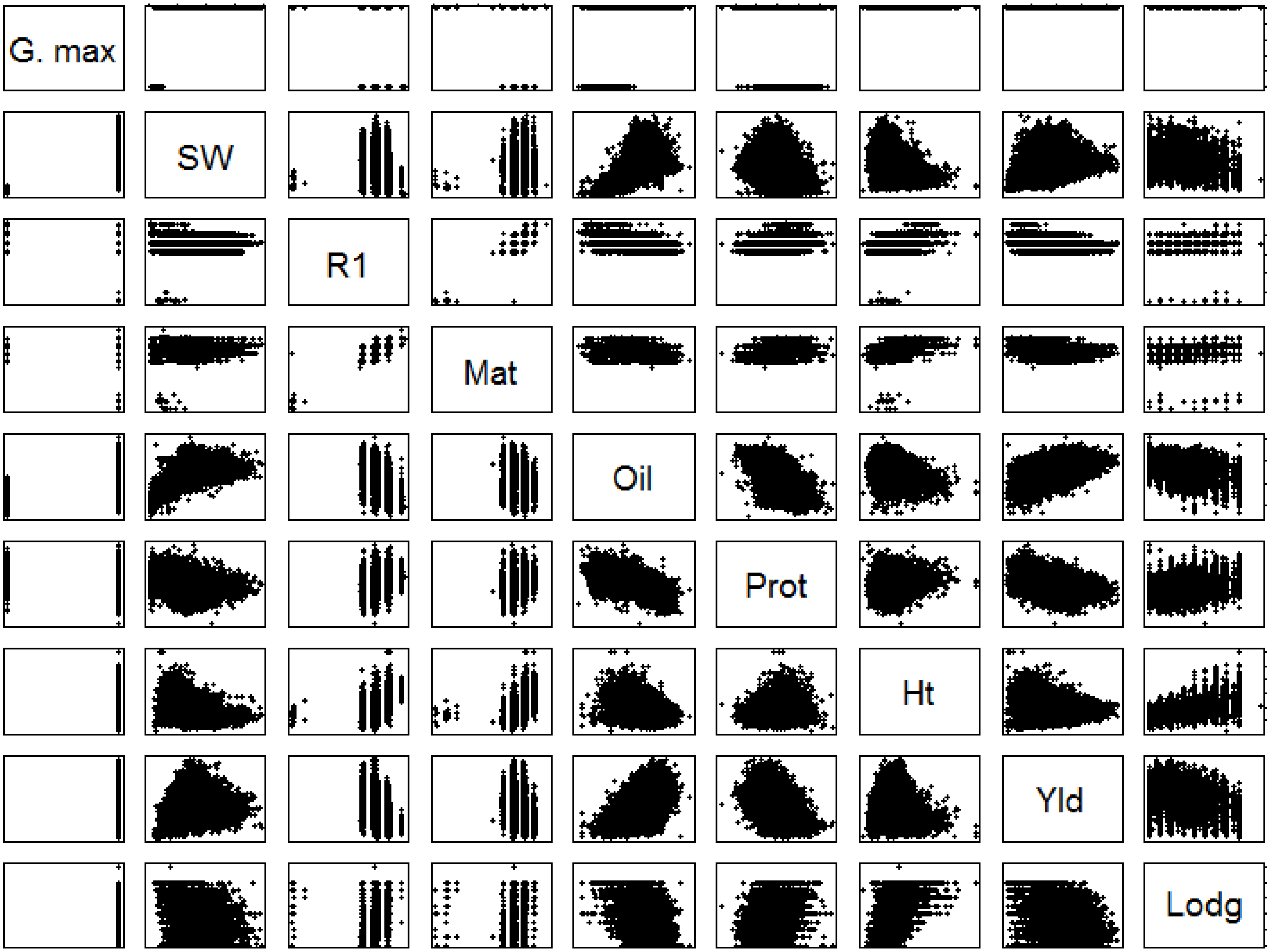
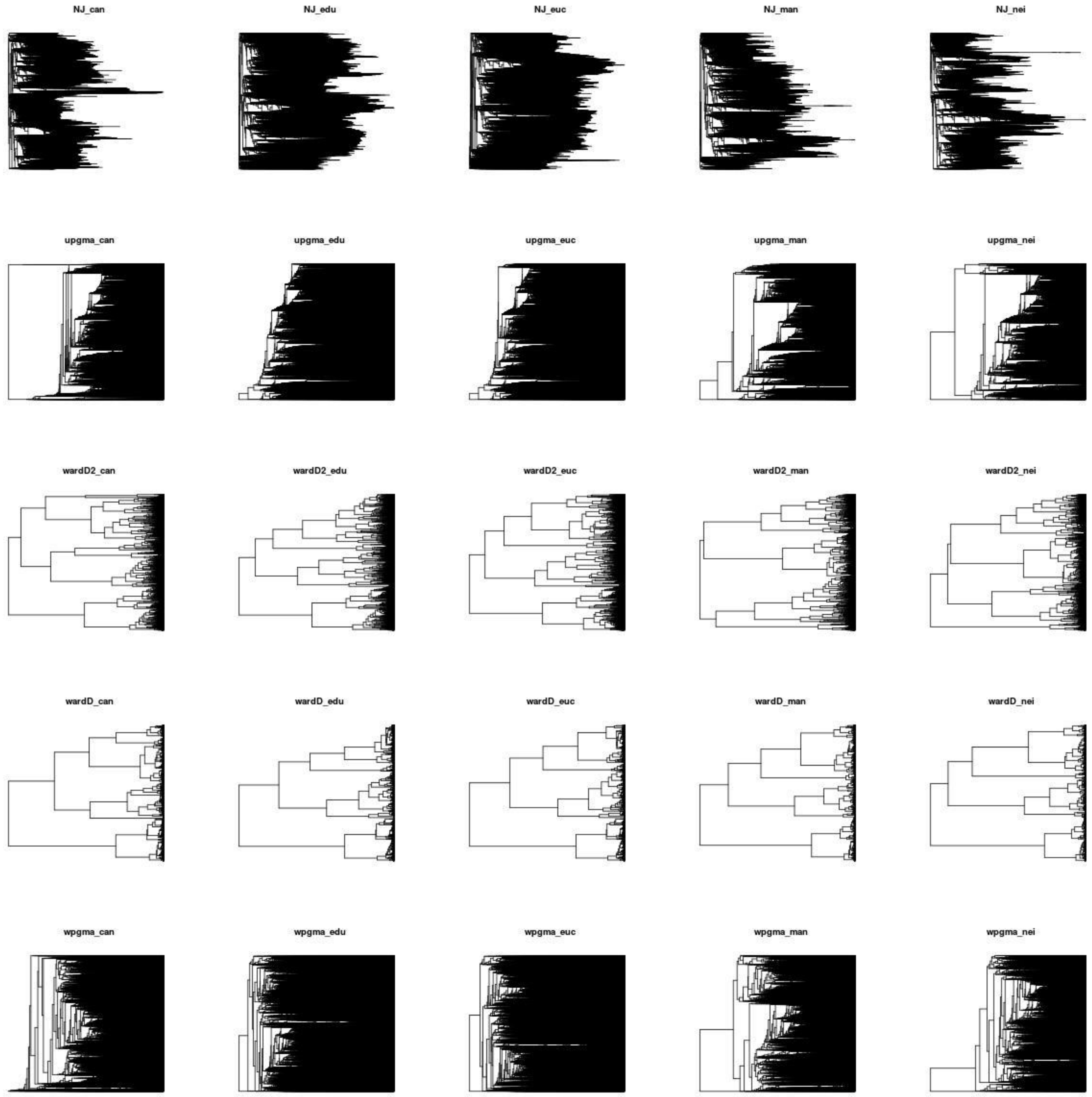
Supplementary File S1

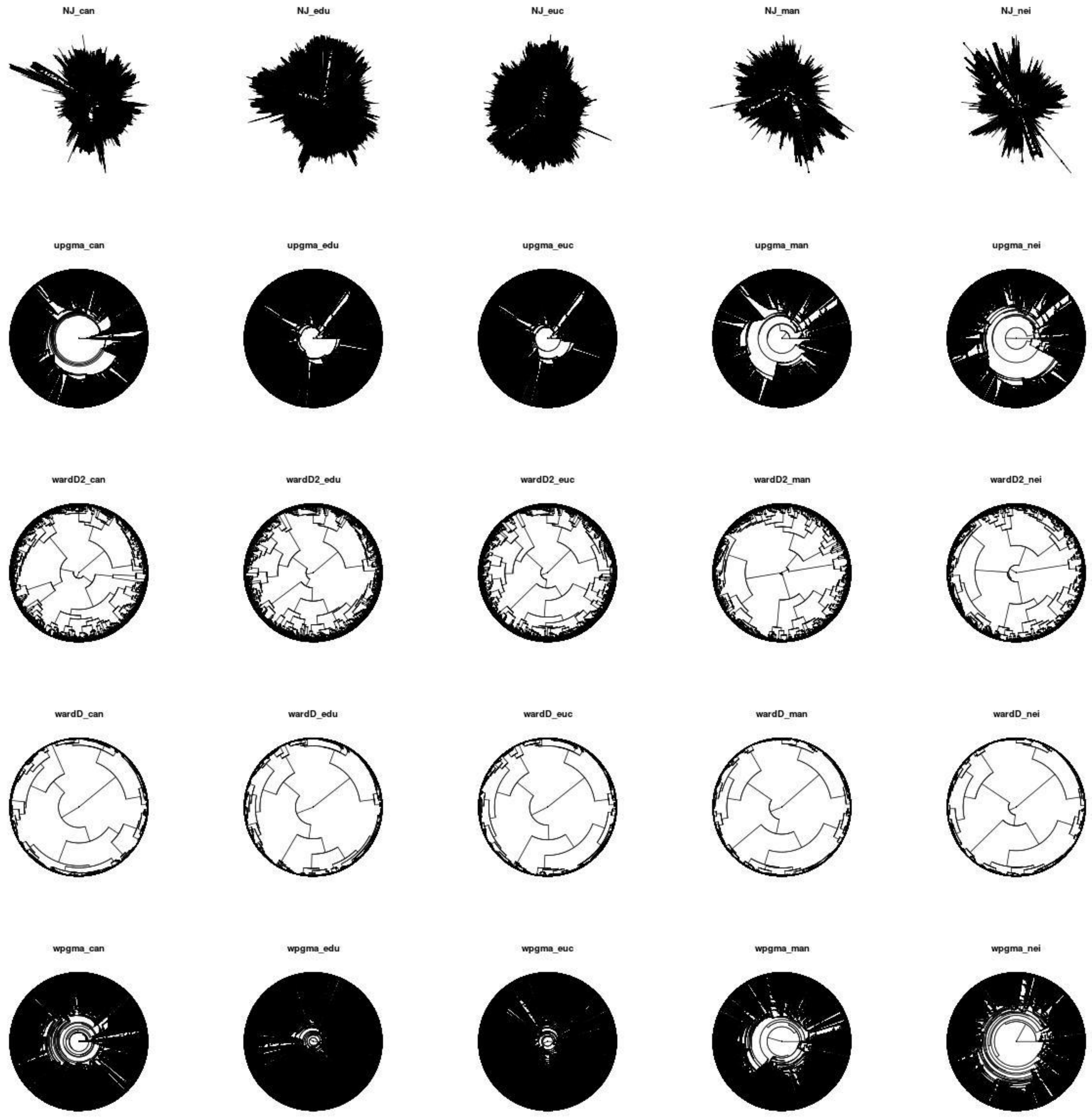
Scatterplot of observed phenotypes: Domesticated soybeans (MAX); Weight of 100 seeds (SW); Flowering date (R1); Maturity date (R8); Percentage of seed oil content (Oil); Percentage of seed protein content (Prot); Grain yield in grams per plant (Yld); Plant height in centimeters (Ht); Lodging score 1 to 5 (Lodg).



Outcome cladograms from unsupervised analyses. Below, each combination between clustering method (rows: NJ, UPGMA, WPGMA, Ward D and Ward D2) and distance metric (columns: Canberra, Edward's, Euclidean, Manhattan and Nei's).



Radial representation of cladograms from unsupervised analyses. Below, each combination between clustering method (rows: NJ, UPGMA, WPGMA, Ward D and Ward D_2) and distance metric (columns: Canberra, Edward's, Euclidean, Manhattan and Nei's).



Validation of effective population size in the germplasm collection

Consider the expression of the drift-mutation-migration model

$$F_{ST} = \frac{1}{1 + 4N_e(\mu + m)}$$

Which can be reformulated into

$$N_e = \frac{1}{4(\mu + m)F_{ST}} - F_{ST}$$

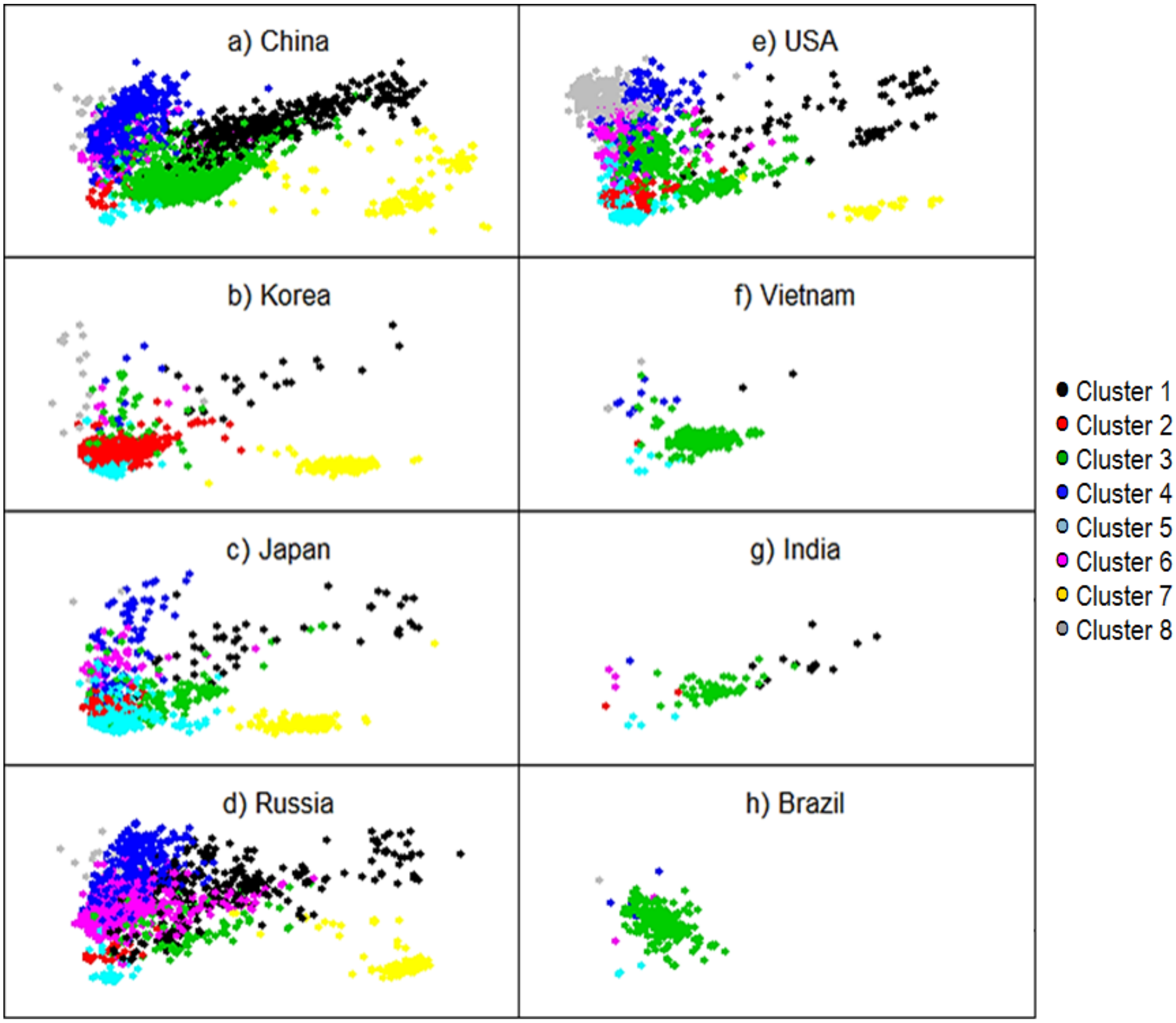
For a mutation rate $\mu = 1/2500$ (Tsuda et al. 2015) and migration rate assumed to be $m = 1\%,$ which can be interpreted as the crosses in soybeans designated to incorporate exotic material. We have estimated the average fixation index across all markers to be $F_{ST} = 0.2217$.

$$N_e = \frac{1}{4 \times (0.004 + 0.01) \times 0.2217} - 0.2217 = 108.21$$

Where the value $N_e = 108$ is consistent to the effective population size estimated from linkage disequilibrium ($N_e = 106$).

Tsuda M, Kaga A, Anai T, Shimizu T, Sayama T, Takagi K, Machita K, Watanabe S, Nishimura M, Yamada N, Mori S. Construction of a high-density mutant library in soybean and development of a mutant retrieval method using amplicon sequencing. BMC genomics. 2015 Dec;16(1):1014.

The eight countries providing the greatest contribution of entries to the germplasm collection are, in descending order, China, Korea, Japan, Russia, USA, Vietnam, India, and Brazil. The plot presents the MDS of each country while the table below it shows the genetic distance among the germplasm.

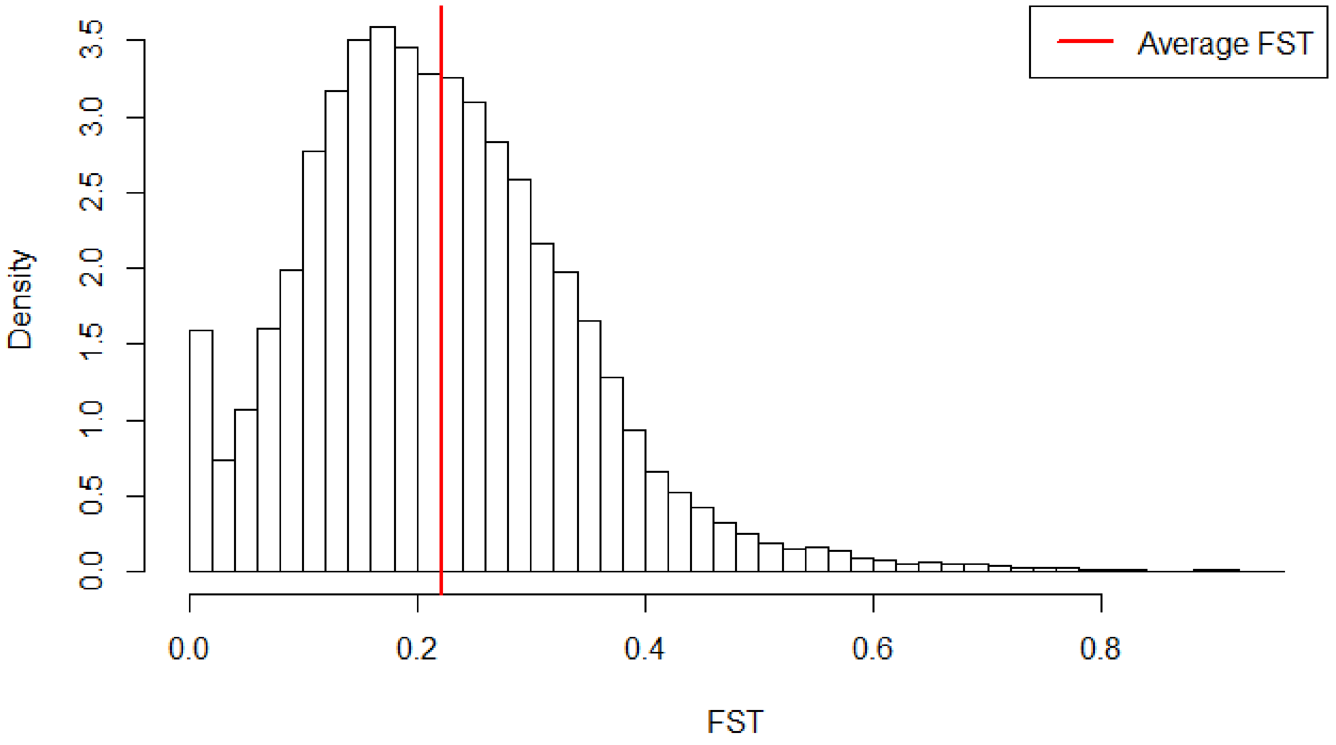


Nei's standard genetic distance among the germplasm from the eight major contributors

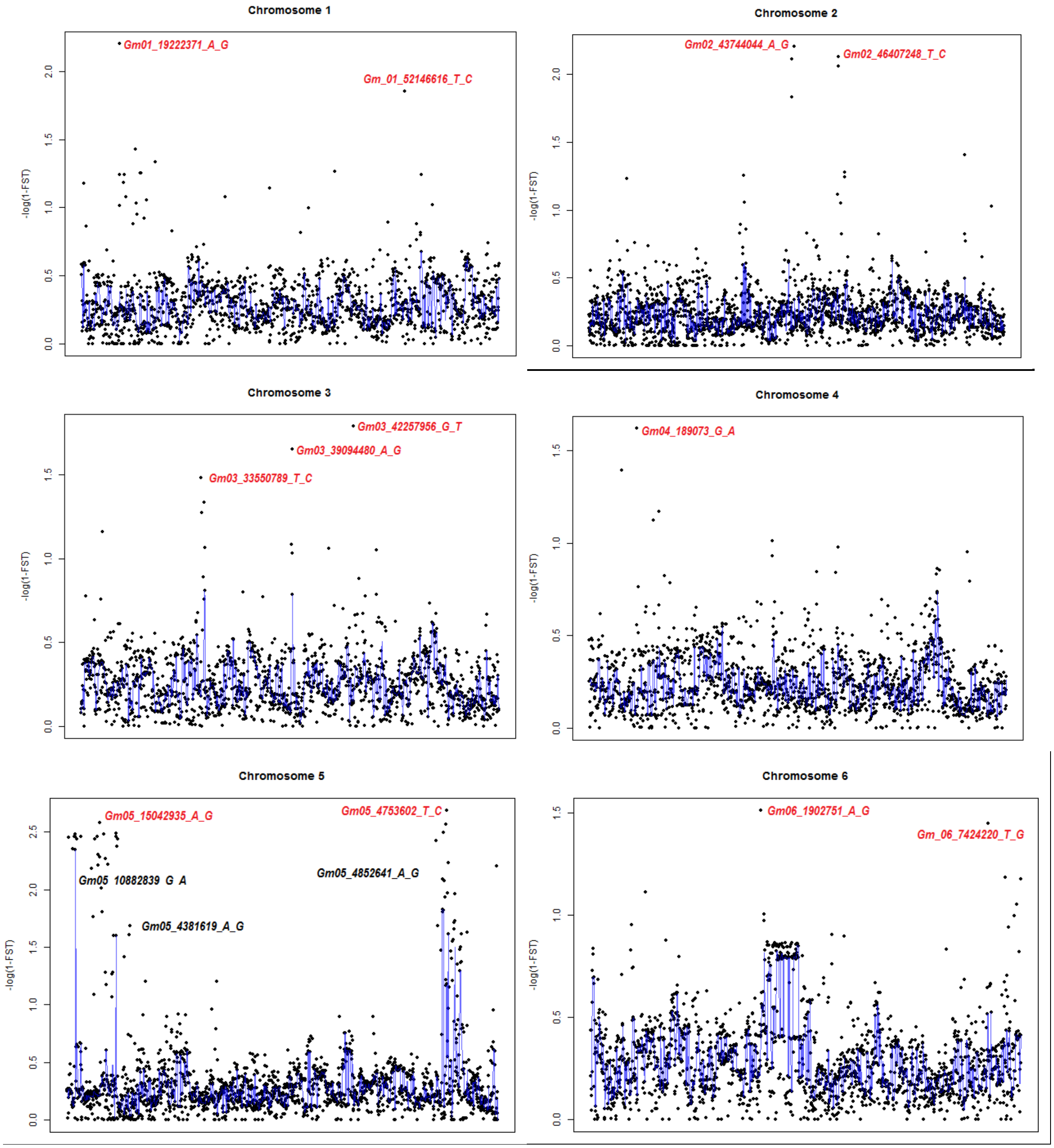
	Korea	Japan	Russia	USA	Vietnam	India	Brazil
China	0.048	0.050	0.030	0.043	0.021	0.050	0.048
Korea		0.022	0.059	0.064	0.065	0.098	0.074
Japan			0.064	0.067	0.066	0.099	0.072
Russia				0.033	0.073	0.089	0.070
USA					0.080	0.103	0.057
Vietnam						0.054	0.071
India							0.097

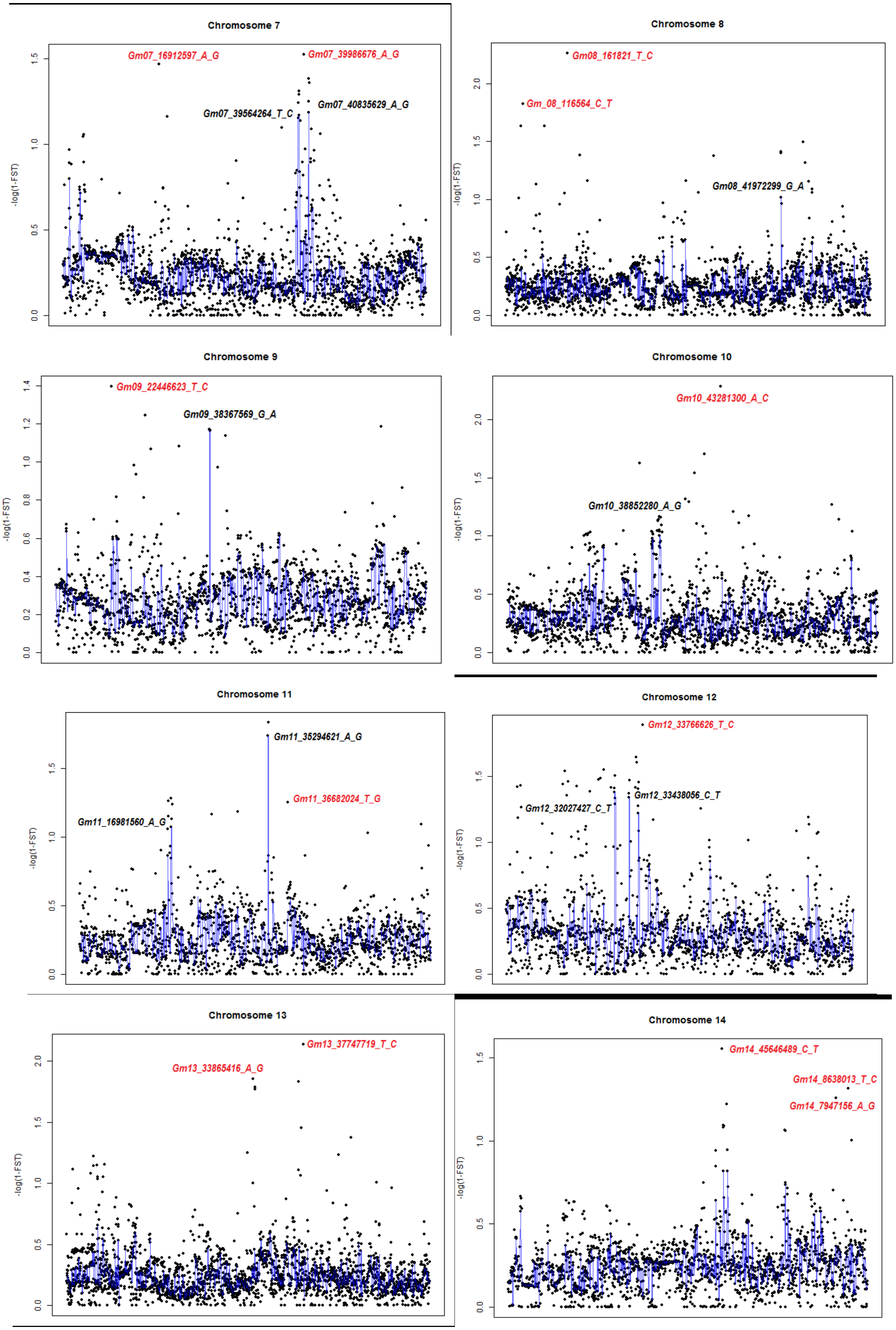
The study evaluated the genomic distance using Canberra, Manhattan, Euclidean, Edwards' (Cavalli-Sforza and Edwards, 1967) and Nei's standard genetic distance (Nei 1972). Canberra, Manhattan, and Euclidean distances were computed with a built-in function in R. We adapted algorithms to calculate Edwards' and Nei's genetic distance from the implementation of Jombart and Ahmed (2011). Tree construction employed pre-assessment of a set of hierarchical agglomerative clustering techniques, including the following methods: Ward's D_1 and D_2 (Murtagh and Legendre 2014), neighbor joining (Saitou and Nei, 1987), and weighted and unweighted pair group methods with arithmetic mean (WPGMA and UPGMA). The last method is analogous to the Fitch-Margoliash method (Fitch and Margoliash 1967). The figure below shows the cladogram of each combination of hierarchical clustering methods (rows) and the distance among the genotypes from the USDA soybean germplasm collection (columns).

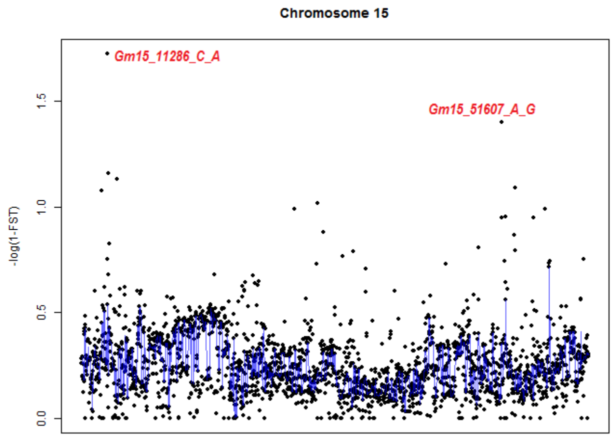
Histogram of Fixation Index (F_{ST}) by chromosome.

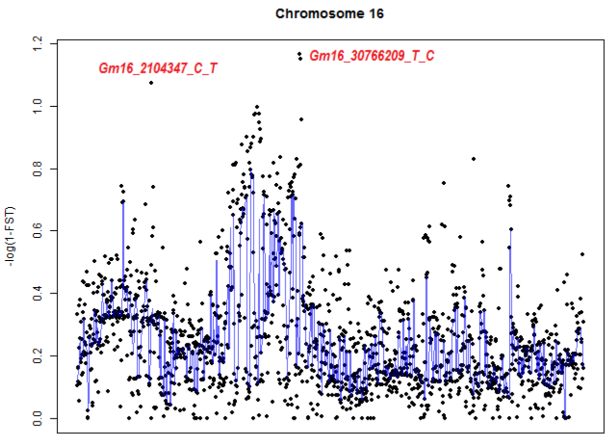


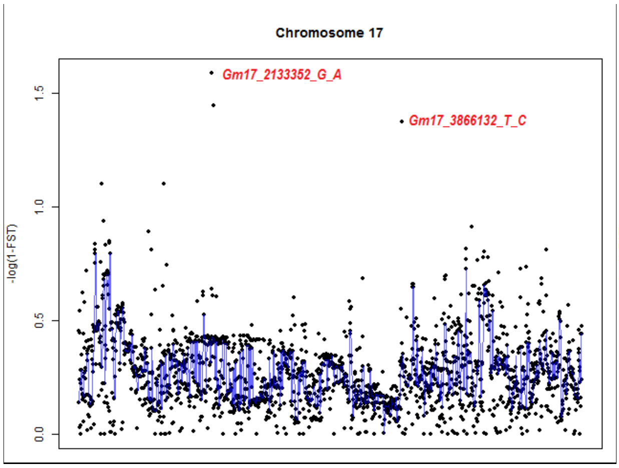
The following plots show the Fixation Index (F_{ST}) by chromosome. Black indicates SNPs on picks of observed selection events, while red indicates isolated picks.

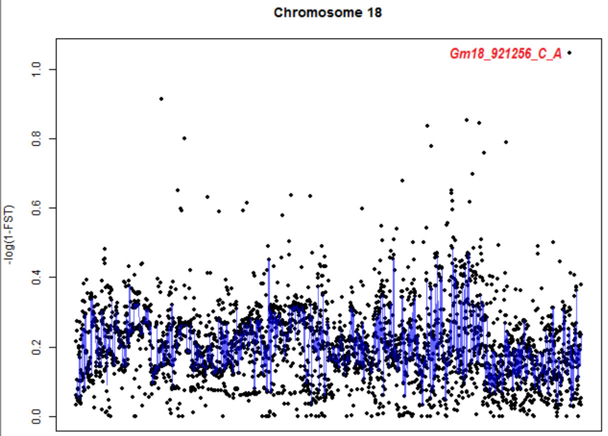

Chromosome 1
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
Gm01_19222371_A_G
Gm_01_52146616_T_C
Chromosome 2
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
Gm02_43744044_A_G
Gm02_46407248_T_C
Chromosome 3
-log(1-FST)
0.0
0.5
1.0
1.5
Gm03_42257956_G_T
Gm03_39094480_A_G
Gm03_33550789_T_C
Chromosome 4
-log(1-FST)
0.0
0.5
1.0
1.5
Gm04_189073_G_A
Chromosome 5
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
2.5
Gm05_15042935_A_G
Gm05_4753602_T_C
Gm05 10882839 G A
Gm05_4852641_A_G
Gm05_4381619_A_G
Chromosome 6
-log(1-FST)
0.0
0.5
1.0
1.5
Gm06_1902751_A_G
Gm_06_7424220_T_G

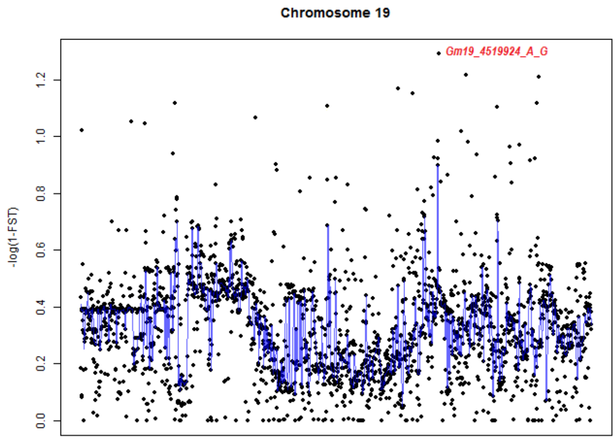

Chromosome 7
-log(1-FST)
0.0
0.5
1.0
1.5
Gm07_16912597_A_G
Gm07_39986676_A_G
Gm07_39564264_T_C
Gm07_40835629_A_G
Chromosome 8
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
Gm08_161821_T_C
Gm_08_116564_C_T
Gm08_41972299_G_A
Chromosome 9
-log(1-FST)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
Gm09_22446623_T_C
Gm09_38367569_G_A
Chromosome 10
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
Gm10_43281300_A_C
Gm10_38852280_A_G
Chromosome 11
-log(1-FST)
0.0
0.5
1.0
1.5
Gm11_35294621_A_G
Gm11_36682024_T_G
Gm11_16981560_A_G
Chromosome 12
-log(1-FST)
0.0
0.5
1.0
1.5
Gm12_33766626_T_C
Gm12_33438056_C_T
Gm12_32027427_C_T
Chromosome 13
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
Gm13_37747719_T_C
Gm13_33865416_A_G
Chromosome 14
-log(1-FST)
0.0
0.5
1.0
1.5
Gm14_45646489_C_T
Gm14_8638013_T_C
Gm14_7947156_A_G

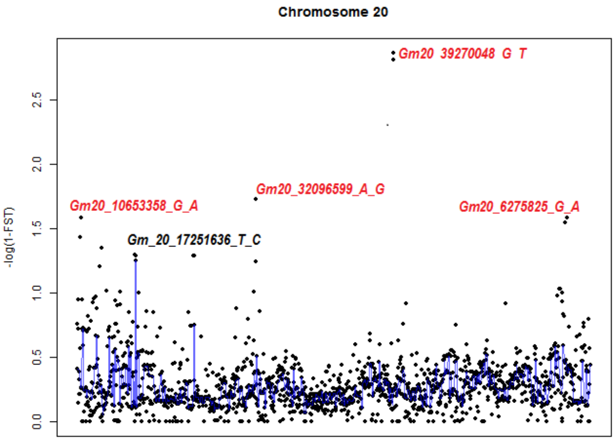

Chromosome 15
Gm15_11286_C_A
Gm15_51607_A_G
-log(1-FST)
0.0
0.5
1.0
1.5


Chromosome 16
Gm16_2104347_C_T
Gm16_30766209_T_C
-log(1-FST)
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Chromosome 17
Gm17_2133352_G_A
Gm17_3866132_T_C
-log(1-FST)
0.0
0.5
1.0
1.5


Chromosome 18
Gm18_921256_C_A
-log(1-FST)
0.0
0.2
0.4
0.6
0.8
1.0


Chromosome 19
Gm19_4519924_A_G
-log(1-FST)
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Chromosome 20
Gm20 39270048 G T
Gm20_10653358_G_A
Gm20_32096599_A_G
Gm20_6275825_G_A
Gm_20_17251636_T_C
-log(1-FST)
0.0
0.5
1.0
1.5
2.0
2.5