THE GENETIC ARCHITECTURE OF SEED COMPOSITION IN SOYBEAN IS REFINED BY GENOME-WIDE ASSOCIATION SCANS ACROSS MULTIPLE POPULATIONS

Justin N. Vaughn[*], Randall L. Nelson[§], Qijian Song[†], Perry B. Cregan[†], and Zenglu Li[*]

[*]Center for Applied Genetic Technologies and Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602

[§]Soybean Maize Germplasm, Pathology, and Genetics Research Unit, US DA, Agricultural Research Service, and Department of Crop Sciences University of Illinois, Urbana, IL 61801

[†]Soybean Genomics and Improvement Laboratory, U.S. Department of Agriculture, Agricultural Research Service, Beltsville, MD 20705

Corresponding author:

Zenglu Li

Center for Applied Genetic Technologies

The University of Georgia

111 Riverbend Road

Athens, GA 30602-6810

phone: 706/542-9805

email: zli@uga.edu

A

Methionine: MS-1997

1(7.5) *5(5)* *3(6.2)* *2(6.5)* *8(4.2)* *7(4.1)* *4(5.4)* *10(4.4)* *2(3.9)* *6(5.3)* *9(4.1)*

Methionine: IL-1996

1(11.5) *2(8.7)* *6(4.9)* *3(7.5)* *7(4.4)* *5(4.9)* *8(4.3)* *4(3.9)* *11(3.8)* *(7.4)*

B

Threonine: MS-1997

1(10.7) *10(3.9)* *2(4.2)* *11(3.7)* *9(3.4)* *(3.8)* *6(3.9)* *3(4)* *8(4)* *7(3.7)*

Threonine: IL-1996

1(6.9) *5(3.7)* *6(4.1)* *8(3.8)* *3(4.1)* *2(4.3)* *7(4.1)* *4(4)* *10(3.8)* *9(3.9)* *11(3.7)*

C

Lysine: MS-1997

1(11) *4(6.2)* *2(8.9)* *5(5.3)* *3(6.3)* *8(4.9)* *6(5.6)* *7(5.1)* *11(4.1)* *9(4.6)* *10(4.3)*

Lysine: IL-1996

1(11.1) *3(6.2)* *5(4.5)* *2(7)* *6(4.1)* *10(4.4)* *4(4.5)* *8(4.5)* *7(4.3)* *9(4.4)* *(16)*

D

Cysteine: MS-1997

1(12.3) *2(9.9)* *3(4.3)* *7(4.2)* *10(3.8)* *9(4)* *4(4.5)* *5(4.1)* *8(4.1)* *6(4.1)* *11(4)*

Cysteine: IL-1996

1(5.4) *2(3.9)* *4(3.9)* *8(4.3)* *10(3.7)* *3(3.8)* *7(4.2)* *9(3.9)* *5(3.9)* *6(4)* *11(3.5)*

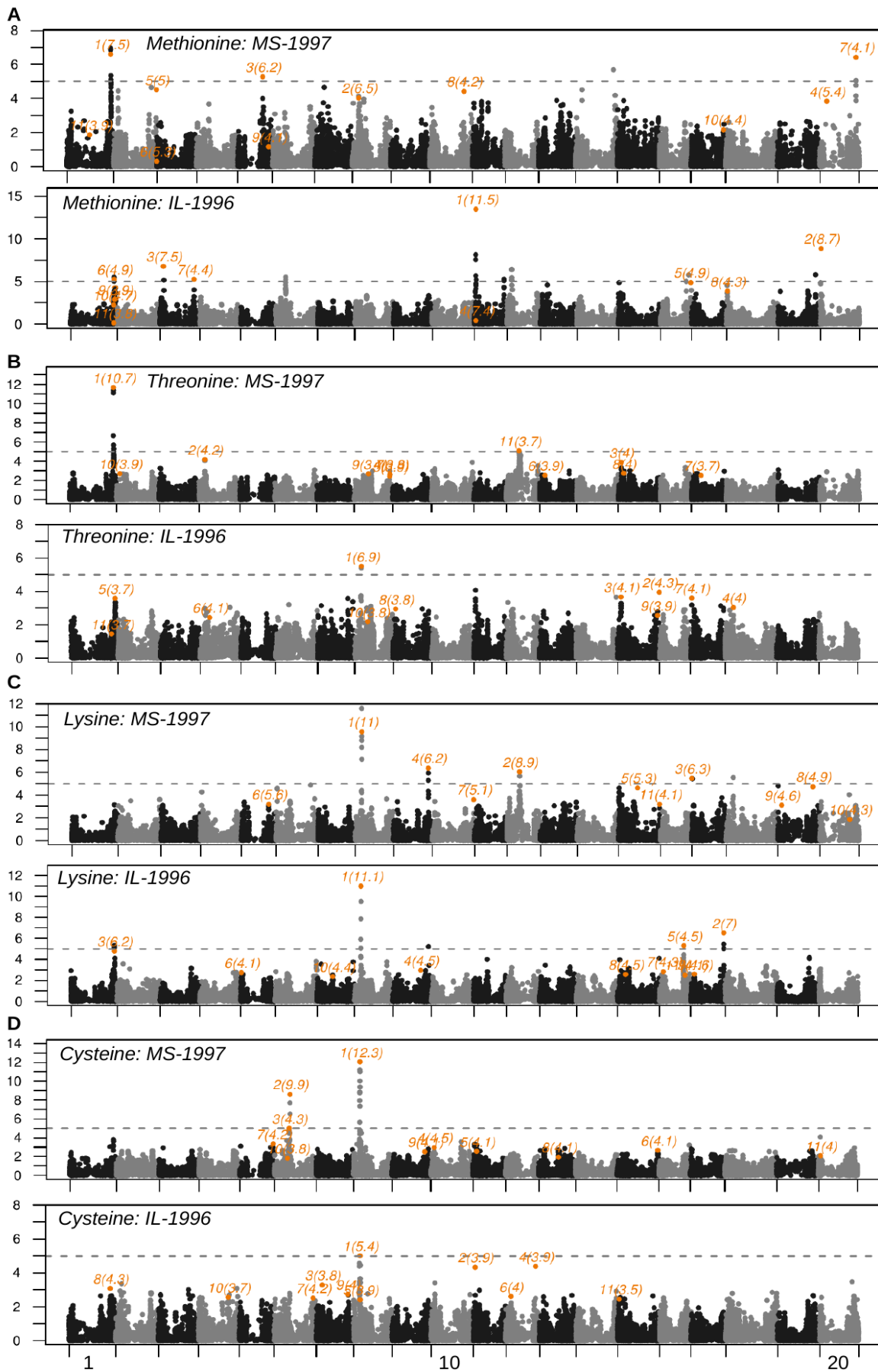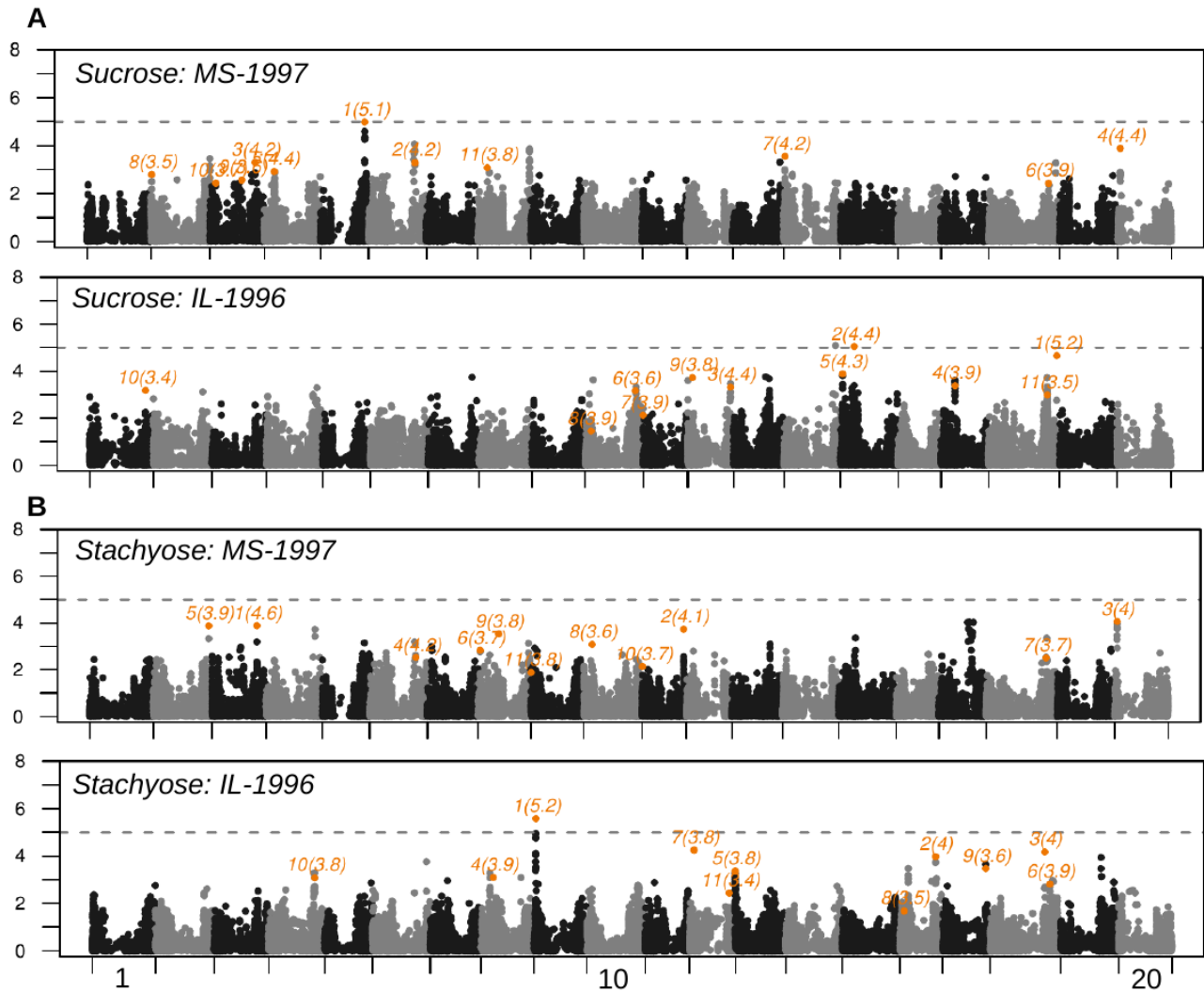1   10   20   2 SI

Figure S1 (previous page): GWA scan results for multiple essential amino acid levels in environment/population datasets MS-1997 and IL-1996. Each marker is plotted with its -log(p-val), as assessed using the CMLM method, on the y-axis and its physical position on the x-axis. Orange coloration indicates markers also identified by the MLMM method; their rank and -log(p-val) are also indicated in adjacent orange font. Chromosomes are indicated by alternating black and gray coloration and are plotted in order, 1 through 20.   A significance threshold of p-val$<10^{-5}$ is indicated by a dotted line; though arbitrarily chosen, it allows comparison across all GWAS scans in this study. A) Methionine. B) Threonine. C) Lysine. D) Cysteine.

Figure S2: GWA scan results for sucrose and stachyose levels in environment/population datasets MS-1997 and IL-1996. Each marker is plotted with its -log(p-val), as assessed using the CMLM method, on the y-axis and its physical position on the x-axis. Orange coloration indicates markers also identified by the MLMM method; their rank and -log(p-val) are also indicated in adjacent orange font. Chromosomes are indicated by alternating black and gray coloration and are plotted in order, 1 through 20. A significance threshold of p-val$<10^{-5}$ is indicated by a dotted line; though arbitrarily chosen, it allows comparison across all GWAS scans in this study. A) Sucrose. B) Stachyose.
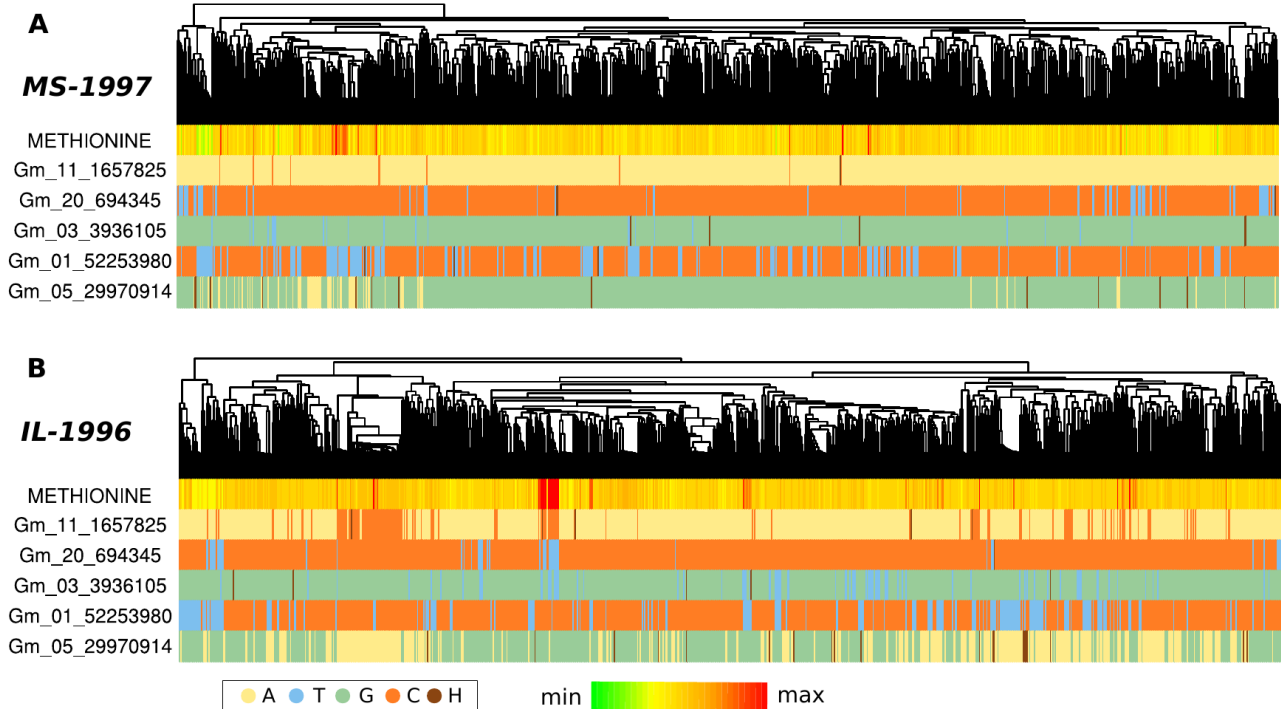
Figure S3: Allele distribution as it relates to population structure and methionine levels. Genotypes were clustered based on genetic distance across all markers (not just those depicted here). Each genotype used in the study represents a leaf in the dendrogram at the top of each panel. Methionine level is represented as a heat map with red being the highest value within a population and green being the lowest. For a given marker, a genotype is color coded according to the nucleotide for which it is homozygous; heterozygotes are shown as brown. A) MS-1997 population. B) IL-1996 population.
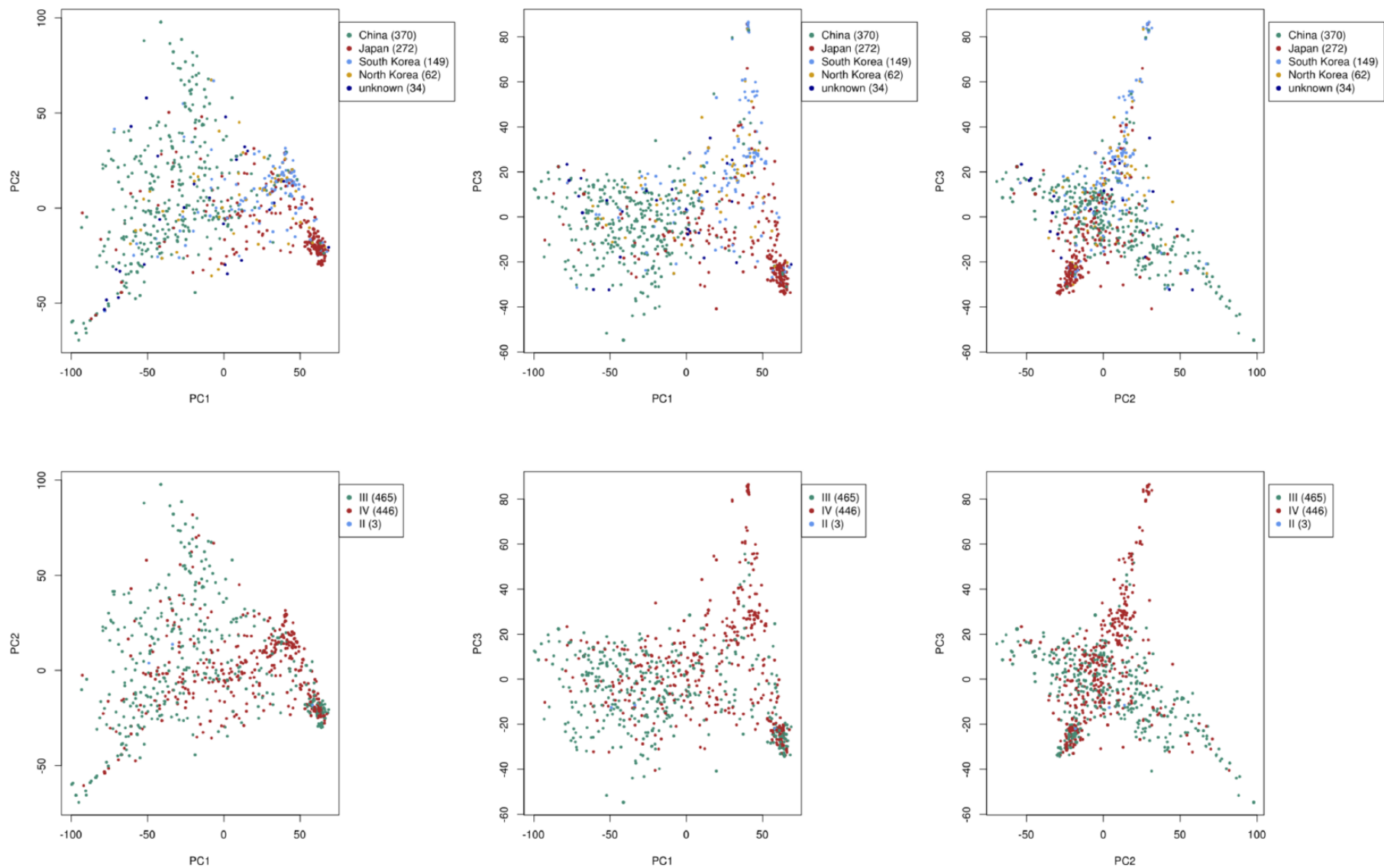
Figure S4: Principal component analysis (PCA) plots based on SNP information in the IL-1966 population used for protein and oil GWAS. Country of origin is indicated in the top row. Maturity group is indicated in the bottom row.
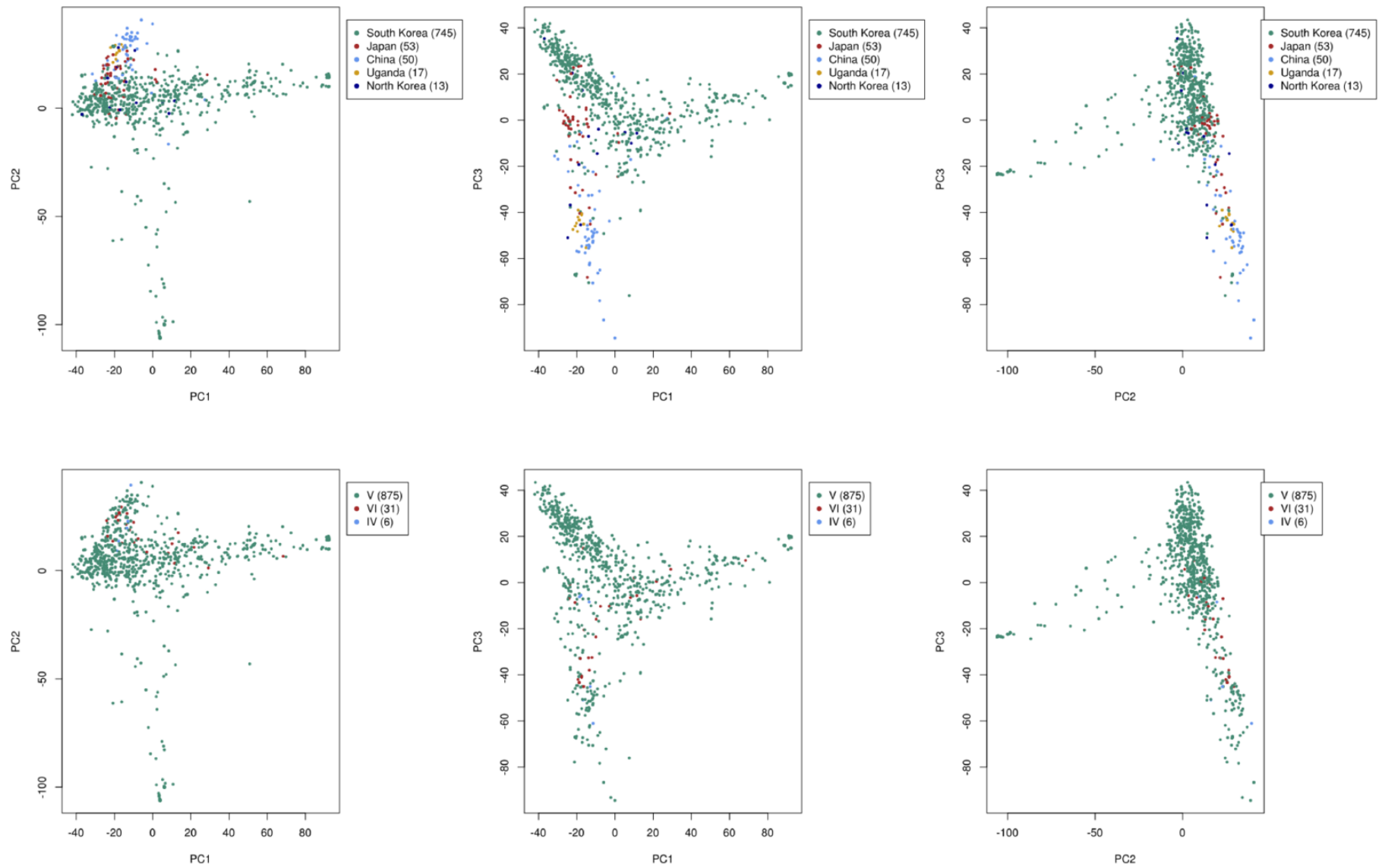
Figure S5: PCA plots based on SNP information in the MS-2000 population used for protein and oil GWAS. Country of origin is indicated in the top row. Maturity group is indicated in the bottom row.
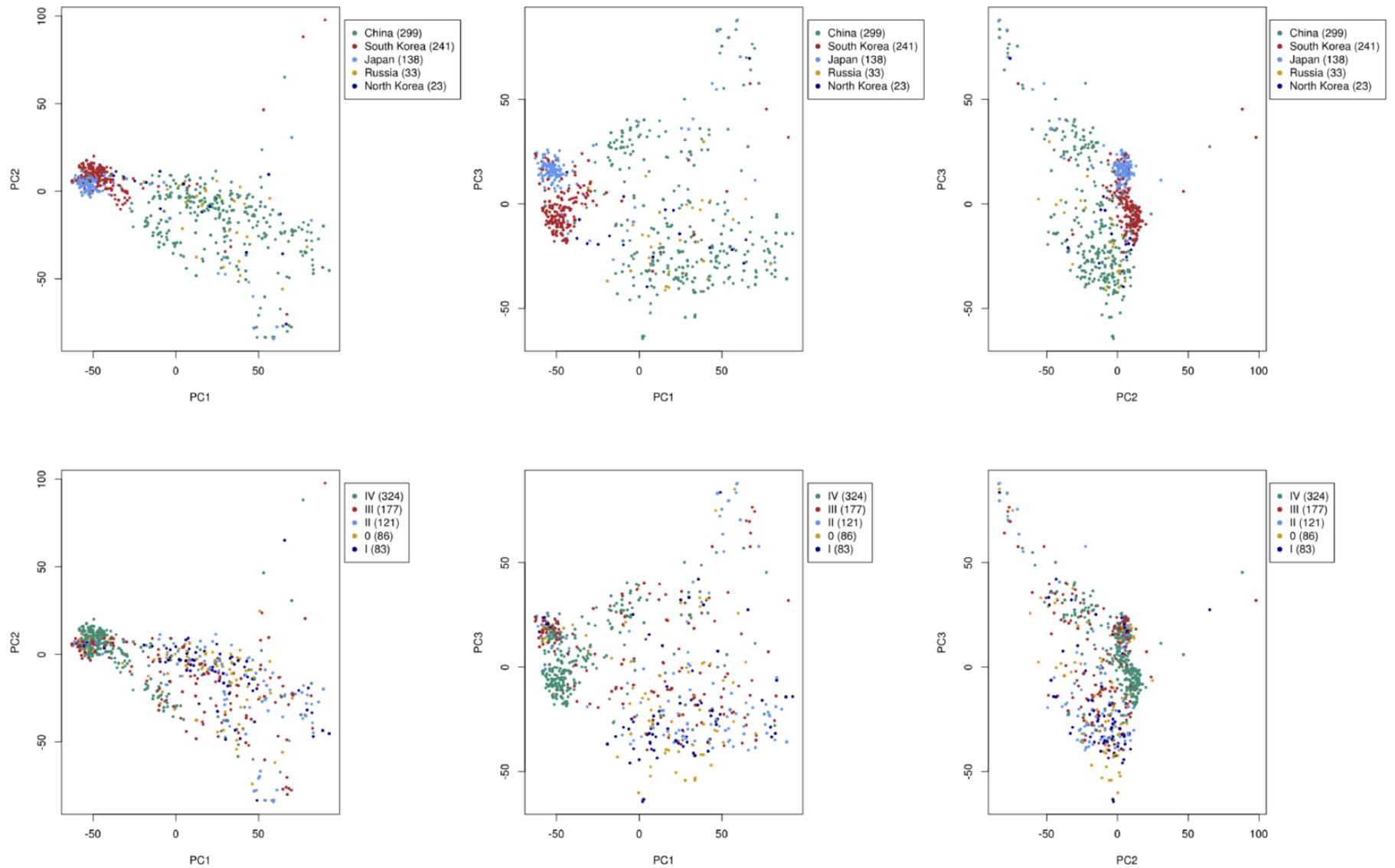
Figure S6: PCA plots based on SNP information in the IL-1996 population used for amino acid GWAS. Country of origin is indicated in the top row. Maturity group is indicated in the bottom row.
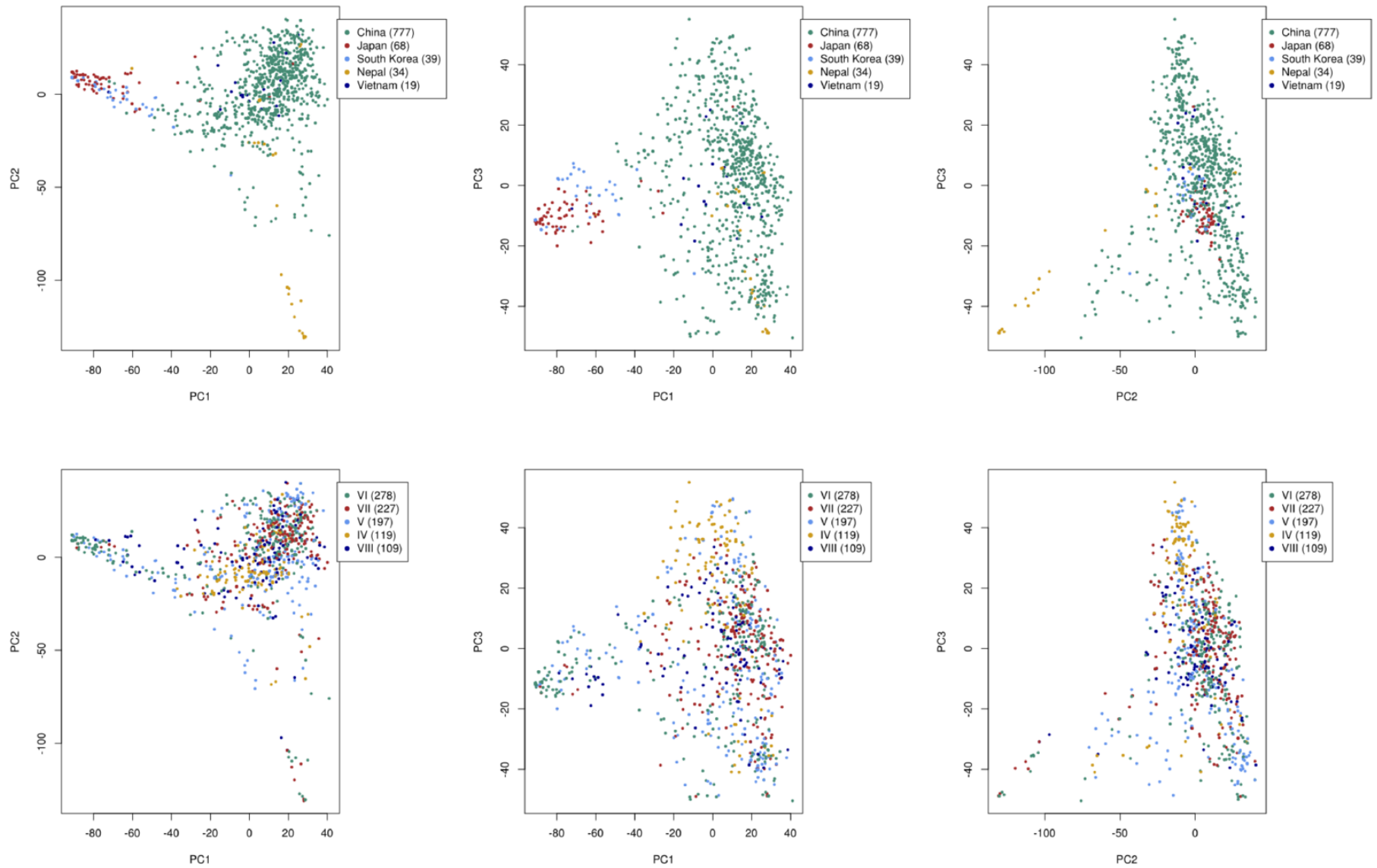
Figure S7: PCA plots based on SNP information in the MS-1997 population used for amino acid GWAS. Country of origin is indicated in the top row. Maturity group is indicated in the bottom row.

Table S1. Phenotyping locations and dates for all populations reported in this study. Populations within a particular "traits analyzed" category do not overlap with regard to genotype.

Traits analyzed	Population name	Location	Replication structure	Number of genotypes	Number of SNPs used
Protein and oil	IL-1964	Urbana, IL	1964(2), average	619	32,396
	IL-1966	Urbana, IL	1965(1) & 1966(1), average	977	33,265
	MS-1996	Stoneville, MS	1996(1) and 1997(1), average	728	32,328
	MS-2000	Stoneville, MS	1999(1) & 2001(1), average	934	28,622
Met, Cys, Lys, Thr, sucrose, and stachyose	IL-1996	Urbana, IL	1996(1)	900	34,479
	MS-1997	Stoneville, MS	1997(1)	978	32,112

[a] Value in parentheses is the number of replications phenotyped in that year

File S1

Averaged phenotypic values used in this study

Available for download as an Excel file at
http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.114.013433/-/DC1