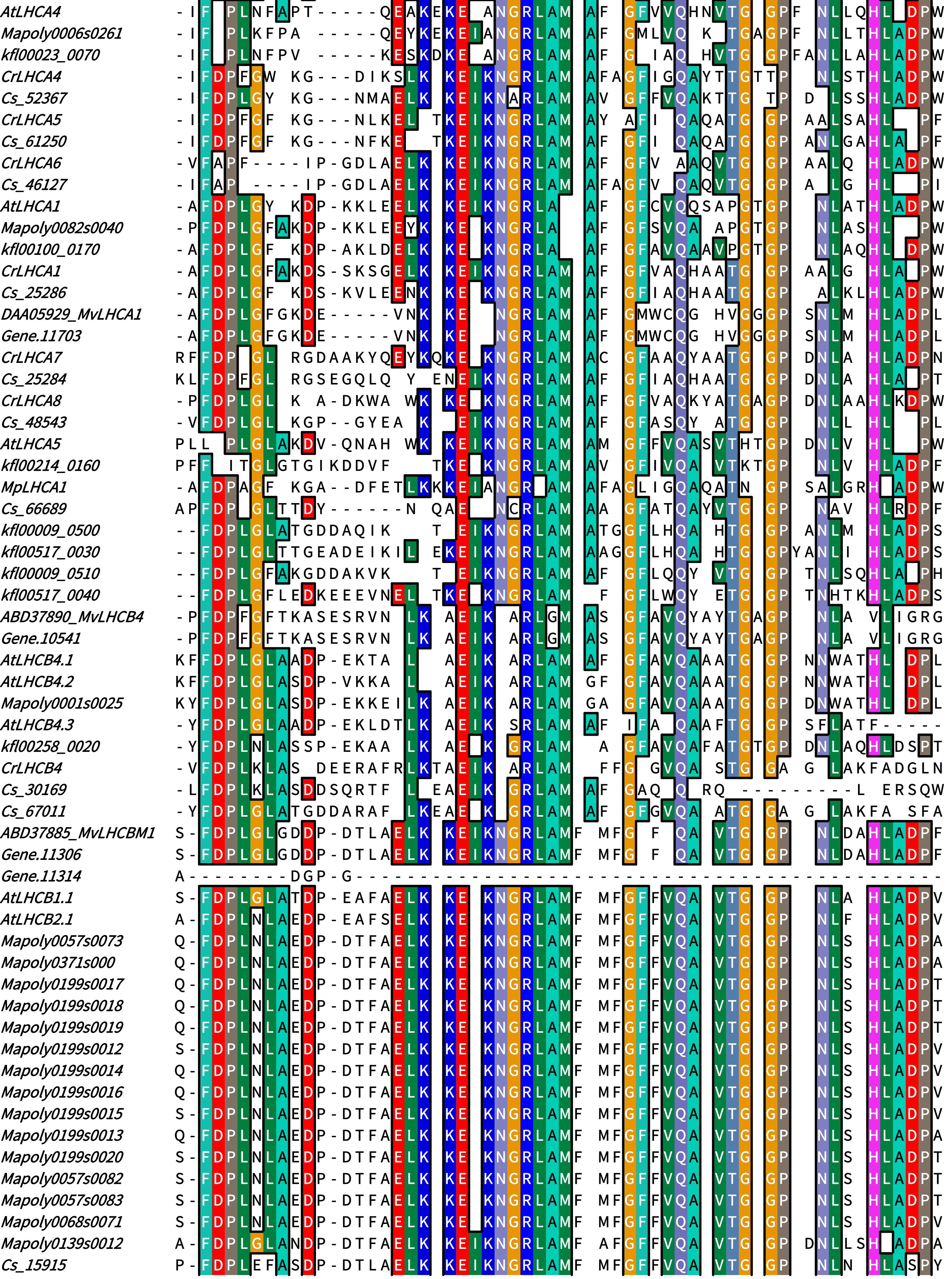
```
Gene.11715          S--SPF-D QT  A    F  FAT EGFRISTWKKTGK------------------PG-
ABD37884_MvLHCA5.2  P-- PL---WG  AT   F   FGF EMKRLEGWKKTGQNGL--------------LNW P--
DAA05927_MvLHCA9    P-- PL---WG  AT   F   FGF EMKRLEGWKKTGQNGL--------------LNW P--
Gene.10884          P-- PL---WG  AT   F   FGF EMKRLEGWKKTGQNGL--------------LNW P--
CrLHCA9             D-- PA---QA  TP   F   MGF EIKRYQGFKQTGTSGF--------------INS P--
Cs_44136            G--FPP---LA  S    F L  GF ELKRYQGFKKTGKSGF--------------LDS P--
MpLHCA9             D--FPI---PA  A    FP  MGI ELKRIRGWLATGKSGV-------------- SD P--
Ol_K08911           G--FPP---QA  A    P   MGY ENKRIQGWLATGSSGV-------------- ET P--
CrLHCA2             A-- DN---QT  AL   A  FAV EGKRYEIYKKTGETGF--------------LSFAP--
Cs_7387             P-- PY---WPG V FHA  YAAFEVKRFDNFQKYGETGL--------------LGFVP--
MpLHCA2             G--WDL---KT  A        MGF EAARIRGFYATGESGV--------------VGN P--
Ol_PTHR21649        G-- DF---QQ  GF      AMSI EAARTRGFMKSGESGV--------------LNA P--
ABD37882_MvLHCA4.2  P--AEW---KS  FI      LMGF EARRWADINTPGSANVDPVFNNNKL-PDSGTPGYPGG
DAA05926_MvLHCA5    P--AEW---KS  FI      LMGF EARRWADINTPGSANVDPVFNNNKL-PDSGTPGYPGG
Gene.11585          P--AEW---KS  FI      LMGF EARRWADINTPGSANVDPVFNNNKL-PDSGTPGYPGG
DAA05920_MvLHCA4    Q--ADW---KT  V       LHAW EGRRWADFHSPGSANADPF--GNSL-PSTATTGYPGG
Gene.11349          Q--ADW---KT  V       LHAW EGRRWADFHSPGSANADPF--GNSL-PSTATTGYPGG
AtLHCA2             F--TDK---TT  FV      L GWAEGRRWADIIKPGSVNTDPVFPNNK--LTGT VGYPGG
Mapoly0083s0003     F--ADQ---TT  FI      FMGWAEGRRWADILKPGSVNTDPVFPNNK--LTGT VGYPGG
AtLHCA6             F--ADS---TT  FVA     LMGWAEGRRWADLIKPGSVDIEPKYPHKV--NPKP VGYPGG
kfl00564_0030       F--ADK---KV  TI     NLMAWAEIRRWADINNPGSVNVDPVFPNNKL-PSGN VGYPGG
MpLHCA4             M--ASP---NT  FV   F  MNWAEVRRWQDMKNPGTMNEDPLTGANKG-DTNT VGYPKG
Ol_K08910           F--TDA---TT  FW      MNWAEVRRWQDIRKPGSVSEDPPFSNAK--LPAGVVGYPGG
Ol_K08908           W--APG---GT  FF   A  FNWAEIRRWQDMKNPGSVNTDPLFGYNSN-DTNT VGYPKG
AtLHCA4             F--ASS---ST  FV   F LFHY EIRRWQDIKNPGSVNQDPIFKQYS--LPKG VGYPGG
Mapoly0006s0261     F--APA---ST  FV   F LFHY EIRRWQDIKYPGSVSQDPFFKSYK--LPPG VGYPGG
kfl00023_0070       F--APA---ST  FL      LMGW EGRRYYDIVNPGSVYQDPIFSSNK--LPPG VGYPGG
CrLHCA4             F--APA---SS  FG      LFAW EIRRYQDFVKPGSANQDPIFTNNKL-PDGN PGYPGG
Cs_52367            F--APT---ST  FI      LFAW EVRRYQDMVKPGSTNQDPIFSQYSL-PSGN PGYPGG
CrLHCA5             ---TNI-N GG  AW   F LMHW EVRRWQDYKNFGSVNEDPIFKGNK--VPNP MGYPGG
Cs_61250            E--GKV-N GG  AW   F LFHW EVRRWQDIKKKDSVNQDPIFKGNK--VPNP LGYPGG
CrLHCA6             E--SPL-GPLG  A   FFLMHW EVRRWQDLRKPGSVDQDPIFSQYK--LPPH VGYPGG
Cs_46127            ---SPA-A LG  AF   FAMHF EVRRAYDFKNPGSQDQDPIFSNNK--LPKH VGYPGG
AtLHCA1             P-- PWGT PT  A    F A AF EHQRSME------------------KDPEKKKYPGG
Mapoly0082s0040     P-- PWGT PI  A    F A AF EAKRGEE------------------PDHEKRKYPGG
kfl00100_0170       A-- PWGN PT  A    FAA AWAESQRNNE------------------SDAEKKKYPGG
CrLHCA1             E-- PF-D NA  AF  F AMAAAEGQR---------------------GDAGGVVYPGG
Cs_25286            P-- PF-N ST  A    F A AGAEAQRNAE------------------PDAEKRKYPGG
DAA05929_MvLHCA1    E-- PY-GFGT  A    FFC SW EISRGLE------------------KDPMKRIYPGG
Gene.11703          E-- PY-GFGT  A    FFC SW EISRGLE------------------KDPMKRIYPGG
CrLHCA7             N-- DF---PT  V    FYLMGWAETKRWYDFKNPGSQADGSFLGFTEE-FKGL NGYPGG
Cs_25284            N--FPF---AA  F       LTGW ETKRWLDFKNPGSQADGSFLGVTDD-FKGVANGYPGG
CrLHCA8             F--APW---GS  A       FLCGF EAKRWQDIRKPGSQGEPGFLGFEASLKGTS LGYPGG
Cs_48543            P-- PF---NT  A      FAFNF EIKRWQDFKKPGSQGEEGFVGLEGFFKGTG NGYPGG
AtLHCA5             F--AST---KT  V    F LMGFAETKRYMDFVSPGSQAKEGFFGLEAA-LEGL PGYPGG
kfl00214_0160       F--APT---GT  FL      LMGF ETKRYYDFVKPGSQKDAWFFGIEAT-LEGL PGYPGG
MpLHCA1             E--SPF-P PV  A    F  MAF EQQRSNE------------------KDPAKRLYPGG
Cs_66689            --- PVGTYVAG VTTH   FAL EKKRLENWQEKG------------------ AG FGL
kfl00009_0500       PFA PT---SP  GLGIFH  AIAEGARLFRETEAGYDSSID-----------TGIYPG-
kfl00517_0030       PFA PT---SY  GIGLFHL AI ELSRLFREAKNGYDDDIE------------VGIYPG-
kfl00009_0510       PVG PA---EWPGIAA F  AGAAESLRLRAEITKGYGSYDD----------VSSFLYPG-
kfl00517_0040       PVG PT---SG  GLAF     GACECIRLNAEVRIGYESYDD----------VSSFVYPG-
ABD37890_MvLHCB4    P-- PW-T TQ  I       L GWAEILRNNE------------------LDPEKRLYPGG
Gene.10541          P-- PW-T TQ  I       L GWAEILRNNE------------------LDPEKRLYPGG
AtLHCB4.1           P-- PF-S ST  W          GY EFQRNAE------------------LDSEKRLYPGG
AtLHCB4.2           P-- PF-S ST  W          GY EFQRNAE------------------LDSEKRLYPGG
Mapoly0001s0025     P-- PF-S SA  W       L  GY EFQRNAE------------------LDPERRLYPGG
AtLHCB4.3           P-- PF-S TT  W          GY EFQRNSE------------------LDPEKRIYPGG
kfl00258_0020       P-- PF-D PT  TI      L  GY EFARNGV------------------TDPE RIYPGG
```



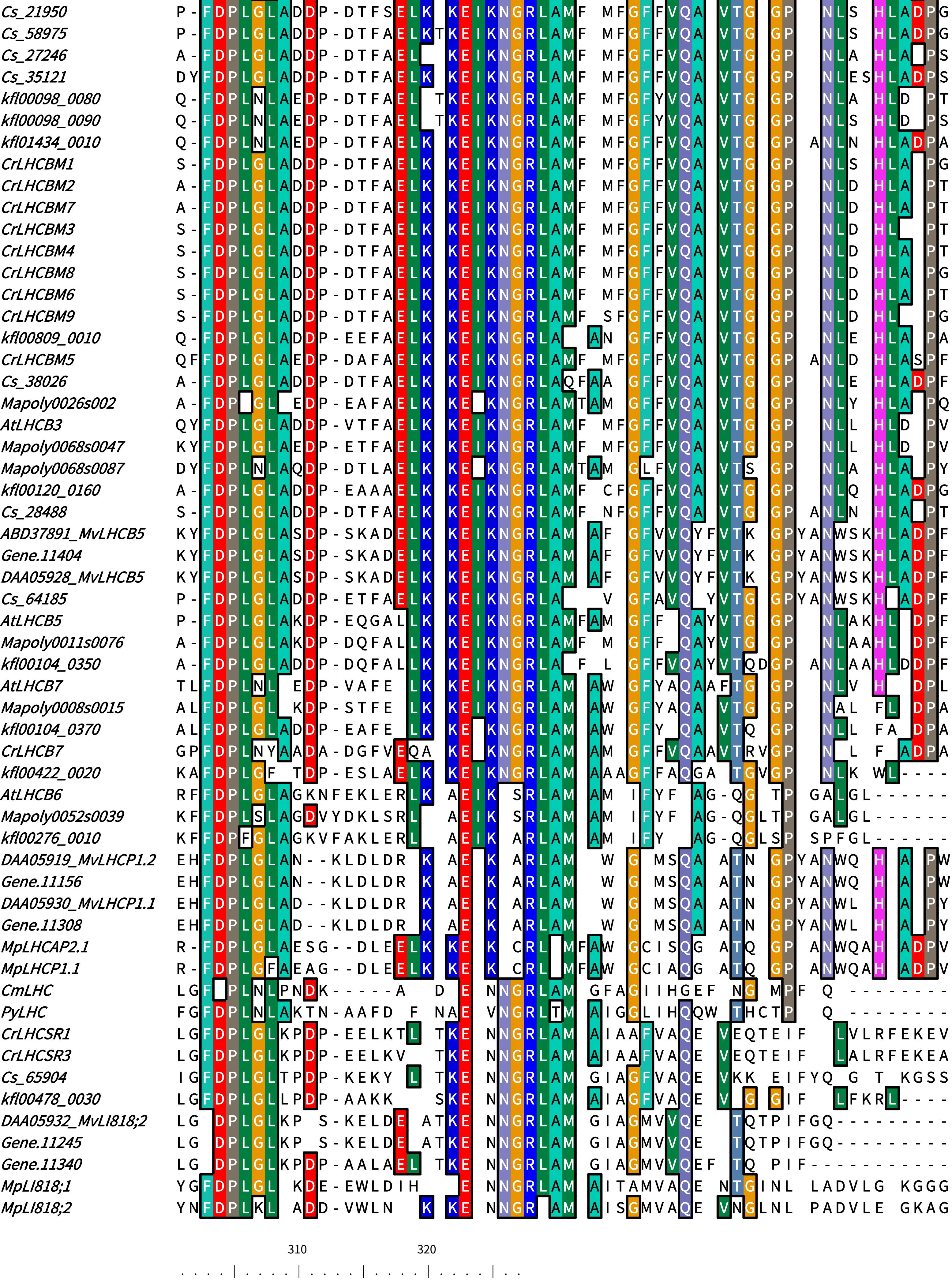
```
CrLHCB4             S-- PF-S TQ  W    L GGAEFYRNSE-------------------TNPEKRCYPGG
Cs_30169            S-- PF-S SQ  W    L GGAEIYRNRS-------------------TDLQ RIYPGG
Cs_67011            S-- PF-S SQ  W  A L GGAEIYRNRE-------------------LEPQARIYPGG
ABD37885_MvLHCBM1   PSL HAQS IAT F  A LMGL EGYRVGG----GPLG------------EAG KMYPGG
Gene.11306          PSL HAQS IAT F  A LMGL EGYRVGG----GPLG------------EAG KMYPGG
Gene.11314          P--------------------- PGAR--------PVH------------HR-HRL PGS
AtLHCB1.1           PSL HAQS LA WAT   LMGA EGYRVAGN---GPLG------------EAE LLYPGG
AtLHCB2.1           PNL HAQS LA WA    LMGF EGYRIGG----GPLG------------EGL PLYPGG
Mapoly0057s0073     SGL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0371s000      SGL HAQS LA WAC   LMGA EGYRVSG----GPLG------------EVS PIYPGG
Mapoly0199s0017     SGL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0199s0018     SGL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0199s0019     SGL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0199s0012     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------DVV PIYPGG
Mapoly0199s0014     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0199s0016     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0199s0015     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------DVV PIYPGG
Mapoly0199s0013     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------EVS PIYPGG
Mapoly0199s0020     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------DVV PIYPGG
Mapoly0057s0082     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------DVV PIYPGG
Mapoly0057s0083     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------DVV PIYPGG
Mapoly0068s0071     SSL HAQS LA WAC   LMGA EGYRVAG----GPLG------------DVV PIYPGG
Mapoly0139s0012     ENL HAQS LA  GC   LMGL EGYRVGG----GPLG------------ADL PIYPGG
Cs_15915            PNL HAQS VAT A    LMGSAEAYRAAGS---APGV------------DGL KLYPGG
Cs_21950            SSL HAQS IAT AC   LMGGAEAYRAAGE---GPGL------------EGL SLYPGG
Cs_58975            PSL HAQS IAT AC   LMGGAEAYRANGE---GPGV------------EGL SLYPGG
Cs_27246            PSL HAQS IA  FS   LMGL EGYRVNG----GPAG------------EGL ALYPGE
Cs_35121            PSL HAQS IA  AS   LMGA EGYRVYG----GPGG------------EGL KVYPGG
kfl00098_0080       PSL HAQS LA  AS   LMGA EAYRVNG----GPLG------------EVE PLYPGG
kfl00098_0090       PSL HAQS LA  AS   LMGA EAYRVNG----GPLG------------EVE PLYPGG
kfl01434_0010       PSL HAQS LA  GS   LMGA EGYRVNG----GPLG------------EIE PLYPGG
CrLHCBM1            PNL HAQS LA  GT   LMGA EGYRVNG----GPLG------------EGL KLYPGG
CrLHCBM2            ENL HAQS IAT AF    MGLAEAYRANG----GPLG------------EGL PL PGG
CrLHCBM7            ENL HAQS IAT AF    MGLAEAYRANG----GPLG------------EGL PL PGG
CrLHCBM3            PSL HAQN VAT A    LMGL EGYRVNG----GPAG------------EGL PLYPGE
CrLHCBM4            PSL HAQN VAT A    LMGL EGYRVNG----GPAG------------EGL PLYPGE
CrLHCBM8            PSL HAQN VAT A    LMGL EGYRVNG----GPAG------------EGL PLYPGE
CrLHCBM6            PSL HAQN VAT A    LMGL EGYRVNG----GPAG------------EGL PLYPGE
CrLHCBM9            PSL HAQN VAT A    L GL EGYRVNG----GPAG------------EGL PLYPGE
kfl00809_0010       PGL HAQS LAT AT   L GA EGYRVNG----GPLG------------EVE PLYPGG
CrLHCBM5            PGL HAQS LAT A    LMGA EGYRVNG----GPAG------------EGL KL PGG
Cs_38026            PSL HAQS IAT AT   L GLAEGYRVNG----GPAG------------EIT PLYPGG
Mapoly0026s002      PNL HAQS LA WAS   LMGL EGYRIGG----GPLG------------DAG GLYPGG
AtLHCB3             PNL HAQS LA  GF   LMGL EGFRINGL---DGVG-------------EG DLYPGG
Mapoly0068s0047     PNL HAQS LA WA    LMGA EGYRQNGL---PGIG-------------EGGELYPGG
Mapoly0068s0087     PNL HAQS LA WAC   LMGL EGYRSGG----GPLG------------KVT PLYPGG
kfl00120_0160       SSL HAQS LAT AC   LMGA EGYRVNGL---EGFQ------------ERD PLYPGG
Cs_28488            PGL NAKN IAT A    LMGA EGYRVNG----GPAG------------EGL KVYPGE
ABD37891_MvLHCB5    P--NPV-P ALA IF  A FFL ENYRYQQD---GPWG------------TGL PLYPGG
Gene.11404          P--NPV-P ALA IF  A FFL ENYRYQQD---GPWG------------TGL PLYPGG
DAA05928_MvLHCB5    P--NPV-P ALA IF  A FFL ENYRYQQD---GPWG------------TGL PLYPGG
Cs_64185            PWGNNPLP FL  V   AL GA ERYRQSGE---GPPGYSPVGKFDSSIFSGL NLYPGG
AtLHCB5             N-- PI-N VLA VA   L GGAEYYRITNGL------------------DFE KL PGG
Mapoly0011s0076     S-- PV-N AAA IA   L GGAEYYRSTNK---SPLG------------SDL RL PGG
kfl00104_0350       T-- PL-N AAATIA   L GGAEYYRSANK---SPLG------------SDL PL PGG
AtLHCB7             PGLAGSQG IV AIC   LMVGPEYARYCGIEALEPLGIYL----------PG INYPGG
Mapoly0008s0015     PGLAGAQG LV AFC   LMVGPEYARYCGIEALEPLGVFL----------PG INYPGG
kfl00104_0370       PGLAGGQG LI AFC A LMLGPEYARATGIAALEPVGLYL----------PG EN PGG
CrLHCB7             EGFAGKQG GL AAC A LMGGPEYARYVGIRSLEPVGVFL----------PG QNYPGG
```



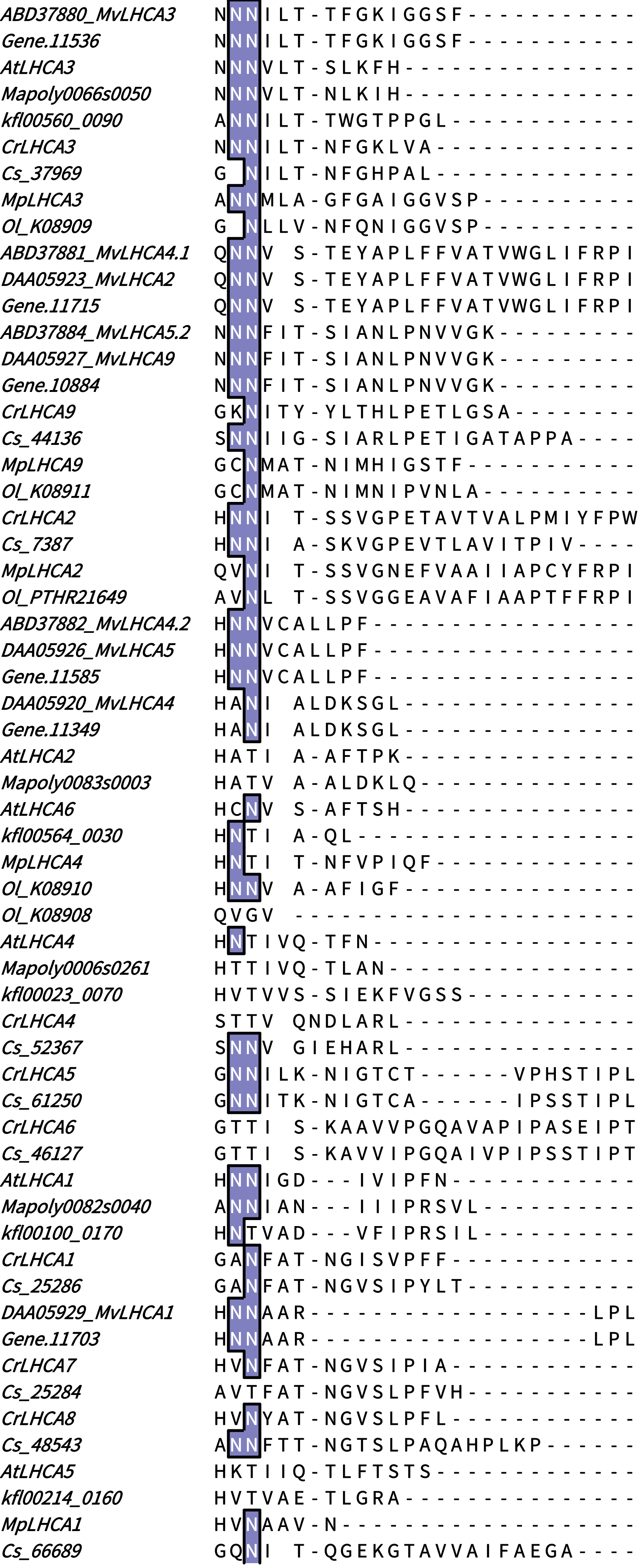
```
kfl00422_0020        PNV NAHS IA AAF A LMGGAEFARLKAPK-------------------EM GLYPGG
AtLHCB6              PDA APFSFGS  GT   LMGW ESKRWVDFFNPDSQSVEWATPWSKTFANTG QGYPGG
Mapoly0052s0039      PSA APFSFGT  GT   LMGW EGKRWADYVNPNSQLVDWATPWSRTFGNTGLQGYPGG
kfl00276_0010        PGA APFTFGT  GT   LMSW EGKRWVDFYNPSSQSVEWATPWSKTFANTGQQGYPGG
DAA05919_MvLHCP1.2   PAGSGFPNFYIQ A ST AMGLAEGYRGGLIDSC--------------FPETVGDL PGG
Gene.11156           PAGSGFPNFYIQ A ST AMGLAEGYRGGLIDSC--------------FPETVGDL PGG
DAA05930_MvLHCP1.1   PEGSGFPNFYIQ G S   MGLAEGYRGGLIDNV--------------FPEEVGDL PGG
Gene.11308           PEGSGFPNFYIQ G S   MGLAEGYRGGLIDNV--------------FPEEVGDL PGG
MpLHCAP2.1           PEGSGYPSFWA  A    L GLAEAYRTGLTDPA--------------FDETVGDVSPGG
MpLHCP1.1            PEGSGYPSFWA  A   FL GSAECYRTGLFENP--------------FPEEL-SVTPGG
CmLHC                ---SGAMQ LAF GFLEFL HRG VLYSDMEW---------------------KGRKPGE
PyLHC                ---QGAMQ LLW CGLE   GVPAVL M-MQG---------------------SGRRPGE
CrLHCSR1             --- GQG----FWEPL  A GVAESYRVAVGWTPTGTGFN----------SLK DYEPGD
CrLHCSR3             --- GQG----FWEPL  A GVAESYRVAVGWTPTGTGFN----------SLK DYEPGD
Cs_65904             --- EYK-GAIFWEPL FS GLAEAWRIGVGWNPSSDKFN----------QLR DYSPGE
kfl00478_0030        --- ENE-GAIFWEPL FA ALAESYRVGLGWTPDSTNFN----------TLR DYEPGN
DAA05932_MvLI818;2   --- PQP----FWE V  A GLAEGFRINRGWSPAEAFFSI--------GVLK SYTPGT
Gene.11245           --- PQP----FWE V  A GLAEGFRINRGWSPAEAYFSI--------GVLK SYTPGT
Gene.11340           --- PQP----FWE VC G AMAEAYRLQEGWNPRDGGYY----------KLRPGYQPGD
MpLI818;1            --- PLP----FWLGLGAL FI EASRVQIAWSPDASRLG----------LMK DHTPGD
MpLI818;2            --- PGT----FWG LGA  FT EANRVQTAWNPEADKLF----------LLK DYTPGD

                              250       260       270       280       290       300
                     ....|....|....|....|....|....|....|....|....|....|....|....|
ABD37880_MvLHCA3     AIF FLGYGKDE-KSMN LK KE KNGRLAM AT GFF QA ATG GPF NLT HLADPF
Gene.11536           AIF FLGYGKDE-KSMN LK KE KNGRLAM AT GFF QA ATG GPF NLT HLADPF
AtLHCA3              PFF PLGFGKDE-KSLKELK KE KNGRLAM AI G F QG VTGVGPY NLL HLADPV
Mapoly0066s0050      PLF FLGFGRDE-KSMK LK KE KNGRLAM AV G F QA FTGVGPF NLL HL DPA
kfl00560_0090        PFF FAGFGKDE-KSKK L DKE  NGRLAM AV G FVQG STG GPF NLL HL SPT
CrLHCA3              PFF LFNLGKTE-AAMKELK KEIKNGRLAM AM G GAQA  TG GPF NLV HLADPV
Cs_37969             QFF LFNLGKSD---LKELKTKE KNGRLAM AVFG GAQA  TA GPF NLT HL DPT
MpLHCA3              KYF FANLGKTD---MA   KKEIKNGRLAM AF GIAVQA ATG GPF N V H  DPF
Ol_K08909            -FF I GQGKSD---MA    KEIKNGRLAM ACFACGAQA  TG GP  NLI H  DPF
ABD37881_MvLHCA4.1   -FLDI G DS------P    KEIKNGRLAMFAF GFVAQYFVNGMGP  GLL HLE PQ
DAA05923_MvLHCA2     -FLDI G DS------P    KEIKNGRLAMFAF GFVAQYFVNGMGP  GLL HLE PQ
Gene.11715           -FLDI G DS------P    KEIKNGRLAMFAF GFVAQYFVNGMGP  GLL HLE PQ
ABD37884_MvLHCA5.2   --FDPLG  ------NSKE M KE KNGRLAMFAWCGFMVQA VTR GP TNLEKH ADPF
DAA05927_MvLHCA9     --FDPLG  ------NSKE M KE KNGRLAMFAWCGFMVQA VTR GP TNLEKH ADPF
Gene.10884           --FDPLG  ------NSKE M KE KNGRLAMFAWCGFMVQA VTR GP TNLEKH ADPF
CrLHCA9              --FDPAG  ------NSPS ATKE KNGRLAM AF GFCVQA ATRTQP  GLTAHLADPF
Cs_44136             --FDPANLDS------EANAEKEIKNGRLAM AF GFVGAA VCR GP  ALQSHL DPF
MpLHCA9              -- DP G  ------NNDA A KEIKNGRLAM AF GVVVQA VYRTGP AALK H VDPF
Ol_K08911            --FDP G GS----KDE  K KEIKNGRAAM AF GIVVQG VYR GP AALK H A PF
CrLHCA2              --FDP G KS------EE K KE KNGRLAM AF GFCSQAAVYG GP  TLQLHLADPG
Cs_7387              --FDPLN RDDY------K QSE  NGRLAM AF GFASQAANTG GP  NLK H ADPT
MpLHCA2              --FDPSGQ------DSPA K KEIKNGRLAM  F GMVSQYAVTGTSP  GLKAH A PT
Ol_PTHR21649         --FDPAGL------DAP K   KE KNARLAM AF GMVSQYAVTGLSP  GLEAH A PQ
ABD37882_MvLHCA4.2   -MFDPLGF KG---NLEESKWKE  NGRLAM AS GMF QYDATG SP TNLA HLADPW
DAA05926_MvLHCA5     -MFDPLGF KG---NLEESKWKE  NGRLAM AS GMF QYDATG SP TNLA HLADPW
Gene.11585           -MFDPLGF KG---NLEESKWKE  NGRLAM AS GMF QYDATG SP TNLA HLADPW
DAA05920_MvLHCA4     -PFDPLGF KG---NLPELK KEIKNGRLAM AS GLFVQYSATGASP DNLAAHLADPG
Gene.11349           -PFDPLGF KG---NLPELK KEIKNGRLAM AS GLFVQYSATGASP DNLAAHLADPG
AtLHCA2              -LFDPLGWGSGSPAKLKEL TKEIKNGRLAM AV GAWFQH YTGTGP DNLFAHLADPG
Mapoly0083s0003      -FFDPLGWGAGGAAKVKEL TKEIKNGRLAM AV GAWFQA YTGTGP DNLFAHLADPG
AtLHCA6              -LFDF MWGRGSPEPVMVL TKEIKNGRLAM AF GFCFQATYTS DP  NLMAHLADPG
kfl00564_0030        -IFDPFGYAKDS-SKAGELK KEIKNARLAM AMAGFFFQA YTGVGP DNLTTHLADPG
MpLHCA4              -LFDP G AKDP-EKVKVLK KEIANGRLAM AV GCI QNEATGVGP ANLKAHLADPA
Ol_K08910            -IFDPLGYAKG---DLKTLKAKEIANGRLAM AFAGIMVQYDHTGVGP ANLVSH ADPA
Ol_K08908            -LFDKFGWAKDE-KTTAELK KEIKNGRLAM AF GICAQY QTGVGP  NLFSH G PG
```



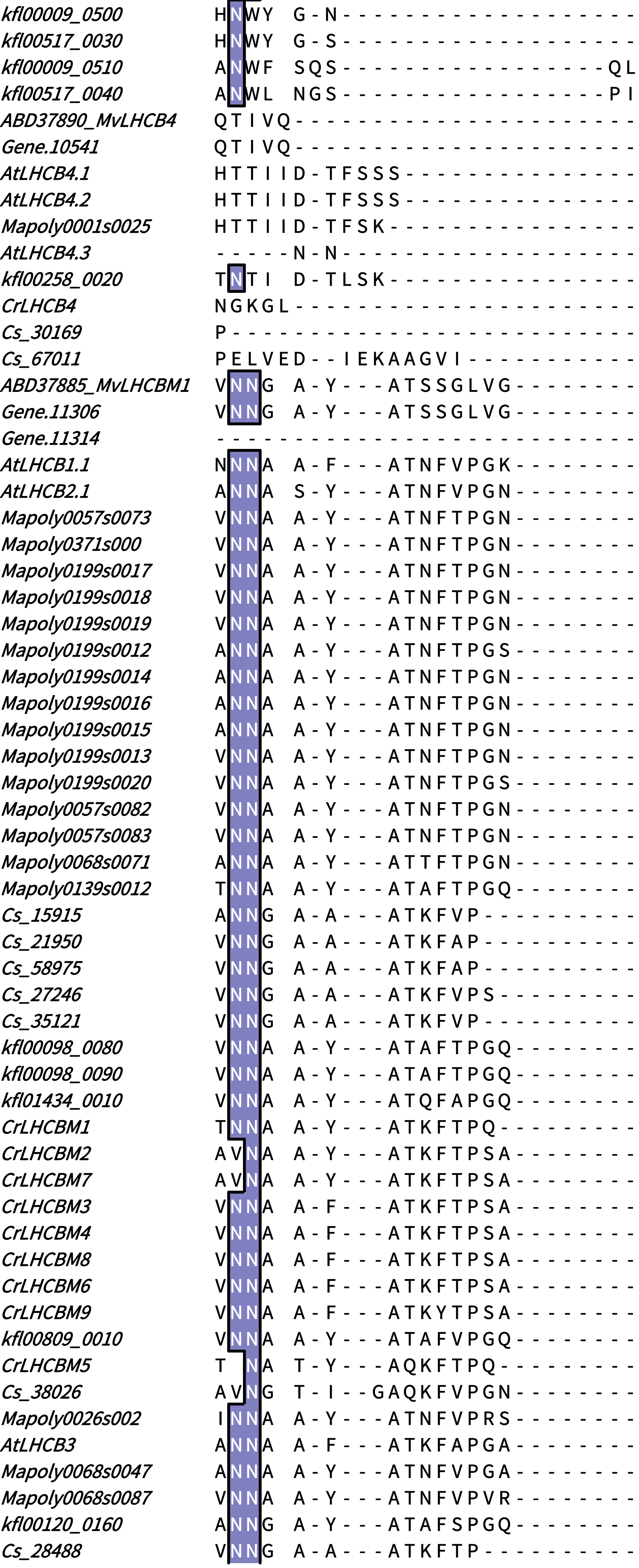
```
AtLHCA4             -IF PLNFAPT-----QEAKEKE ANGRLAM AF GFVVQHNVTG GPF NLLQHL DPW
Mapoly0006s0261     -IF PLKFPA -----QEYKEKEIANGRLAM AF GMLVQ K TGAGPF NLLTHLADPW
kfl00023_0070       -IF PLNFPV -----KESKDKE ANGRLAM AF G IAQ HVTG GPFANLLAHLADPW
CrLHCA4             -IFDPFGW KG---DIKSLK KEIKNGRLAM AFAGFIGQAYTTGTTP  NLSTHLADPW
Cs_52367            -IFDPLGY KG---NMAELK KEIKNARLAM AV GFFVQAKTTG TP D LSSHLADPW
CrLHCA5             -IFDPFGF KG---NLKEL TKEIKNGRLAM AY AFI QAQATG GP AALSAHL  PF
Cs_61250            -IFDPFGF KG---NFKE  TKEIKNGRLAM AF GFI QAQATG GP ANLGAHLA PF
CrLHCA6             -VFAPF----IP-GDLAELK KEIKNGRLAM AF GFV AAQVTG GP AALQ HLADPW
Cs_46127            -IFAP ----IP-GDLAELK KEIKNGRLAM AFAGFV QAQVTG GP A LG HL  PI
AtLHCA1             -AFDPLGY KDP-KKLEELK KEIKNGRLA  AF GFCVQQSAPGTGP  NLATHLADPW
Mapoly0082s0040     -PFDPLGFAKDP-KKLEEYK KE KNGRLA  AF GFSVQA APGTGP  NLASHL  PW
kfl00100_0170       -AFDPLGF KDP-AKLDELK KE KNGRLA  AF GFAVQAAVPGTGP  NLAQHL DPW
CrLHCA1             -AFDPLGFAKDS-SKSGELK KEIKNGRLAM AF GFVAQHAATG GP AALG HLA PW
Cs_25286            -AFDPLGF KDS-KVLEENK KE KNGRLAM AF GFIAQHAATG GP  ALKLHLADPW
DAA05929_MvLHCA1    -AFDPLGFGKDE-----VNK KE  NGRLAM AF GMWCQG HVGGGP SNLM HLADPL
Gene.11703          -AFDPLGFGKDE-----VNK KE  NGRLAM AF GMWCQG HVGGGP SNLM HLADPL
CrLHCA7             RFFDP GL RGDAAKYQEYKQKE KNGRLAM AC GFAAQYAATG GP DNLA HLADPN
Cs_25284            KLFDPFGL RGSEGQLQ Y ENEIKNGRLAM AF GFIAQHAATG GP DNLA HLA PT
CrLHCA8             -PFDPLGL K A-DKWA WK KE KNGRLAM AF GFVAQKYATGAGP DNLAAHLKDPW
Cs_48543            -VFDPLGL KGP--GYEA K  EIKNGRLAM AF GFASQY ATG GP  NLL HL  PL
AtLHCA5             PLL PLGLAKDV-QNAH WK KEIKNGRLAM AM GFFVQASVTHTGP DNLV HL  PW
kfl00214_0160       PFF ITGLGTGIKDDVF   TKE KNGRLAM AV GFIVQA VTKTGP  NLV HLADPF
MpLHCA1             -AFDPAGF KGA--DFETLKKKEIANGR AM AFAGLIGQAQATN GP SALGRH ADPW
Cs_66689            APFDP GLTTDY------N QAE  NCRLAM AA GFATQAYVTG GP  NAV HLRDPF
kfl00009_0500       --FDPLGLATGDDAQIK   T EIKNGRLAM ATGGFLHQA HTG GP ANLM HLADPS
kfl00517_0030       --FDPLGLTTGEADEIKIL EKEIKNGRLAM AAGGFLHQA HTG GPYANLI HLADPS
kfl00009_0510       --FDPLGFAKGDDAKVK   T EIKNGRLAM AF GFLQQY VTG GP TNLSQHLA PH
kfl00517_0040       --FDPLGFLEDKEEEVNEL TKE KNGRLAM  F GFLWQY ETG GP TNHTKHLADPS
ABD37890_MvLHCB4    -PFDPFGFTKASESRVN LK AEIK ARLGM AS GFAVQYAYTGAGP  NLA VLIGRG
Gene.10541          -PFDPFGFTKASESRVN LK AEIK ARLGM AS GFAVQYAYTGAGP  NLA VLIGRG
AtLHCB4.1           KFFDPLGLAADP-EKTA L  AEIK ARLAM AF GFAVQAAATG GP NNWATHL DPL
AtLHCB4.2           KFFDPLGLASDP-VKKA L  AEIK ARLAM GF GFAVQAAATG GP NNWATHL DPL
Mapoly0001s0025     KYFDPLGLASDP-EKKEILK AEIK ARLAM GA GFAVQAAATG GP DNWATHL DPL
AtLHCB4.3           -YFDPLGLAADP-EKLDTLK AEIK SRLAM AF IFA QAAFTG GP SFLATF-----
kfl00258_0020       -YFDPLNLASSP-EKAA LK AE K GRLAM  A GFAVQAFATGTGP DNLAQHLDSPT
CrLHCB4             -VFDPLKLAS DEERAFRLKTAEIK ARLAM  FFG GVQA STG GA G LAKFADGLN
Cs_30169            -LFDPLKLASDDSQRTF L EAEIK GRLAM AF GAQ Q RQ ---------L ERSQW
Cs_67011            -YFDPLGLATGDDARAF LKEAE K GRLAM AF GFGVQA ATG GA G LAKFA SFA
ABD37885_MvLHCBM1   S-FDPLGLGDDP-DTLAELK KEIKNGRLAMF MFG F QA VTG GP  NLDAHLADPF
Gene.11306          S-FDPLGLGDDP-DTLAELK KEIKNGRLAMF MFG F QA VTG GP  NLDAHLADPF
Gene.11314          A--------DGP-G----------------------------------------------
AtLHCB1.1           S-FDPLGLATDP-EAFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLA HLADPV
AtLHCB2.1           A-FDPLNLAEDP-EAFSELK KE KNGRLAMF MFGFFVQA VTG GP  NLF HLADPV
Mapoly0057s0073     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPA
Mapoly0371s000      Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPA
Mapoly0199s0017     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPT
Mapoly0199s0018     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPT
Mapoly0199s0019     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPT
Mapoly0199s0012     S-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPV
Mapoly0199s0014     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPV
Mapoly0199s0016     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPV
Mapoly0199s0015     S-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPV
Mapoly0199s0013     Q-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPA
Mapoly0199s0020     S-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPT
Mapoly0057s0082     S-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPT
Mapoly0057s0083     S-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPT
Mapoly0068s0071     S-FDPLNLAEDP-DTFAELK KE KNGRLAMF MFGFFVQA VTG GP  NLS HLADPV
Mapoly0139s0012     A-FDPLGLANDP-DTFAELK KEIKNGRLAMF AFGFFVQA VTG GP DNLLSH ADPA
Cs_15915            P-FDPLEFASDP-DTLAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLN HLASPY
```



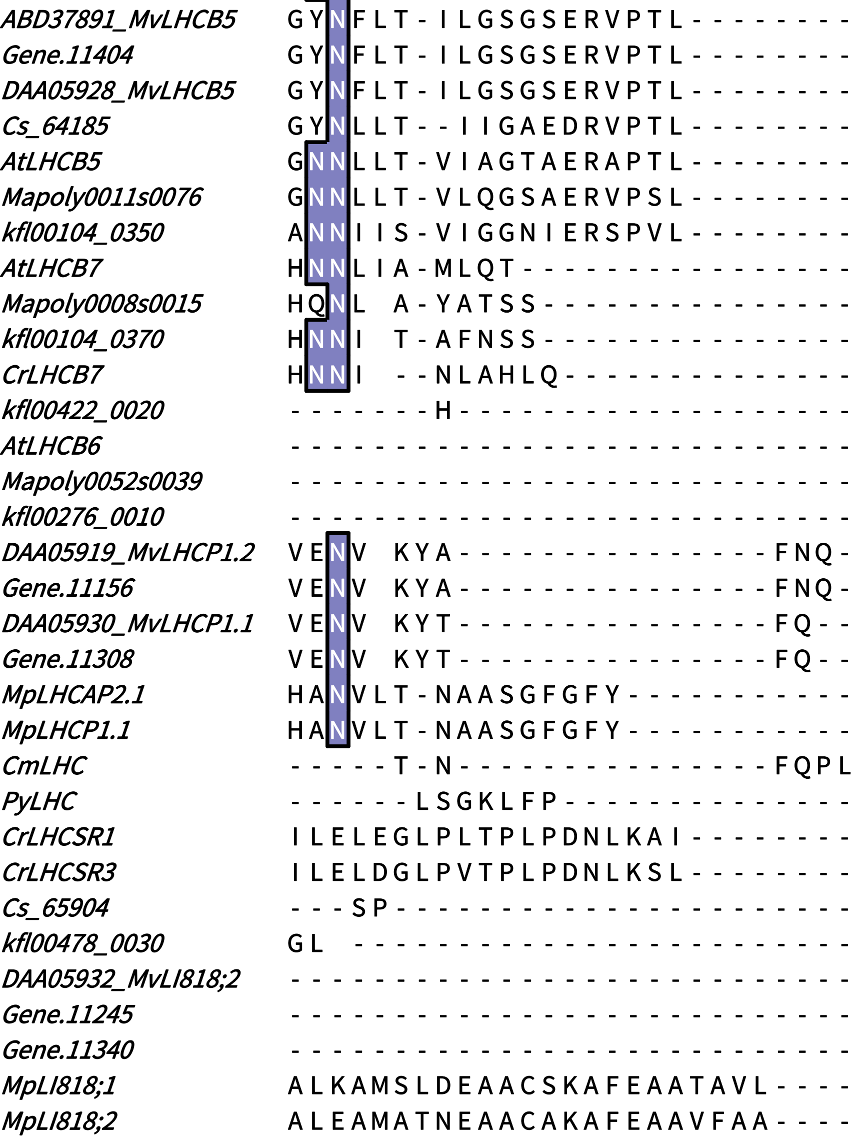
```
Cs_21950            P-FDPLGLADDP-DTFSELK KEIKNGRLAMF MFGFFVQA VTG GP  NLS HLADPG
Cs_58975            P-FDPLGLADDP-DTFAELKTKEIKNGRLAMF MFGFFVQA VTG GP  NLS HLADPG
Cs_27246            A-FDPLGLADDP-DTFAEL  KEIKNGRLAMF MFGFFVQA VTG GP  NLA HLA PS
Cs_35121            DYFDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLESHLADPS
kfl00098_0080       Q-FDPLNLAEDP-DTFAEL TKEIKNGRLAMF MFGFYVQA VTG GP  NLA HLD PT
kfl00098_0090       Q-FDPLNLAEDP-DTFAEL TKEIKNGRLAMF MFGFYVQA VTG GP  NLS HLD PS
kfl01434_0010       Q-FDPLNLAEDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP ANLN HLADPA
CrLHCBM1            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLS HLA PG
CrLHCBM2            A-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLD HLA PT
CrLHCBM7            A-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLD HLA PT
CrLHCBM3            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLD HL  PT
CrLHCBM4            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLD HLA PT
CrLHCBM8            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLD HLA PG
CrLHCBM6            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLD HLA PT
CrLHCBM9            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF SFGFFVQA VTG GP  NLD HL  PG
kfl00809_0010       Q-FDPLGLADDP-EEFAELK KEIKNGRLA  AN GFFVQA VTG GP  NLE HLA PA
CrLHCBM5            QFFDPLGLAEDP-DAFAELK KEIKNGRLAMF MFGFFVQA VTG GP ANLD HLASPF
Cs_38026            A-FDPLGLADDP-DTFAELK KEIKNGRLAQFAA GFFVQA VTG GP  NLE HLADPF
Mapoly0026s002      A-FDP GL EDP-EAFAELK KE KNGRLAMTAM GFFVQA VTG GP  NLY HLA PQ
AtLHCB3             QYFDPLGLADDP-VTFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLL HLD PV
Mapoly0068s0047     KYFDPLGLAEDP-ETFAELK KEIKNGRLAMF MFGFFVQA VTG GP  NLL HLD PV
Mapoly0068s0087     DYFDPLNLAQDP-DTLAELK KE KNGRLAMTAM GLFVQA VTS GP  NLA HLA PY
kfl00120_0160       A-FDPLGLADDP-EAAAELK KEIKNGRLAMF CFGFFVQA VTG GP  NLQ HLADPG
Cs_28488            S-FDPLGLADDP-DTFAELK KEIKNGRLAMF NFGFFVQA VTG GP ANLN HLA PT
ABD37891_MvLHCB5    KYFDPLGLASDP-SKADELK KEIKNGRLAM AF GFVVQYFVTK GPYANWSKHLADPF
Gene.11404          KYFDPLGLASDP-SKADELK KEIKNGRLAM AF GFVVQYFVTK GPYANWSKHLADPF
DAA05928_MvLHCB5    KYFDPLGLASDP-SKADELK KEIKNGRLAM AF GFVVQYFVTK GPYANWSKHLADPF
Cs_64185            P-FDPLGLADDP-ETFAELK KEIKNGRLA   V GFAVQ YVTG GPYANWSKH ADPF
AtLHCB5             P-FDPLGLAKDP-EQGALLK KEIKNGRLAMFAM GFF QAYVTG GP  NLAKHL DPF
Mapoly0011s0076     A-FDPLGLAKDP-DQFALLK KEIKNGRLAMF M GFF QAYVTG GP  NLAAHL DPF
kfl00104_0350       A-FDPLGLADDP-DQFALLK KEIKNGRLA F L GFFVQAYVTQDGP ANLAAHLDDPF
AtLHCB7             TLFDPLNL EDP-VAFE LK KEIKNGRLAM AW GFYAQAAFTG GP  NLV H  DPL
Mapoly0008s0015     ALFDPLGL KDP-STFE LK KEIKNGRLAM AW GFYAQA VTG GP  NAL FL DPA
kfl00104_0370       ALFDPLGLADDP-EAFE LK KE KNGRLAM AW GFY QA VTQ GP  NLL FA DPA
CrLHCB7             GPFDPLNYAADA-DGFVEQA KE KNGRLAM AM GFFVQAAVTRVGP  N L F ADPA
kfl00422_0020       KAFDPLGF TDP-ESLAELK KEIKNGRLAM AAAGFFAQGA TGVGP  NLK WL----
AtLHCB6             RFFDPLGLAGKNFEKLERLK AEIK SRLAM AM IFYF AG-QG TP GALGL------
Mapoly0052s0039     KFFDPLSLAGDVYDKLSRL  AEIK SRLAM AM IFYF AG-QGLTP GALGL------
kfl00276_0010       KFFDPFGLAGKVFAKLERL  AEIK ARLAM AM IFY  AG-QGLSP SPFGL------
DAA05919_MvLHCP1.2  EHFDPLGLAN--KLDLDR K AE K ARLAM  W G MSQA ATN GPYANWQ H A PW
Gene.11156          EHFDPLGLAN--KLDLDR K AE K ARLAM  W G MSQA ATN GPYANWQ H A PW
DAA05930_MvLHCP1.1  EHFDPLGLAD--KLDLDR K AE K ARLAM  W G MSQA ATN GPYANWL H A PY
Gene.11308          EHFDPLGLAD--KLDLDR K AE K ARLAM  W G MSQA ATN GPYANWL H A PY
MpLHCAP2.1          R-FDPLGLAESG--DLEELK KE K CRL MFAW GCISQG ATQ GP ANWQAH ADPV
MpLHCP1.1           R-FDPLGFAEAG--DLEELK KE K CRL MFAW GCIAQG ATQ GP ANWQAH ADPV
CmLHC               LGF PLNLPNDK-----A  D E NNGRLAM GFAGIIHGEF NG MPF Q --------
PyLHC               FGFDPLNLAKTN-AAFD F NAE VNGRLTM AIGGLIHQQW THCTP  Q --------
CrLHCSR1            LGFDPLGLKPDP-EELKTL TKE NNGRLAM AIAAFVAQE VEQTEIF  LVLRFEKEV
CrLHCSR3            LGFDPLGLKPDP-EELKV  TKE NNGRLAM AIAAFVAQE VEQTEIF  LALRFEKEA
Cs_65904            IGFDPLGLTPDP-KEKY L TKE NNGRLAM GIAGFVAQE VKK EIFYQ G T KGSS
kfl00478_0030       LGFDPLGLLPDP-AAKK   SKE NNGRLAM AIAGFVAQE V G GIF  LFKRL----
DAA05932_MvLI818;2  LG DPLGLKP S-KELDE ATKE NNGRLAM GIAGMVVQE  TQTPIFGQ---------
Gene.11245          LG DPLGLKP S-KELDE ATKE NNGRLAM GIAGMVVQE  TQTPIFGQ---------
Gene.11340          LG DPLGLKPDP-AALAEL TKE NNGRLAM GIAGMVVQEF TQ PIF-----------
MpLI818;1           YGFDPLGL KDE-EWLDIH   E NNGRLAM AITAMVAQE NTGINL LADVLG KGGG
MpLI818;2           YNFDPLKL ADD-VWLN  K KE NNGR AM AISGMVAQE VNGLNL PADVLE GKAG

                            310       320
                    ....|....|....|....|....|..
```



```
ABD37880_MvLHCA3     NNNILT-TFGKIGGSF-----------
Gene.11536           NNNILT-TFGKIGGSF-----------
AtLHCA3              NNNVLT-SLKFH---------------
Mapoly0066s0050      NNNVLT-NLKIH---------------
kfl00560_0090        ANNILT-TWGTPPGL------------
CrLHCA3              NNNILT-NFGKLVA-------------
Cs_37969             G NILT-NFGHPAL-------------
MpLHCA3              ANNMLA-GFGAIGGVSP----------
Ol_K08909            G NLLV-NFQNIGGVSP----------
ABD37881_MvLHCA4.1   QNNV S-TEYAPLFFVATVWGLIFRPI
DAA05923_MvLHCA2     QNNV S-TEYAPLFFVATVWGLIFRPI
Gene.11715           QNNV S-TEYAPLFFVATVWGLIFRPI
ABD37884_MvLHCA5.2   NNNFIT-SIANLPNVVGK---------
DAA05927_MvLHCA9     NNNFIT-SIANLPNVVGK---------
Gene.10884           NNNFIT-SIANLPNVVGK---------
CrLHCA9              GKNITY-YLTHLPETLGSA--------
Cs_44136             SNNIIG-SIARLPETIGATAPPA----
MpLHCA9              GCNMAT-NIMHIGSTF-----------
Ol_K08911            GCNMAT-NIMNIPVNLA----------
CrLHCA2              HNNI T-SSVGPETAVTVALPMIYFPW
Cs_7387              HNNI A-SKVGPEVTLAVITPIV----
MpLHCA2              QVNI T-SSVGNEFVAAIIAPCYFRPI
Ol_PTHR21649         AVNL T-SSVGGEAVAFIAAPTFFRPI
ABD37882_MvLHCA4.2   HNNVCALLPF-----------------
DAA05926_MvLHCA5     HNNVCALLPF-----------------
Gene.11585           HNNVCALLPF-----------------
DAA05920_MvLHCA4     HANI ALDKSGL---------------
Gene.11349           HANI ALDKSGL---------------
AtLHCA2              HATI A-AFTPK---------------
Mapoly0083s0003      HATV A-ALDKLQ--------------
AtLHCA6              HCNV S-AFTSH---------------
kfl00564_0030        HNTI A-QL------------------
MpLHCA4              HNTI T-NFVPIQF-------------
Ol_K08910            HNNV A-AFIGF---------------
Ol_K08908            QVGV ----------------------
AtLHCA4              HNTIVQ-TFN-----------------
Mapoly0006s0261      HTTIVQ-TLAN----------------
kfl00023_0070        HVTVVS-SIEKFVGSS-----------
CrLHCA4              STTV QNDLARL---------------
Cs_52367             SNNV GIEHARL---------------
CrLHCA5              GNNILK-NIGTCT------VPHSTIPL
Cs_61250             GNNITK-NIGTCA------IPSSTIPL
CrLHCA6              GTTI S-KAAVVPGQAVAPIPASEIPT
Cs_46127             GTTI S-KAVVIPGQAIVPIPSSTIPT
AtLHCA1              HNNIGD---IVIPFN------------
Mapoly0082s0040      ANNIAN---IIIPRSVL----------
kfl00100_0170        HNTVAD---VFIPRSIL----------
CrLHCA1              GANFAT-NGISVPFF------------
Cs_25286             GANFAT-NGVSIPYLT-----------
DAA05929_MvLHCA1     HNNAAR------------------LPL
Gene.11703           HNNAAR------------------LPL
CrLHCA7              HVNFAT-NGVSIPIA------------
Cs_25284             AVTFAT-NGVSLPFVH-----------
CrLHCA8              HVNYAT-NGVSLPFL------------
Cs_48543             ANNFTT-NGTSLPAQAHPLKP------
AtLHCA5              HKTIIQ-TLFTSTS-------------
kfl00214_0160        HVTVAE-TLGRA---------------
MpLHCA1              HVNAAV-N-------------------
Cs_66689             GQNI T-QGEKGTAVVAIFAEGA----
```



```
kfl00009_0500       HNWY G-N-------------------
kfl00517_0030       HNWY G-S-------------------
kfl00009_0510       ANWF SQS-----------------QL
kfl00517_0040       ANWL NGS-----------------PI
ABD37890_MvLHCB4    QTIVQ----------------------
Gene.10541          QTIVQ----------------------
AtLHCB4.1           HTTIID-TFSSS---------------
AtLHCB4.2           HTTIID-TFSSS---------------
Mapoly0001s0025     HTTIID-TFSK----------------
AtLHCB4.3           -----N-N-------------------
kfl00258_0020       TNTI D-TLSK----------------
CrLHCB4             NGKGL----------------------
Cs_30169            P--------------------------
Cs_67011            PELVED--IEKAAGVI-----------
ABD37885_MvLHCBM1   VNNG A-Y---ATSSGLVG--------
Gene.11306          VNNG A-Y---ATSSGLVG--------
Gene.11314          ---------------------------
AtLHCB1.1           NNNA A-F---ATNFVPGK--------
AtLHCB2.1           ANNA S-Y---ATNFVPGN--------
Mapoly0057s0073     VNNA A-Y---ATNFTPGN--------
Mapoly0371s000      VNNA A-Y---ATNFTPGN--------
Mapoly0199s0017     VNNA A-Y---ATNFTPGN--------
Mapoly0199s0018     VNNA A-Y---ATNFTPGN--------
Mapoly0199s0019     VNNA A-Y---ATNFTPGN--------
Mapoly0199s0012     ANNA A-Y---ATNFTPGS--------
Mapoly0199s0014     ANNA A-Y---ATNFTPGN--------
Mapoly0199s0016     ANNA A-Y---ATNFTPGN--------
Mapoly0199s0015     ANNA A-Y---ATNFTPGN--------
Mapoly0199s0013     VNNA A-Y---ATNFTPGN--------
Mapoly0199s0020     VNNA A-Y---ATNFTPGS--------
Mapoly0057s0082     VNNA A-Y---ATNFTPGN--------
Mapoly0057s0083     VNNA A-Y---ATNFTPGN--------
Mapoly0068s0071     ANNA A-Y---ATTFTPGN--------
Mapoly0139s0012     TNNA A-Y---ATAFTPGQ--------
Cs_15915            ANNG A-A---ATKFVP----------
Cs_21950            VNNG A-A---ATKFAP----------
Cs_58975            VNNG A-A---ATKFAP----------
Cs_27246            VNNG A-A---ATKFVPS---------
Cs_35121            VNNG A-A---ATKFVP----------
kfl00098_0080       VNNA A-Y---ATAFTPGQ--------
kfl00098_0090       VNNA A-Y---ATAFTPGQ--------
kfl01434_0010       VNNA A-Y---ATQFAPGQ--------
CrLHCBM1            TNNA A-Y---ATKFTPQ---------
CrLHCBM2            AVNA A-Y---ATKFTPSA--------
CrLHCBM7            AVNA A-Y---ATKFTPSA--------
CrLHCBM3            VNNA A-F---ATKFTPSA--------
CrLHCBM4            VNNA A-F---ATKFTPSA--------
CrLHCBM8            VNNA A-F---ATKFTPSA--------
CrLHCBM6            VNNA A-F---ATKFTPSA--------
CrLHCBM9            VNNA A-F---ATKYTPSA--------
kfl00809_0010       VNNA A-Y---ATAFVPGQ--------
CrLHCBM5            T NA T-Y---AQKFTPQ---------
Cs_38026            AVNG T-I--GAQKFVPGN--------
Mapoly0026s002      INNA A-Y---ATNFVPRS--------
AtLHCB3             ANNA A-F---ATKFAPGA--------
Mapoly0068s0047     ANNA A-Y---ATNFVPGA--------
Mapoly0068s0087     VNNA A-Y---ATNFVPVR--------
kfl00120_0160       ANNG A-Y---ATAFSPGQ--------
Cs_28488            VNNG A-A---ATKFTP----------
```
]



ABD37891_MvLHCB5	GYNFLT-ILGSGSERVPTL--------
Gene.11404	GYNFLT-ILGSGSERVPTL--------
DAA05928_MvLHCB5	GYNFLT-ILGSGSERVPTL--------
Cs_64185	GYNLLT--IIGAEDRVPTL--------
AtLHCB5	GNNLLT-VIAGTAERAPTL--------
Mapoly0011s0076	GNNLLT-VLQGSAERVPSL--------
kfl00104_0350	ANNIIS-VIGGNIERSPVL--------
AtLHCB7	HNNLIA-MLQT----------------
Mapoly0008s0015	HQNL A-YATSS---------------
kfl00104_0370	HNNI T-AFNSS---------------
CrLHCB7	HNNI --NLAHLQ--------------
kfl00422_0020	-------H-------------------
AtLHCB6	---------------------------
Mapoly0052s0039	---------------------------
kfl00276_0010	---------------------------
DAA05919_MvLHCP1.2	VENV KYA---------------FNQ-
Gene.11156	VENV KYA---------------FNQ-
DAA05930_MvLHCP1.1	VENV KYT---------------FQ--
Gene.11308	VENV KYT---------------FQ--
MpLHCAP2.1	HANVLT-NAASGFGFY-----------
MpLHCP1.1	HANVLT-NAASGFGFY-----------
CmLHC	-----T-N---------------FQPL
PyLHC	------LSGKLFP--------------
CrLHCSR1	ILELEGLPLTPLPDNLKAI--------
CrLHCSR3	ILELDGLPVTPLPDNLKSL--------
Cs_65904	---SP----------------------
kfl00478_0030	GL ------------------------
DAA05932_MvLI818;2	---------------------------
Gene.11245	---------------------------
Gene.11340	---------------------------
MpLI818;1	ALKAMSLDEAACSKAFEAATAVL----
MpLI818;2	ALEAMATNEAACAKAFEAAVFAA----

Fig. S2

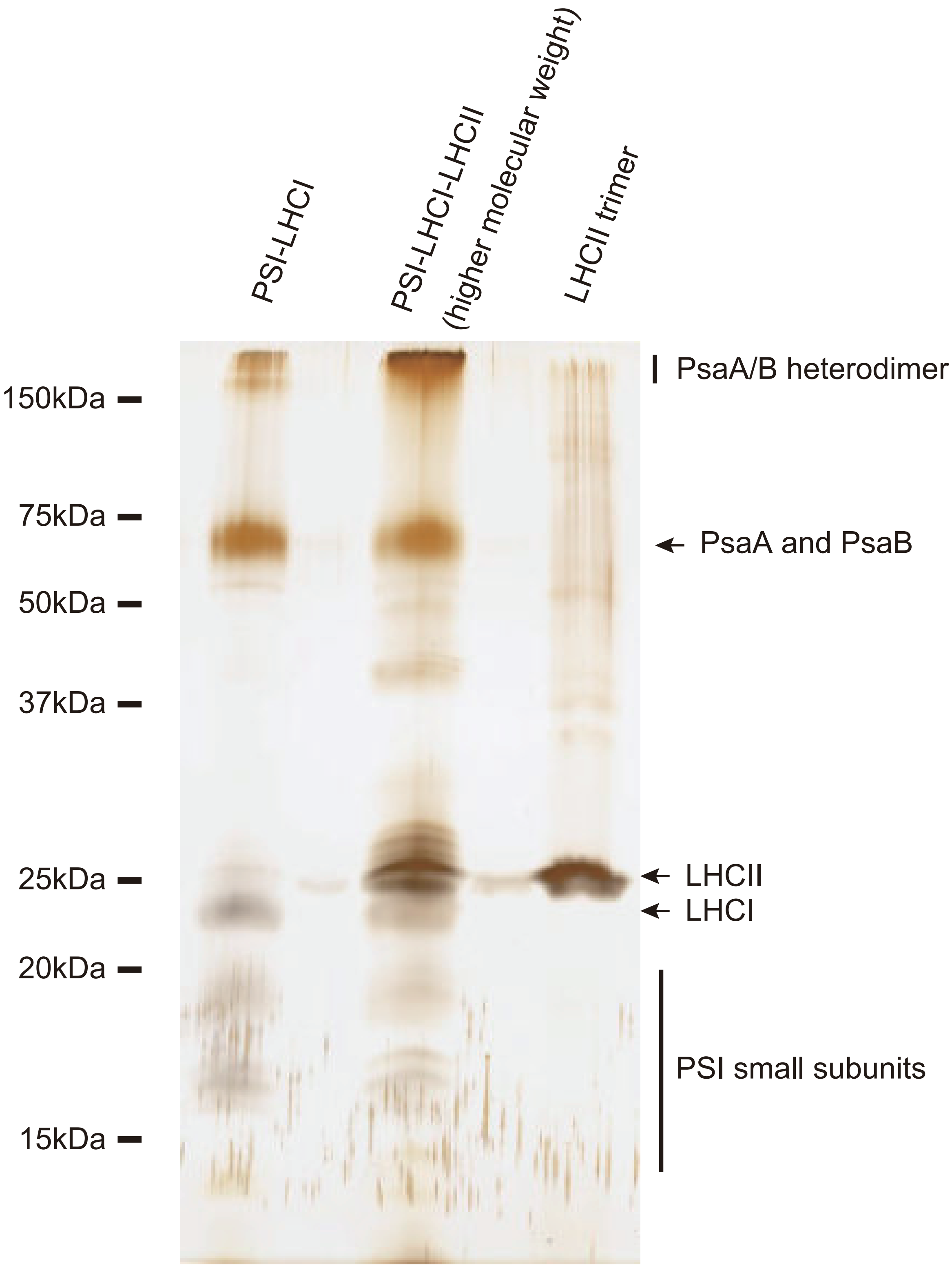


Figure S2. 2D-SDS-PAGE of the high-molecular-weight PSI-LHCI-LHCII, PSI-LHCI, and LHCII trimer.

The high-molecular-weight PSI-LHCI-LHCII, PSI-LHCI, and LHCII trimer bands separated by CN-PAGE after sucrose density gradient centrifugation (Fig. 7) were subjected to 2D-SDS-PAGE. Protein bands were visualized by silver-staining.

Aso et al.

Table S1-1. The list of the PSI proteins predicted by the Iso-seq analysis in this study

	Gene annotation in this study	Accesson No.	The best-hit gene in *A.thliana*			The best-hit gene in *C. reinhardtii*		
Gene3244	PSAA	ICQU01000001	ATCG00350.1	PSAA	Photosystem I, PsaA/PsaB protein	NP_958375.1	PSAA	Photosystem I P700 chlorophyll a apoprotein A1
Gene3504	PSAB	ICQU01000002	ATCG00340.1	PSAB	Photosystem I, PsaA/PsaB protein	NP_958404.1	PSAB	Photosystem I P700 chlorophyll a apoprotein A2
Gene9022	PSAD	ICQU01000006	AT1G03130.1	PSAD-2	photosystem I subunit D-2	Cre05.g238332.t1.1	PSAD	Photosystem I reaction center subunit II, 20 kDa
Gene11867	PSAE	ICQU01000029	AT2G20260.1	PSAE-2	photosystem I subunit E-2	Cre10.g420350.t1.2	PSAE	Photosystem I 8.1 kDa reaction center subunit IV
Gene11501	PSAF	ICQU01000021	AT1G31330.1	PSAF	photosystem I subunit F	Cre09.g412100.t1.2	PSAF	Photosystem I reaction center subunit III
Gene11916	PSAG	ICQU01000031	AT1G55670.1	PSAG	photosystem I subunit G	Cre12.g560950.t1.2	PSAG	Photosystem I reaction center subunit V
Gene11865	PSAH	ICQU01000028	AT1G52230.1	PSAH-2	photosystem I subunit H2	Cre07.g330250.t1.2	PSAH	Subunit H of photosystem I
Gene11650	PSAL	ICQU01000024	AT4G12800.1	PSAL	photosystem I subunit l	Cre12.g486300.t1.2	PSAL	Photosystem I reaction center subunit XI
Gene11942	PSAO	ICQU01000034	AT1G08380.1	PSAO	photosystem I subunit O	Cre07.g334550.t1.2	PSAO1	Photosystem I subunit O

Table S1-2. The list of the PSII proteins predicted by the Iso-seq analysis in this study

	Gene annotation in this study	Accesson No.	**The best-hit gene in *A.thliana***			**The best-hit gene in *C. reinhardtii***		
Gene10292	PSBA	ICQU01000008	ATCG00020.1	PSBA	photosystem II reaction center protein A	NP_958377.1	PSBA	photosystem II protein D1
Gene4825	PSBB	ICQU01000003	ATCG00680.1	PSBB	photosystem II reaction center protein B	NP_958388.1	PSBB	photosystem II 47 kDa protein
Gene6781	PSBC	ICQU01000005	ATCG00280.1	PSBC	photosystem II reaction center protein C	NP_958422.1	PSBC	photosystem II 44 kDa protein
Gene4841	PSBD	ICQU01000004	ATCG00270.1	PSBD	photosystem II reaction center protein D	NP_958420.1	PSBD	photosystem II protein D2
Gene10723	PSBO	ICQU01000010	AT3G50820.1	PSBO-2	photosystem II subunit O-2	Cre09.g396213.t1.1	PSBO	Oxygen-evolving enhancer protein 1 of photosystem II
Gene11499	PSBP	ICQU01000020	AT1G06680.1	PSBP-1	photosystem II subunit P-1	Cre12.g550850.t1.2	PSBP1	Oxygen-evolving enhancer protein 2 of photosystem II
Gene11808	PSBQ	ICQU01000027	AT4G05180.1	PSBQ-2	photosystem II subunit Q-2	Cre08.g372450.t1.2	PSBQ	Oxygen evolving enhancer protein 3
Gene11938	PSBR	ICQU01000033	AT1G79040.1	PSBR	photosystem II subunit R	Cre06.g261000.t1.2	PSBR	10 kDa photosystem II polypeptide
Gene11934	PSBW	ICQU01000032	AT2G30570.1	PSBW	photosystem II reaction center W			
Gene11880	PSBX	ICQU01000030	AT2G06520.1	PSBX	photosystem II subunit X			
Gene9164	PSBY	ICQU01000007	AT1G67740.1	PSBY	photosystem II BY	Cre10.g452100.t1.1	PSBY	Ycf32-related polyprotein of photosystem II

Table S1-3. The list of the LHC proteins predicted by the Iso-seq analysis in this study

	Gene annotation in this study	Accesson No.	**The best-hit gene in *A.thliana***			**The best-hit gene in *C. reinhardtii***		
Gene11245	LHCSR	ICQU01000013	AT1G15820.1	LHCB6	light harvesting complex photosystem II subunit 6	Cre08.g367500.t1.1	LHCSR3.1	Stress-related chlorophyll a/b binding protein 2
Gene11340	LHCSR	ICQU01000017	AT3G61470.1	LHCA2	photosystem I light harvesting complex gene 2	Cre08.g367500.t1.1	LHCSR3.1	Stress-related chlorophyll a/b binding protein 2
Gene11306	LHCBM	ICQU01000014	AT2G05100.1	LHCB2.1	photosystem II light harvesting complex gene 2.1	Cre01.g066917.t1.1	LHCBM1	Chlorophyll a/b binding protein of LHCII
Gene11314	LHCBM	ICQU01000016	AT2G05070.1	LHCB2.2	photosystem II light harvesting complex gene 2.2	Cre01.g066917.t1.1	LHCBM1	Chlorophyll a/b binding protein of LHCII
Gene10541	LHCB4	ICQU01000009	AT2G40100.1	LHCB4.3	light harvesting complex photosystem II	Cre17.g720250.t1.2	LHCB4	Chlorophyll a/b binding protein of photosystem II
Gene11404	LHCB5	ICQU01000019	AT4G10340.1	LHCB5	light harvesting complex of photosystem II 5	Cre16.g673650.t1.1	LHCB5	Minor chlorophyll a/b binding protein of photosystem II
Gene11156	LHCP	ICQU01000012	AT5G54270.1	LHCB3	light-harvesting chlorophyll B-binding protein 3	Cre03.g156900.t1.2	LHCBM5	Chlorophyll a/b binding protein of LHCII
Gene11308	LHCP	ICQU01000015	AT5G54270.1	LHCB3	light-harvesting chlorophyll B-binding protein 3	Cre04.g232104.t1.1	LHCBM3	Light-harvesting complex II chlorophyll a/b binding protein M3
Gene11703	LHCA1	ICQU01000025	AT3G54890.1	LHCA1	photosystem I light harvesting complex gene 1	Cre06.g283050.t1.2	LHCA1	Light-harvesting protein of photosystem I
Gene11349	LHCA2	ICQU01000018	AT3G61470.1	LHCA2	photosystem I light harvesting complex gene 2	Cre16.g687900.t1.2	LHCA7	Light-harvesting protein of photosystem I
Gene11585	LHCA2	ICQU01000023	AT3G47470.1	LHCA4	light-harvesting chlorophyll-protein complex I subunit A4	Cre10.g452050.t1.2	LHCA4	Light-harvesting protein of photosystem I
Gene11536	LHCA3	ICQU01000022	AT1G61520.1	LHCA3	photosystem I light harvesting complex gene 3	Cre11.g467573.t1.1	LHCA3	Chlorophyll a/b binding protein of photosystem I, type III
Gene11715	algae-type LHCA2	ICQU01000026	AT1G45474.1	LHCA5	photosystem I light harvesting complex gene 5	Cre12.g508750.t1.2	LHCA2	Light-harvesting protein of photosystem I
Gene10884	LHCA9	ICQU01000011	AT1G45474.1	LHCA5	photosystem I light harvesting complex gene 5	Cre07.g344950.t1.2	LHCA9	Light-harvesting protein of photosystem I

Table S2. The identified proteins in the PSI-PSII band by MS

					Best-Hit Gene in Arabidopsis		Best-Hit Gene in Chlamydomonas	
Gene ID	NSAF (fmol)	Category	Gene annotation in this study	Accesson No.	Gene ID	Gene Name	Gene ID	Gene Name
Gene11306	4526	LHC	LHCBM	ICQU01000014	AT2G05100.1	LHCB2.1	Cre01.g066917.t1.1	LHCBM1
Gene10292	4034	PSII	PSBA	ICQU01000008	ATCG00020.1	PSBA	NP_958377.1	PSBA
Gene11501	3411	PSI	PSAF	ICQU01000021	AT1G31330.1	PSAF	Cre09.g412100.t1.2	PSAF
Gene11156	3322	LHC	LHCP	ICQU01000012	AT5G54270.1	LHCB3	Cre03.g156900.t1.2	LHCBM5
Gene11585	2942	LHC	LHCA2	ICQU01000023	AT3G47470.1	LHCA4	Cre10.g452050.t1.2	LHCA4
Gene4825	2883	PSII	PSBB	ICQU01000003	ATCG00680.1	PSBB	NP_958388.1	PSBB
Gene11867	2773	PSI	PSAE	ICQU01000029	AT2G20260.1	PSAE-2	Cre10.g420350.t1.2	PSAE
Gene11865	2654	PSI	PSAH	ICQU01000028	AT1G52230.1	PSAH-2	Cre07.g330250.t1.2	PSAH
Gene4841	2605	PSII	PSBD	ICQU01000004	ATCG00270.1	PSBD	NP_958420.1	PSBD
Gene3547	2468	Others			AT5G59970.1		Cre12.g506350.t1.2	HFO18
Gene11916	2444	PSI	PSAG	ICQU01000031	AT1G55670.1	PSAG	Cre12.g560950.t1.2	PSAG
Gene9022	2440	PSI	PSAD	ICQU01000006	AT1G03130.1	PSAD-2	Cre05.g238332.t1.1	PSAD
Gene11404	2412	LHC	LHCB5	ICQU01000019	AT4G10340.1	LHCB5	Cre16.g673650.t1.1	LHCB5
Gene10541	2301	LHC	LHCB4	ICQU01000009	AT2G40100.1	LHCB4.3	Cre17.g720250.t1.2	LHCB4
Gene11715	2235	LHC	algae-type LHCA2	ICQU01000026	AT1G45474.1	LHCA5	Cre12.g508750.t1.2	LHCA2
Gene6781	1782	PSII	PSBC	ICQU01000005	ATCG00280.1	PSBC	NP_958422.1	PSBC
Gene11340	1707	LHC	LHCSR	ICQU01000017	AT3G61470.1	LHCA2	Cre08.g367500.t1.1	LHCSR3.1
Gene6617	1577	Others			AT3G61320.1		Cre06.g261750.t1.2	
Gene3504	1569	PSII	PSAB	ICQU01000002	ATCG00340.1	PSAB	NP_958404.1	PSAB
Gene7944	1382	Others			AT3G61320.1		Cre06.g261750.t1.2	
Gene11703	1355	LHC	LHCA1	ICQU01000025	AT3G54890.1	LHCA1	Cre06.g283050.t1.2	LHCA1
Gene10723	1232	PSII	PSBO	ICQU01000010	AT3G50820.1	PSBO-2	Cre09.g396213.t1.1	PSBO
Gene11536	1228	LHC	LHCA3	ICQU01000022	AT1G61520.1	LHCA3	Cre11.g467573.t1.1	LHCA3
Gene10884	1219	LHC	LHCA9	ICQU01000011	AT1G45474.1	LHCA5	Cre07.g344950.t1.2	LHCA9
Gene7515	1169	Others			AT5G04180.1	ATACA3	Cre09.g415700.t1.2	CAH3
Gene11308	1066	LHC	LHCP	ICQU01000015	AT5G54270.1	LHCB3	Cre04.g232104.t1.1	LHCBM3
Gene7423	858	Others						
Gene8340	839	Others			AT5G17170.1	ENH1	Cre12.g510400.t1.1	CPLD30
Gene9362	780	Others			AT4G38970.1	FBA2	Cre05.g234550.t1.2	FBA3
Gene8224	529	Others			AT5G23860.1	TUB8	Cre12.g549550.t1.2	TUB2
Gene8089	512	Others			ATCG00490.1	RBCL	NP_958405.1	RBCL

Table S3. The identified proteins in the PSII-LHCII band by MS

					Best-Hit Gene in Arabidopsis		Best-Hit Gene in Chlamydomonas	
Gene ID	NSAF(fmol)	Category	Gene annotation in this study	Accesson No.	Gene ID	Gene Name	Gene ID	Gene Name
Gene11306	7753	LHC	LHCBM	ICQU01000014	AT2G05100.1	LHCB2.1	Cre01.g066917.t1.1	LHCBM1
Gene10292	6458	PSII	PSBA	ICQU01000008	ATCG00020.1	PSBA	NP_958377.1	PSBA
Gene10541	3371	LHC	LHCB4	ICQU01000009	AT2G40100.1	LHCB4.3	Cre17.g720250.t1.2	LHCB4
Gene4825	3021	PSII	PSBB	ICQU01000003	ATCG00680.1	PSBB	NP_958388.1	PSBB
Gene11404	1991	LHC	LHCB5	ICQU01000019	AT4G10340.1	LHCB5	Cre16.g673650.t1.1	LHCB5
Gene11865	1757	PSI	PSAH	ICQU01000028	AT1G52230.1	PSAH-2	Cre07.g330250.t1.2	PSAH
Gene6781	1495	PSII	PSBC	ICQU01000005	ATCG00280.1	PSBC	NP_958422.1	PSBC
Gene11938	1341	PSII	PSBR	ICQU01000033	AT1G79040.1	PSBR	Cre06.g261000.t1.2	PSBR
Gene3547	1317	Others			AT5G59970.1		Cre12.g506350.t1.2	HFO18
Gene11585	1159	LHCI	LHCA2	ICQU01000023	AT3G47470.1	LHCA4	Cre10.g452050.t1.2	LHCA4
Gene11867	1142	PSI	PSAE	ICQU01000029	AT2G20260.1	PSAE-2	Cre10.g420350.t1.2	PSAE
Gene11916	1105	PSI	PSAG	ICQU01000031	AT1G55670.1	PSAG	Cre12.g560950.t1.2	PSAG
Gene4841	1095	PSII	PSBD	ICQU01000004	ATCG00270.1	PSBD	NP_958420.1	PSBD
Gene7779	1091	Others			ATCG00430.1	PSBG	Cre12.g492300.t1.2	NUO10
Gene11308	1084	LHC	LHCP	ICQU01000015	AT5G54270.1	LHCB3	Cre04.g232104.t1.1	LHCBM3
Gene9022	1080	PSI	PSAD	ICQU01000006	AT1G03130.1	PSAD-2	Cre05.g238332.t1.1	PSAD
Gene11715	976	LHC	algae-type LHCA2	ICQU01000026	AT1G45474.1	LHCA5	Cre12.g508750.t1.2	LHCA2
Gene10884	937	LHC	LHCA9	ICQU01000011	AT1G45474.1	LHCA5	Cre07.g344950.t1.2	LHCA9
Gene11501	926	PSI	PSAF	ICQU01000021	AT1G31330.1	PSAF	Cre09.g412100.t1.2	PSAF
Gene11349	902	LHC	LHCA2	ICQU01000018	AT3G61470.1	LHCA2	Cre16.g687900.t1.2	LHCA7
Gene11703	858	LHC	LHCA1	ICQU01000025	AT3G54890.1	LHCA1	Cre06.g283050.t1.2	LHCA1
Gene7515	822	Others			AT5G04180.1	ATACA3	Cre09.g415700.t1.2	CAH3
Gene11156	789	LHC	LHCP	ICQU01000012	AT5G54270.1	LHCB3	Cre03.g156900.t1.2	LHCBM5
Gene2597	781	Others			ATCG01100.1	NDHA		
Gene11934	741	PSII	PSBW	ICQU01000032	AT2G30570.1	PSBW		
Gene4851	714	Others			ATCG01110.1	NDHH	Cre09.g405850.t1.1	NUO7
Gene11718	690	Others			ATCG00420.1	NDHJ	Cre07.g327400.t1.1	NUO9
Gene9021	683	Others			ATCG01070.1	NDHE	Cre09.g402552.t1.1	NUO11
Gene11761	676	Others						
Gene6617	618	Others			AT3G61320.1		Cre06.g261750.t1.2	
Gene11650	605	PSI	PSAL	ICQU01000024	AT4G12800.1	PSAL	Cre12.g486300.t1.2	PSAL
Gene8813	572	Others			AT1G20020.1	ATLFNR2	Cre11.g476750.t1.2	FNR1
Gene11139	565	Others			AT2G28720.1		Cre13.g590750.t1.2	HTB21
Gene10723	565	PSII	PSBO	ICQU01000010	AT3G50820.1	PSBO-2	Cre09.g396213.t1.1	PSBO
Gene11808	539	PSII	PSBQ	ICQU01000027	AT4G05180.1	PSBQ-2	Cre08.g372450.t1.2	PSBQ
Gene8340	531	Others			AT5G17170.1	ENH1	Cre12.g510400.t1.1	CPLD30
Gene10856	527	Others			AT5G58260.1			
Gene11340	526	LHC	LHCSR	ICQU01000017	AT3G61470.1	LHCA2	Cre08.g367500.t1.1	LHCSR3.1
Gene5104	522	Others			AT1G15980.1	NDF1		

Table S4. The identified proteins in the PSI-LHCI band by MS

					Best-Hit Gene in Arabidopsis		Best-Hit Gene in Chlamydomonas	
Gene ID	NSAF(fmol)	Category	Gene annotation in this study	Accesson No.	Gene ID	Gene Name	Gene ID	Gene Name
Gene11703	9063	LHC	LHCA1	ICQU01000025	AT3G54890.1	LHCA1	Cre06.g283050.t1.2	LHCA1
Gene11916	6111	PSI	PSAG	ICQU01000031	AT1G55670.1	PSAG	Cre12.g560950.t1.2	PSAG
Gene11867	5182	PSI	PSAE	ICQU01000029	AT2G20260.1	PSAE-2	Cre10.g420350.t1.2	PSAE
Gene11501	5101	PSI	PSAF	ICQU01000021	AT1G31330.1	PSAF	Cre09.g412100.t1.2	PSAF
Gene11536	4709	LHC	LHCA3	ICQU01000022	AT1G61520.1	LHCA3	Cre11.g467573.t1.1	LHCA3
Gene10884	4541	LHC	LHCA9	ICQU01000011	AT1G45474.1	LHCA5	Cre07.g344950.t1.2	LHCA9
Gene11865	2669	PSI	PSAH	ICQU01000028	AT1G52230.1	PSAH-2	Cre07.g330250.t1.2	PSAH
Gene11349	2593	LHC	LHCA2	ICQU01000018	AT3G61470.1	LHCA2	Cre16.g687900.t1.2	LHCA7
Gene9022	2205	PSI	PSAD	ICQU01000006	AT1G03130.1	PSAD-2	Cre05.g238332.t1.1	PSAD
Gene11585	1980	LHC	LHCA2	ICQU01000023	AT3G47470.1	LHCA4	Cre10.g452050.t1.2	LHCA4
Gene11715	1861	LHC	algae-type LHCA2	ICQU01000026	AT1G45474.1	LHCA5	Cre12.g508750.t1.2	LHCA2
Gene11306	1695	LHC	LHCBM	ICQU01000014	AT2G05100.1	LHCB2.1	Cre01.g066917.t1.1	LHCBM1
Gene3547	1333	Others			AT5G59970.1		Cre12.g506350.t1.2	HFO18
Gene3504	1286	PSI	PSAB	ICQU01000002	ATCG00340.1	PSAB	NP_958404.1	PSAB
Gene11650	1115	PSI	PSAL	ICQU01000024	AT4G12800.1	PSAL	Cre12.g486300.t1.2	PSAL
Gene11761	914	Others						
Gene11139	868	Others			AT2G28720.1		Cre13.g590750.t1.2	HTB21
Gene4841	822	PSII	PSBD	ICQU01000004	ATCG00270.1	PSBD	NP_958420.1	PSBD
Gene3244	820	PSI	PSAA	ICQU01000001	ATCG00350.1	PSAA	NP_958375.1	PSAA
Gene10292	783	PSII	PSBA	ICQU01000008	ATCG00020.1	PSBA	NP_958377.1	PSBA
Gene9567	782	Others						
Gene6617	765	Others			AT3G61320.1		Cre06.g261750.t1.2	
Gene7515	730	Others			AT5G04180.1	ATACA3	Cre09.g415700.t1.2	CAH3
Gene4825	682	PSII	PSBB	ICQU01000003	ATCG00680.1	PSBB	NP_958388.1	PSBB
Gene10541	675	LHC	LHCB4	ICQU01000009	AT2G40100.1	LHCB4.3	Cre17.g720250.t1.2	LHCB4
Gene5134	575	Others						
Gene9362	538	Others			AT4G38970.1	FBA2	Cre05.g234550.t1.2	FBA3
Gene6781	535	PSII	PSBC	ICQU01000005	ATCG00280.1	PSBC		