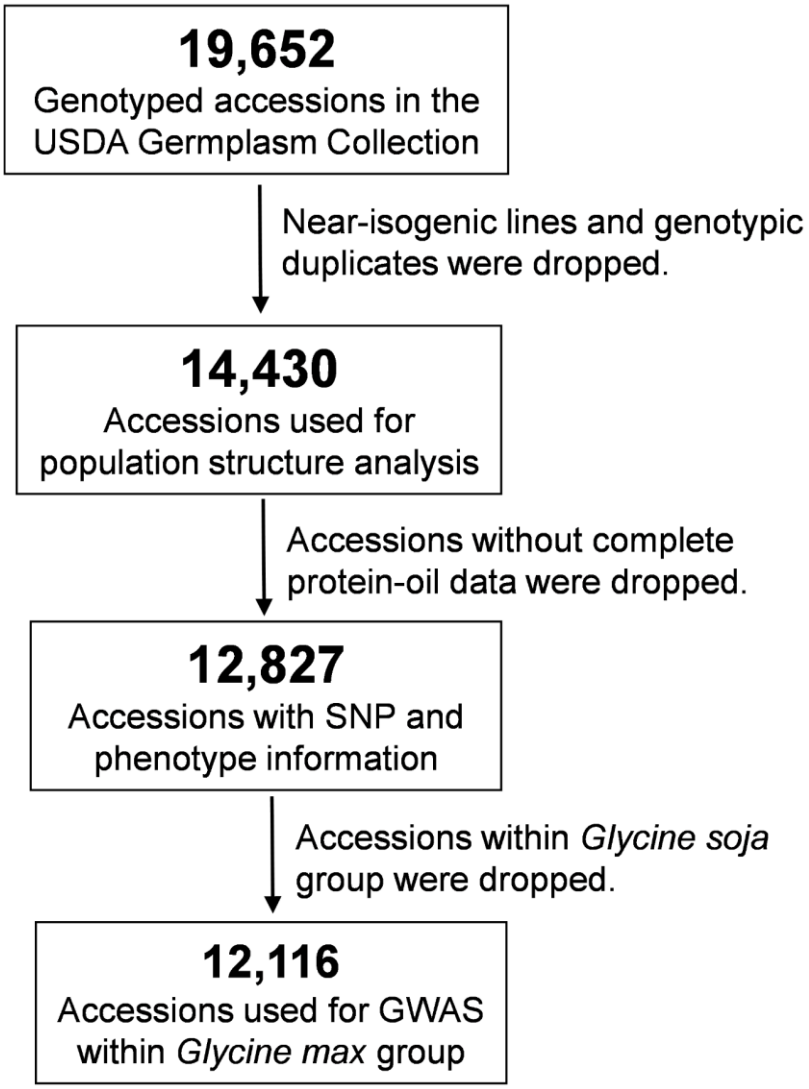
A Population Structure and Genome-wide Association Analysis on the USDA Soybean Germplasm Collection

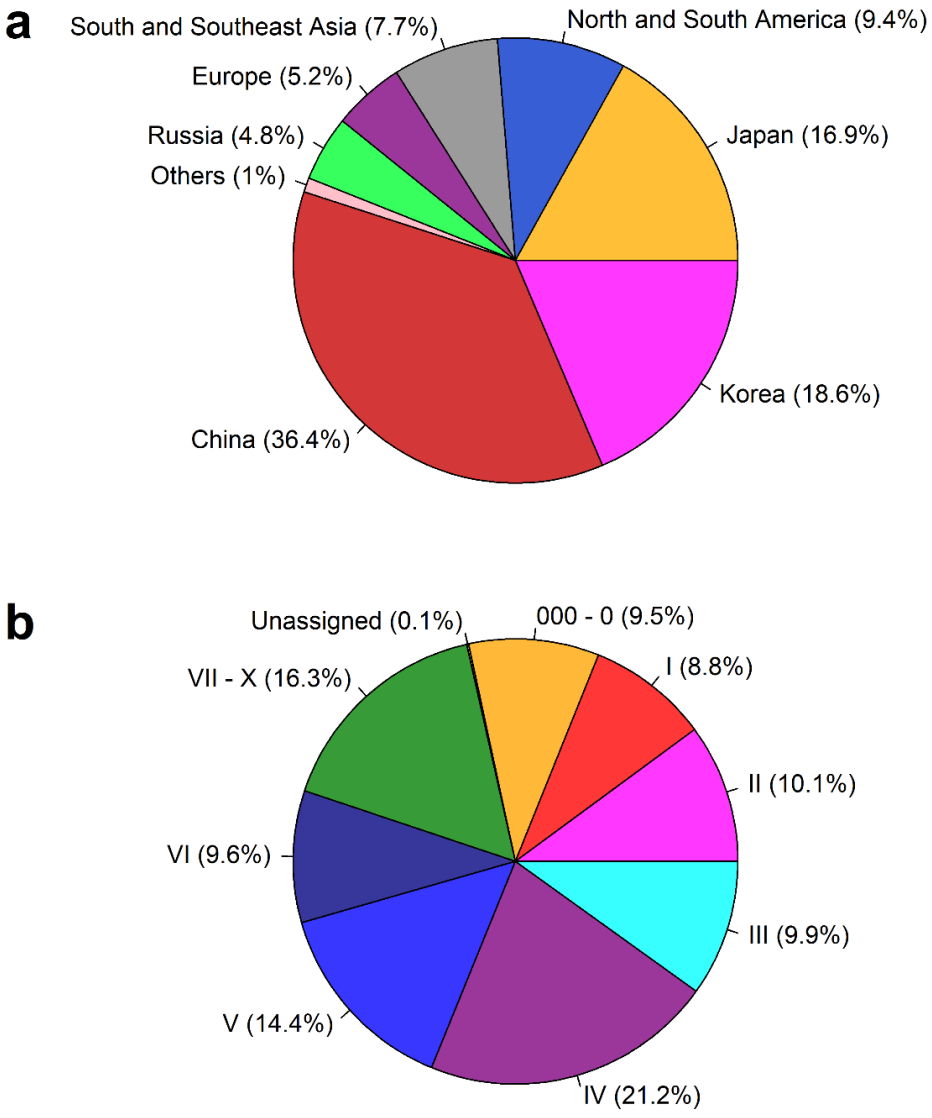
Nonoy Bandillo, Diego Jarquin, Qijian Song, Randall Nelson, Perry Cregan, Jim Specht, and Aaron Lorenz*

N. Bandillo, D. Jarquin, J. Specht and A. Lorenz, Dept. of Agronomy & Horticulture, University of Nebraska-Lincoln, Keim Hall Lincoln, NE 68583-0915; Q. Song and P. Cregan, Soybean Genomics and Improvement Laboratory Beltsville Agricultural Research Center, Beltsville, MD 20705; R. Nelson, USDA-ARS, Room 232, 1101 Peabody Drive, Urbana, IL 61801-0000. Corresponding author (alorenz2@unl.edu).

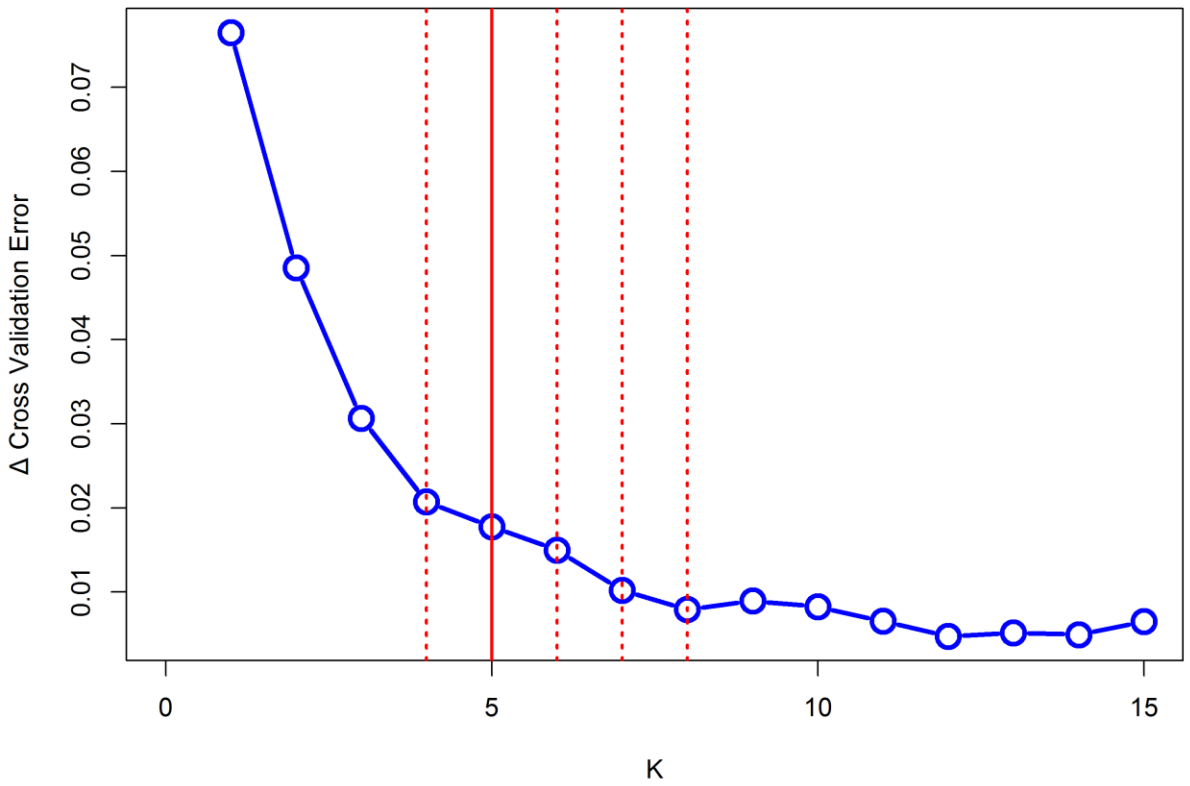
Supplementary Figures (S1-S6)



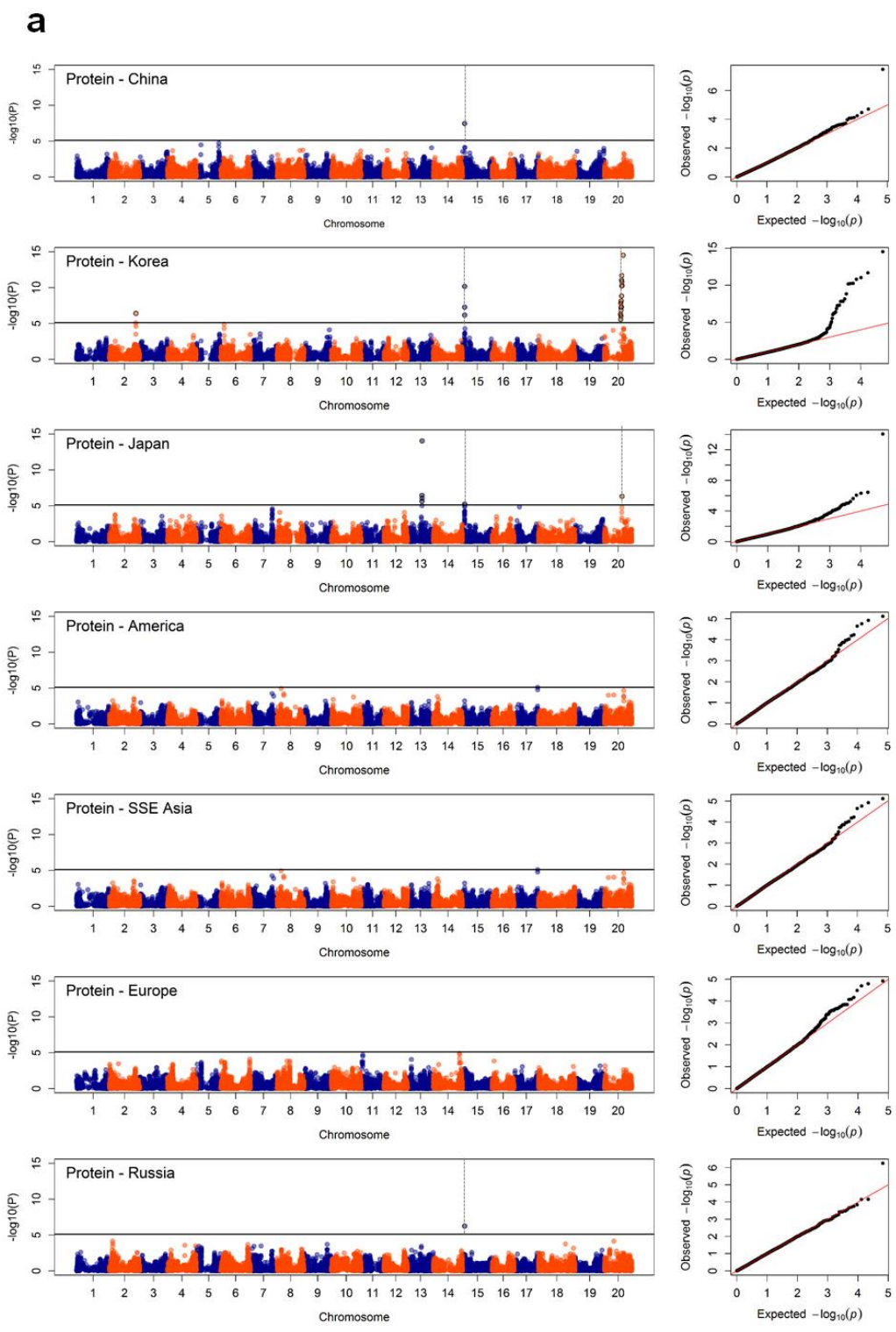
Supplementary Figure S1. The stepwise filtering of *G. max* and *G. soja* accessions held in the USDA Germplasm Collection for analysis of population structure and genome-wide association mapping.



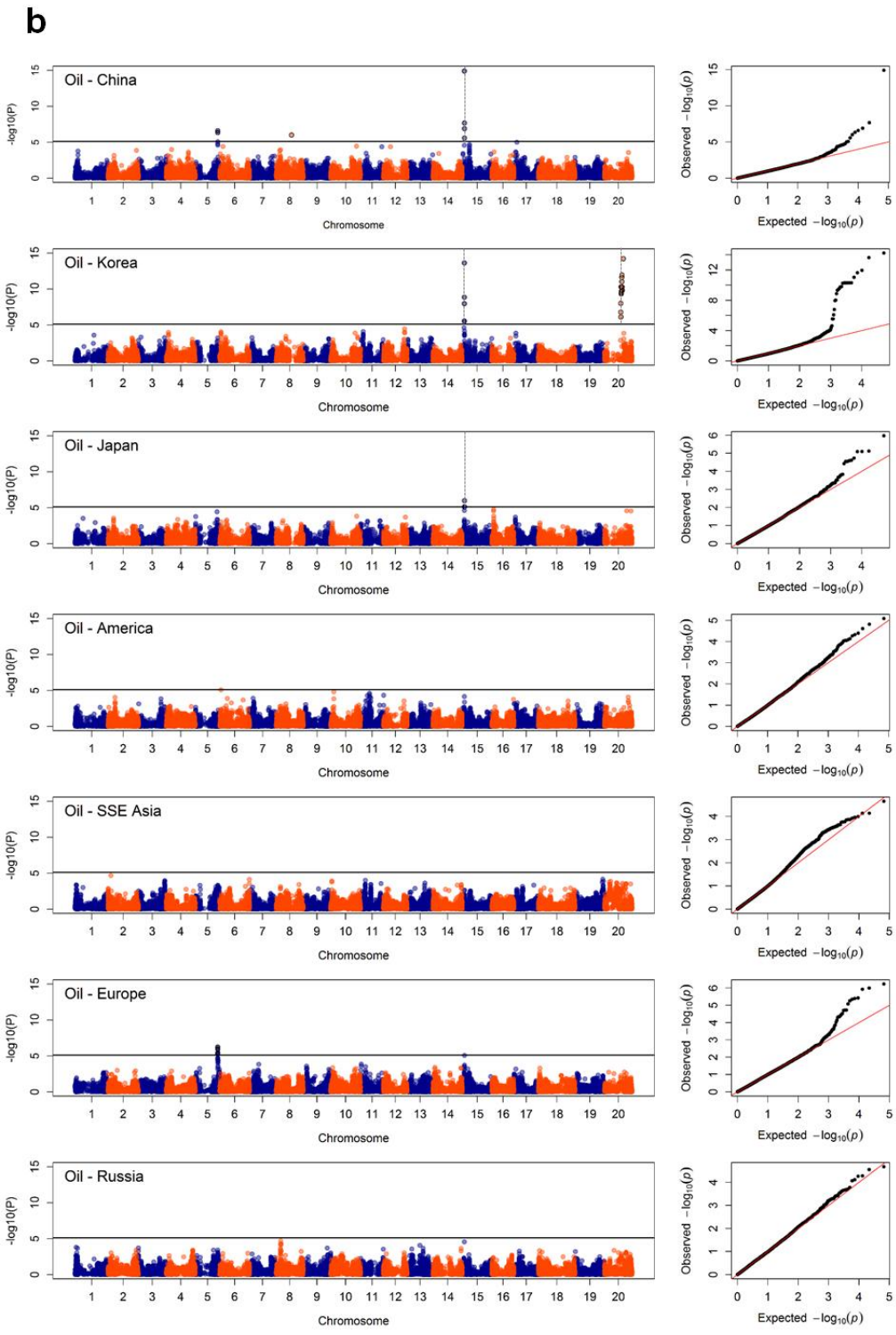
Supplementary Figure S2. Percentage distribution of the 14,430 soybean accessions used in the population structure analysis according to world region (panel a) and maturity group class (panel b).



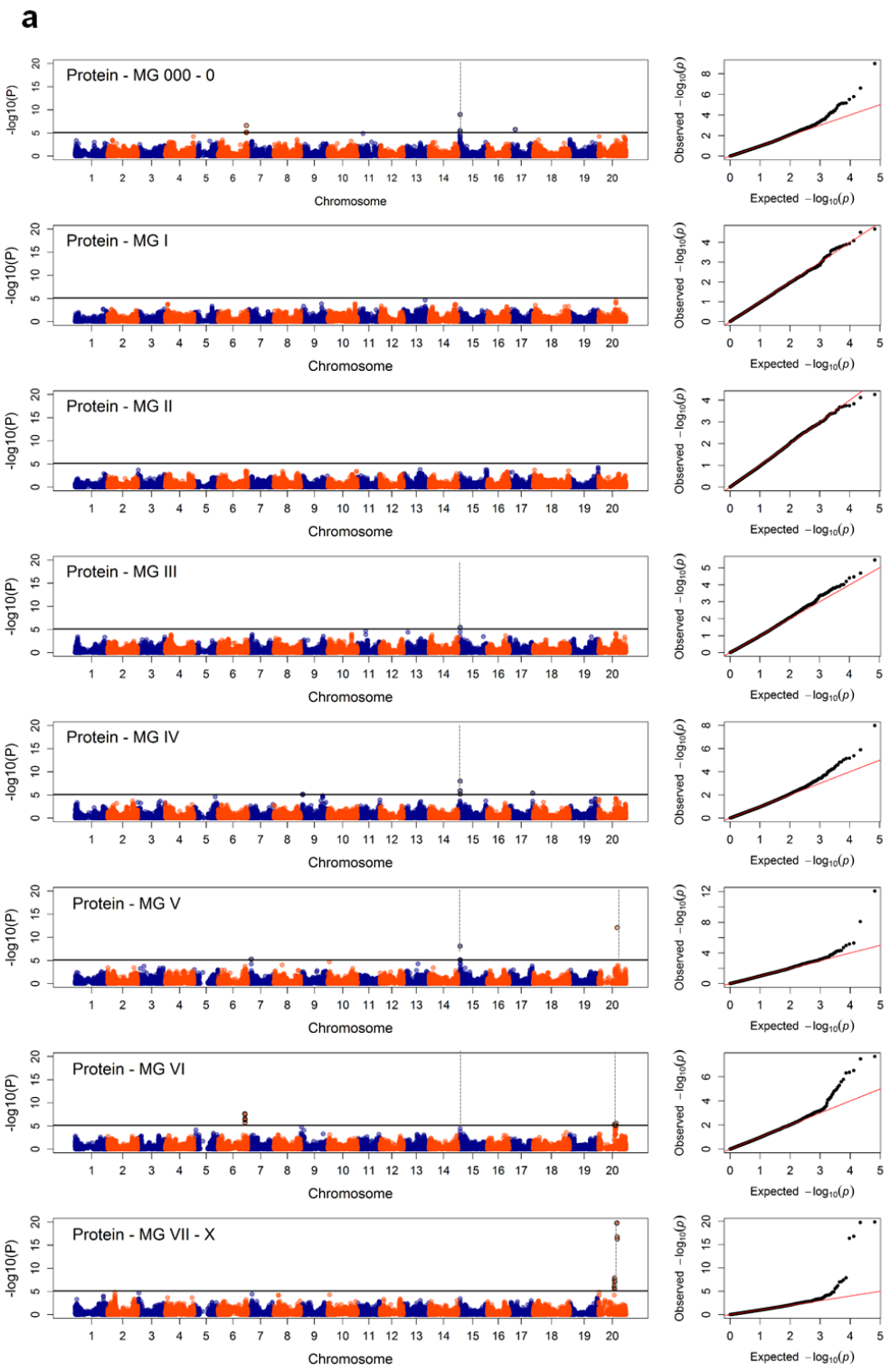
Supplementary Figure S3. Exploration of the optimal number of genetic subpopulations (K) using Δ cross-validation error values in the soybean germplasm collection. A solid line denotes the choice of K=5 which represents the most likely number of subpopulations within the soybean germplasm collection.



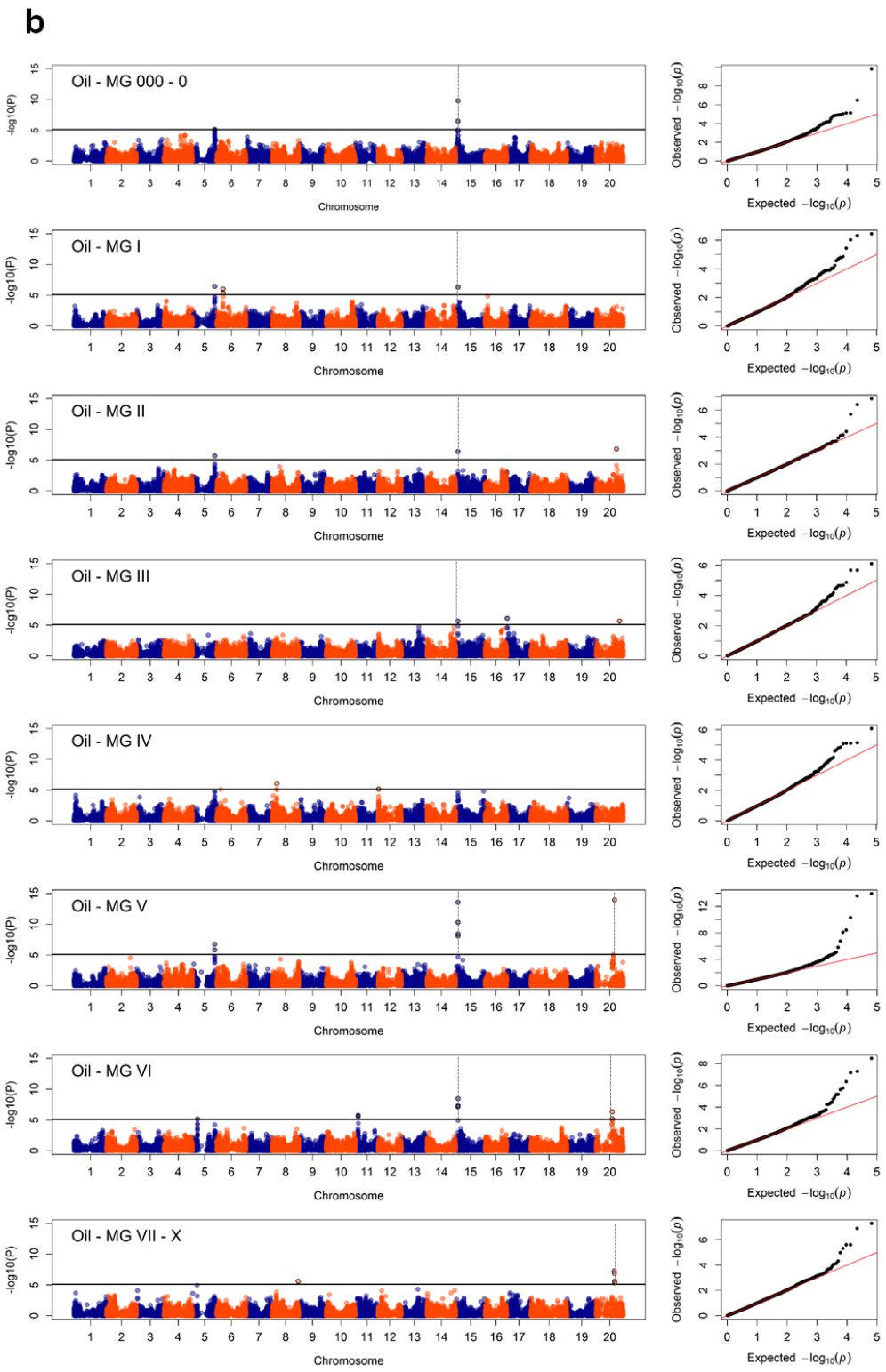
Supplementary Figure S4. Genome-wide association study for protein (panel a) and oil (panel b) within each world region class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$(P value) (right) are vertically arranged in each panel. Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring a significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the SNP detected for either protein or oil using 12, 116 *G. max* accessions.



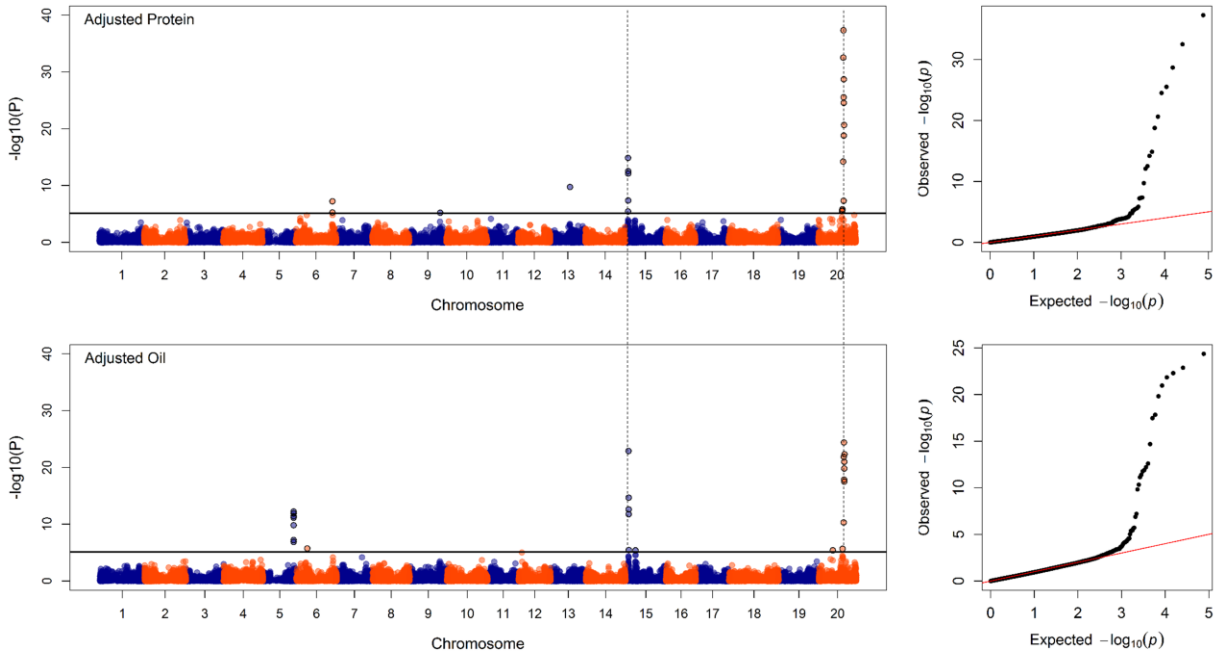
Supplementary Figure S4. Genome-wide association study for protein (panel a) and oil (panel b) within each world region class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$(P value) (right) are vertically arranged in each panel. Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring a significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the SNP detected for either protein or oil using 12, 116 *G. max* accessions.



Supplementary Figure S5. Genome-wide association study for protein (panel a) and oil (panel b) within each MG class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$ (P value) (right). Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the SNP detected for either protein or oil using 12, 116 *G. max* accessions.



Supplementary Figure S5. Genome-wide association study for protein (panel a) and oil (panel b) within each MG class. Manhattan plot for association within each subpopulation (left) and quantile-quantile plots of -$\log_{10}$ (P value) (right). Markers are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line indicates the calculated threshold value for declaring significant association. The dashed vertical lines indicate the same significant associations detected in two or more MGs that co-localized with the SNP detected for either protein or oil using 12, 116 *G. max* accessions.



Supplementary Figure S6. Genome-wide association scans for *G. max* accessions using adjusted phenotype data for seed oil and protein content. Manhattan plots show the associations for seed protein and oil with SNP markers that are plotted on the x-axis according to their physical position on each chromosome. The solid horizontal line denotes the calculated threshold value for declaring significant association. The dashed vertical lines indicate that the significant association positions on chromosome 15 and 20 for protein were the same as for those oil.