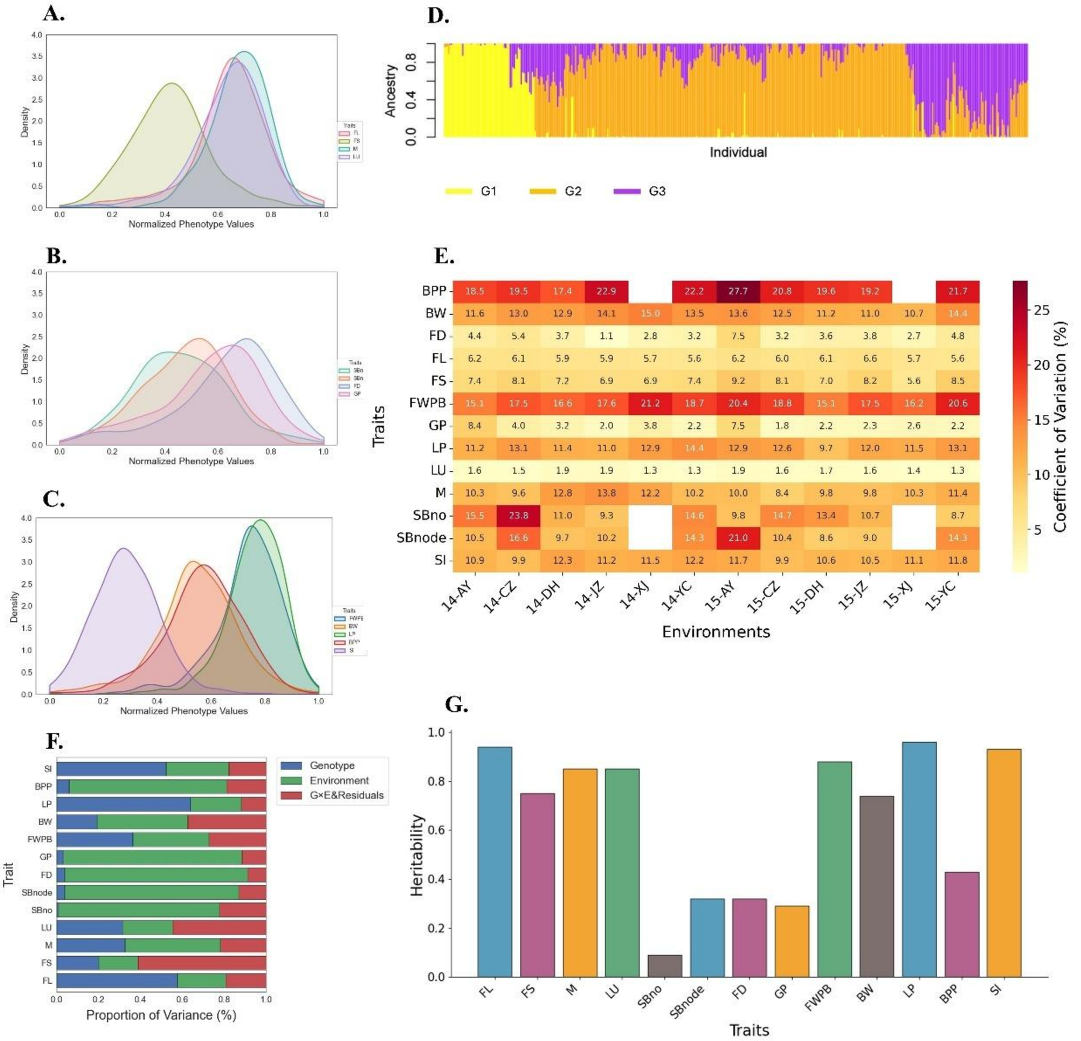
Supplementary information

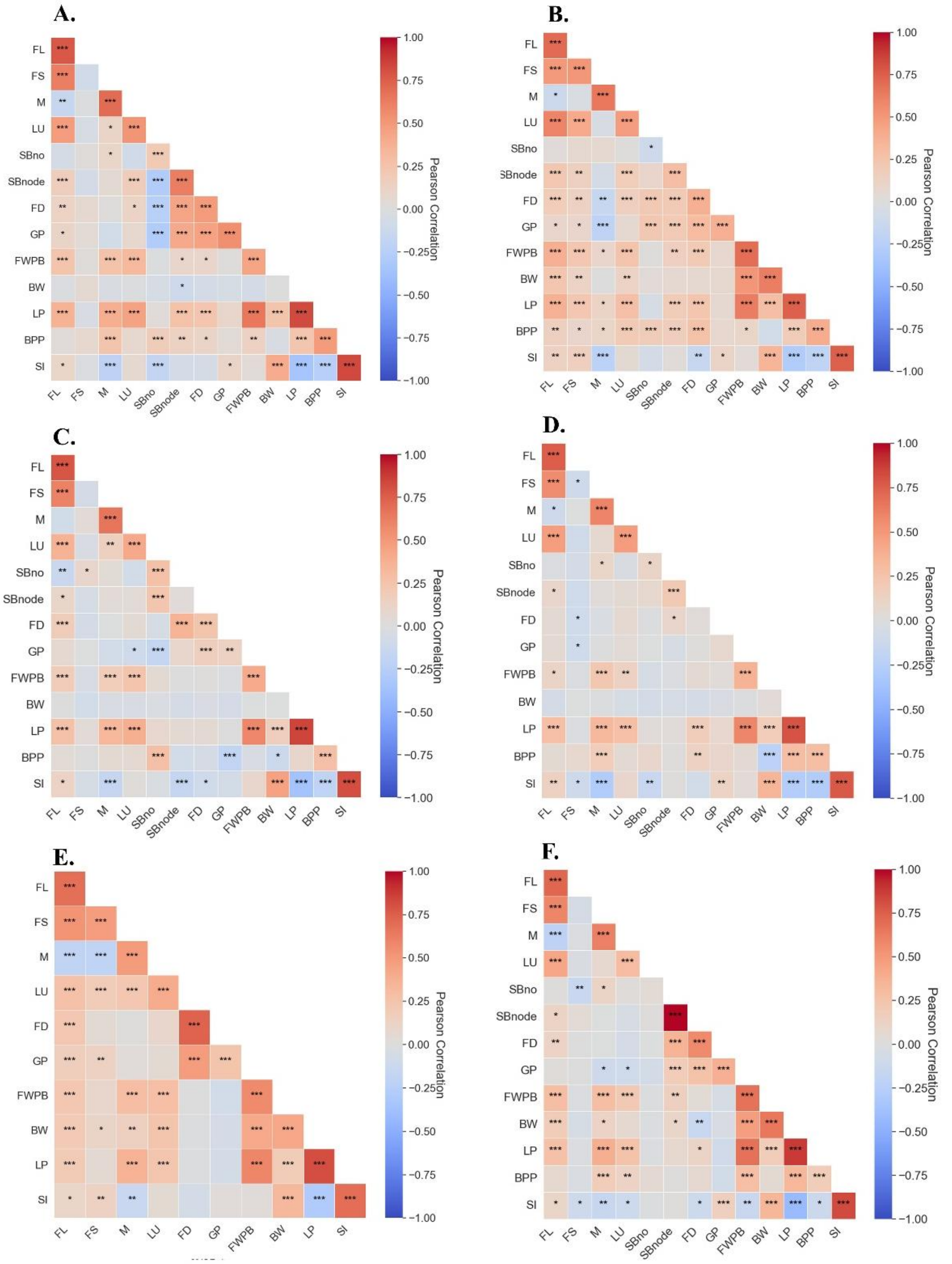
Deep learning-based phenotype prediction analysis of genotype-environment interactions and mining of environmentally stable germplasm and elite loci in cotton

Supplementary Figures

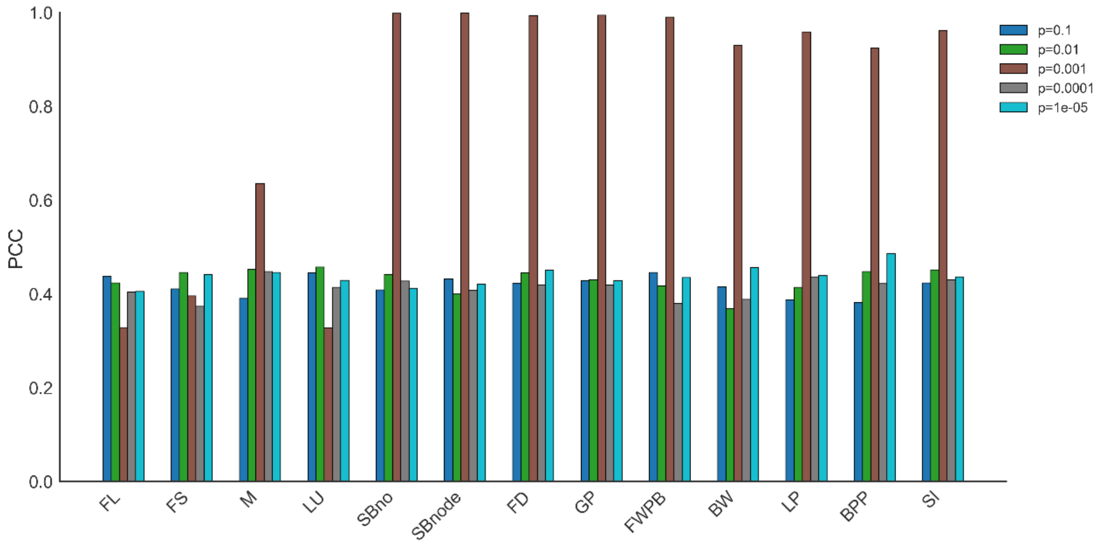


Traits	14-AY	14-CZ	14-DH	14-JZ	14-XJ	14-YC	15-AY	15-CZ	15-DH	15-JZ	15-XJ	15-YC
BPP	18.5	19.5	17.4	22.9		22.2	27.7	20.8	19.6	19.2		21.7
BW	11.6	13.0	12.9	14.1	15.0	13.5	13.6	12.5	11.2	11.0	10.7	14.4
FD	4.4	5.4	3.7	1.1	2.8	3.2	7.5	3.2	3.6	3.8	2.7	4.8
FL	6.2	6.1	5.9	5.9	5.7	5.6	6.2	6.0	6.1	6.6	5.7	5.6
FS	7.4	8.1	7.2	6.9	6.9	7.4	9.2	8.1	7.0	8.2	5.6	8.5
FWPB	15.1	17.5	16.6	17.6	21.2	18.7	20.4	18.8	15.1	17.5	16.2	20.6
GP	8.4	4.0	3.2	2.0	3.8	2.2	7.5	1.8	2.2	2.3	2.6	2.2
LP	11.2	13.1	11.4	11.0	12.9	14.4	12.9	12.6	9.7	12.0	11.5	13.1
LU	1.6	1.5	1.9	1.9	1.3	1.3	1.9	1.6	1.7	1.6	1.4	1.3
M	10.3	9.6	12.8	13.8	12.2	10.2	10.0	8.4	9.8	9.8	10.3	11.4
SBno	15.5	23.8	11.0	9.3		14.6	9.8	14.7	13.4	10.7		8.7
SBnode	10.5	16.6	9.7	10.2		14.3	21.0	10.4	8.6	9.0		14.3
SI	10.9	9.9	12.3	11.2	11.5	12.2	11.7	9.9	10.6	10.5	11.1	11.8

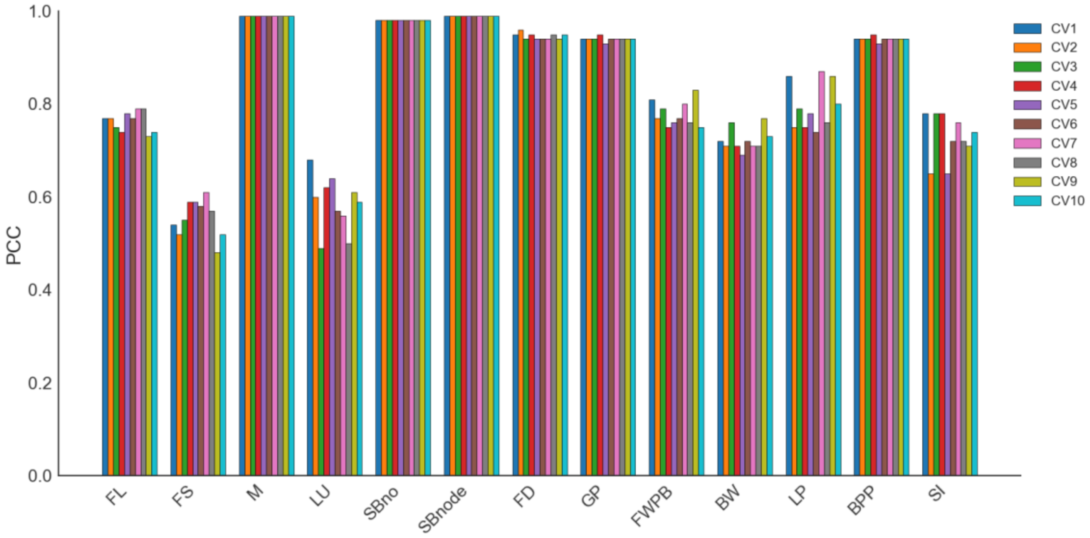
Supplementary Fig. 1. Phenotype diversity and population structure analysis of 419 upland cotton. (A) Phenotype distribution of fiber quality-related traits. (B) Phenotype distribution of growth period-related traits. (C) Phenotype distribution of fiber yield-related traits. (D) Population structure analysis of 419 upland cotton. (E) Heat map of coefficients of variation of 13 traits in 12 environments for 419 upland cotton. (F) Analysis of variance components (ANOVA) of 13 traits of 419 upland cotton. (G) Heritability of 13 traits of 419 upland cotton.



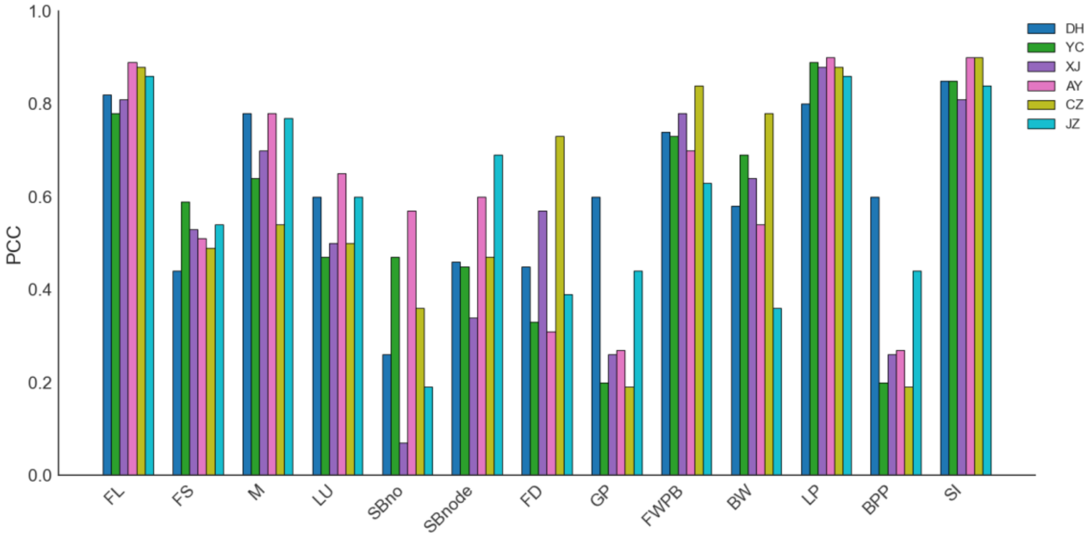
Supplementary Fig. 2. Correlation heatmap of 13 traits of 419 upland cotton in 6 locations. (A~F) Correlation heatmap of 13 traits of 419 upland cotton in AY, CZ, DH, JZ, XJ, YC, respectively. Traits of 2014 are shown at the bottom of each heatmap, and those for 2015 are shown on the left side. Phenotype value of traits SBno, SBnode and BPP missing in Xinjiang. *P-values<0.05, **P-values<0.01, ***P-values<0.001.



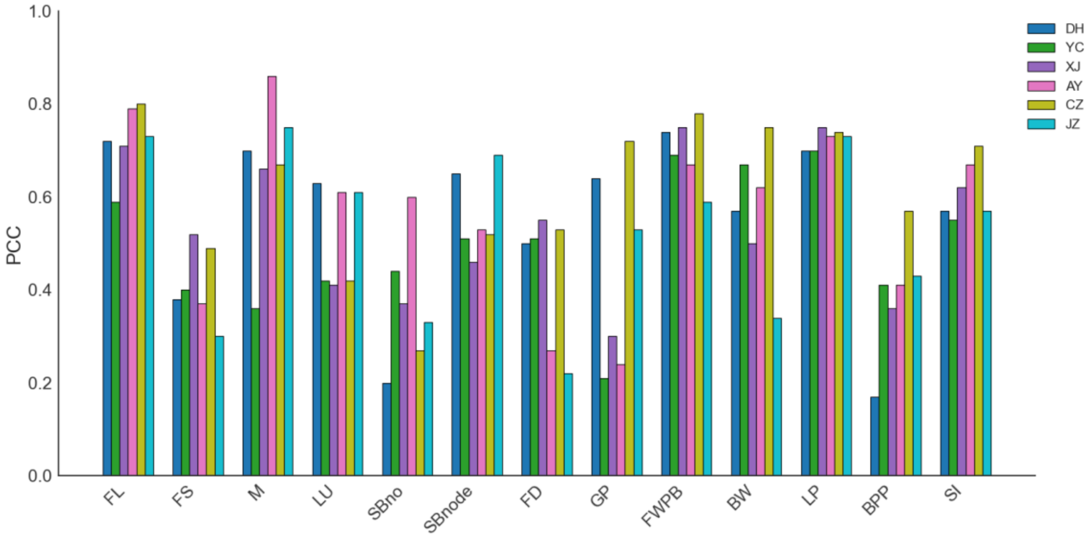
Supplementary Fig. 3. Test results of AttGEI-Net model prediction of 13 trait phenotype values based on different p-values. (p=0.1, 0.01, 0.001, 0.0001, 1e-05)



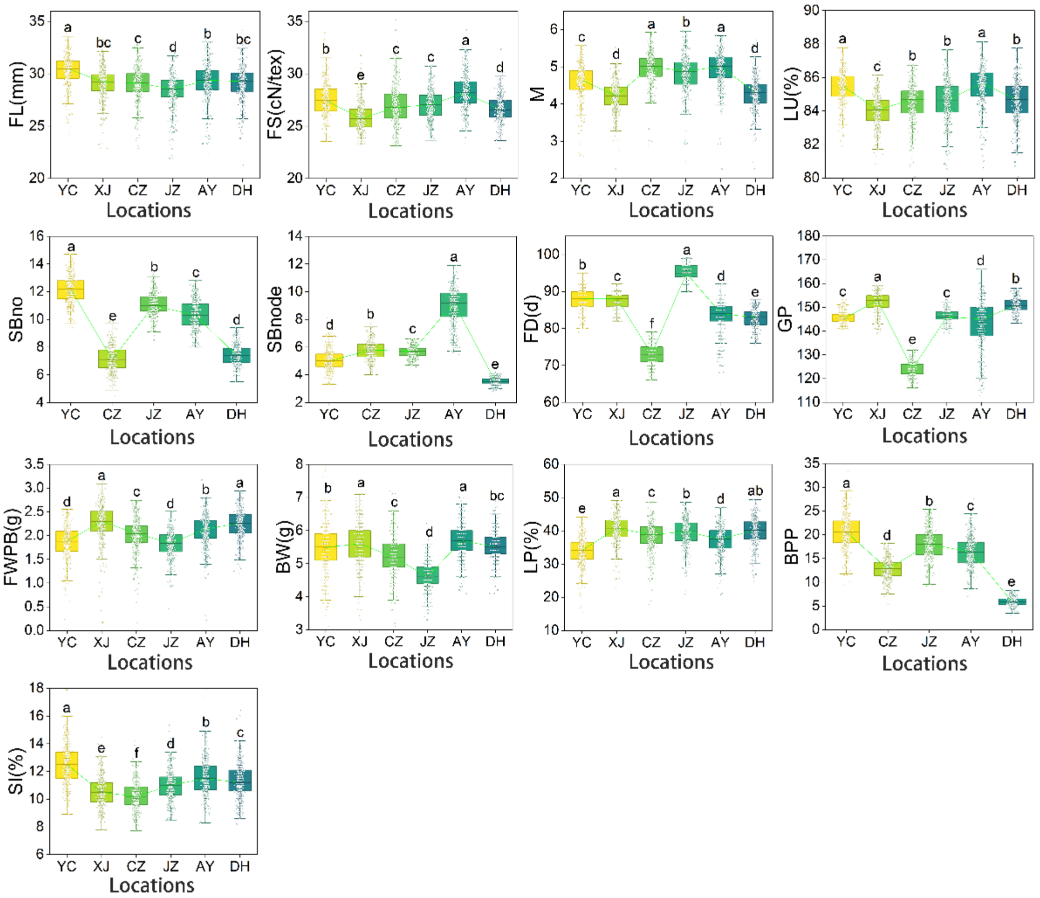
Supplementary Fig. 4. Prediction accuracy of AttGEI-Net model in scenario 1. The different colored columns represent each of the ten cross-validation.



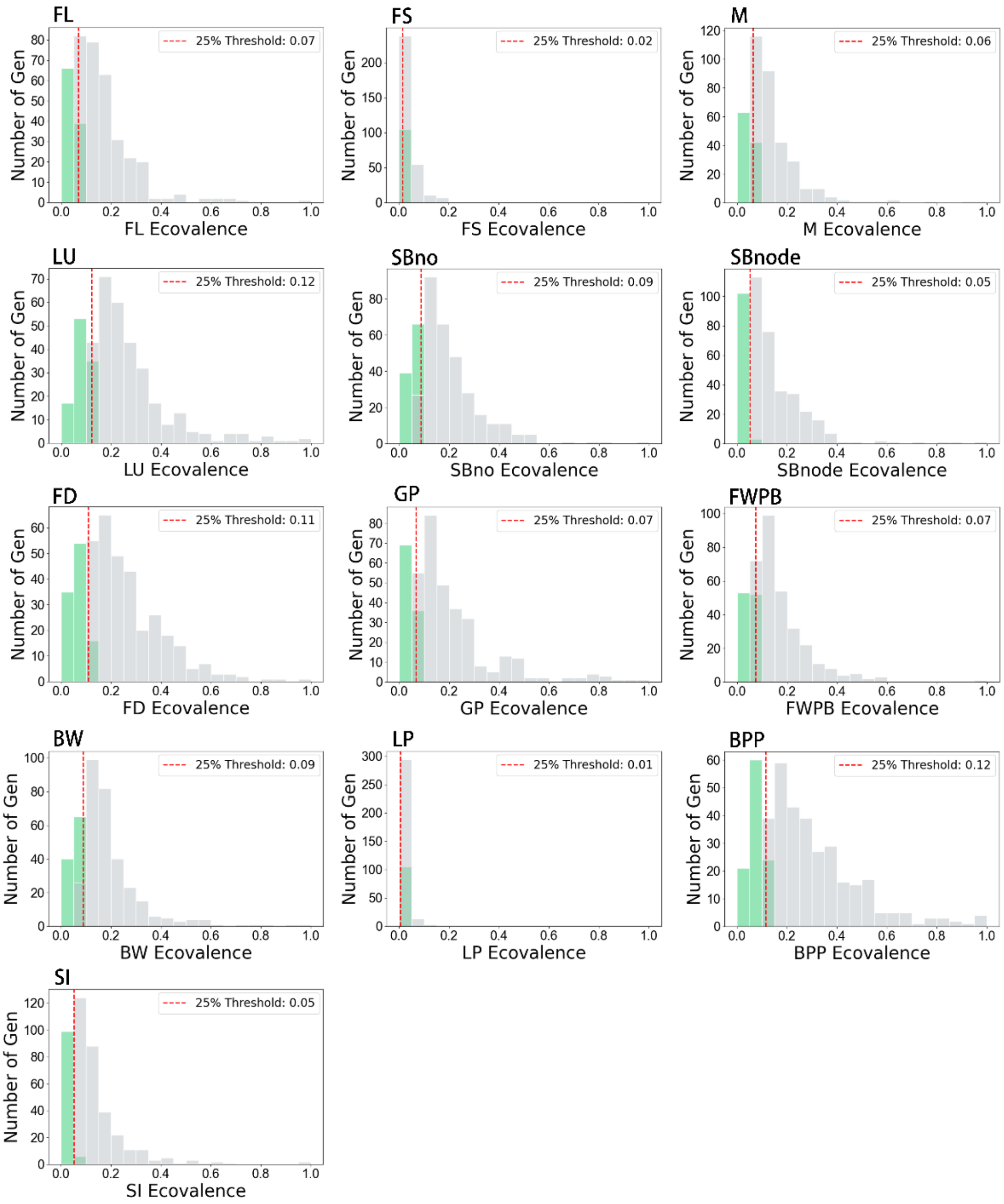
Supplementary Fig. 5. Prediction accuracy of AttGEI-Net model in scenario 2. The different colored columns represent different locations.



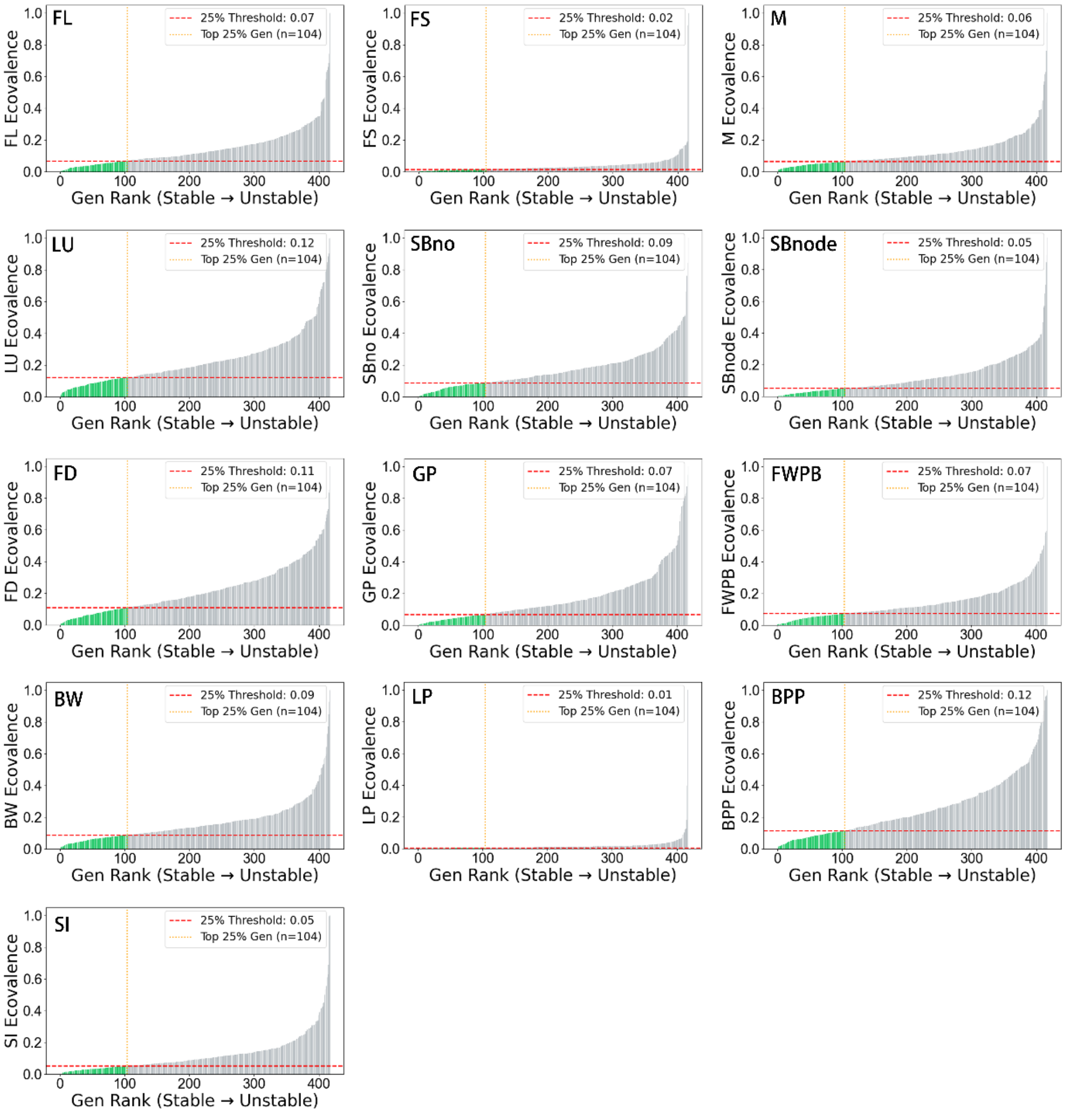
Supplementary Fig. 6. Prediction accuracy of AttGEI-Net model in scenario 3. The different colored columns represent different locations.



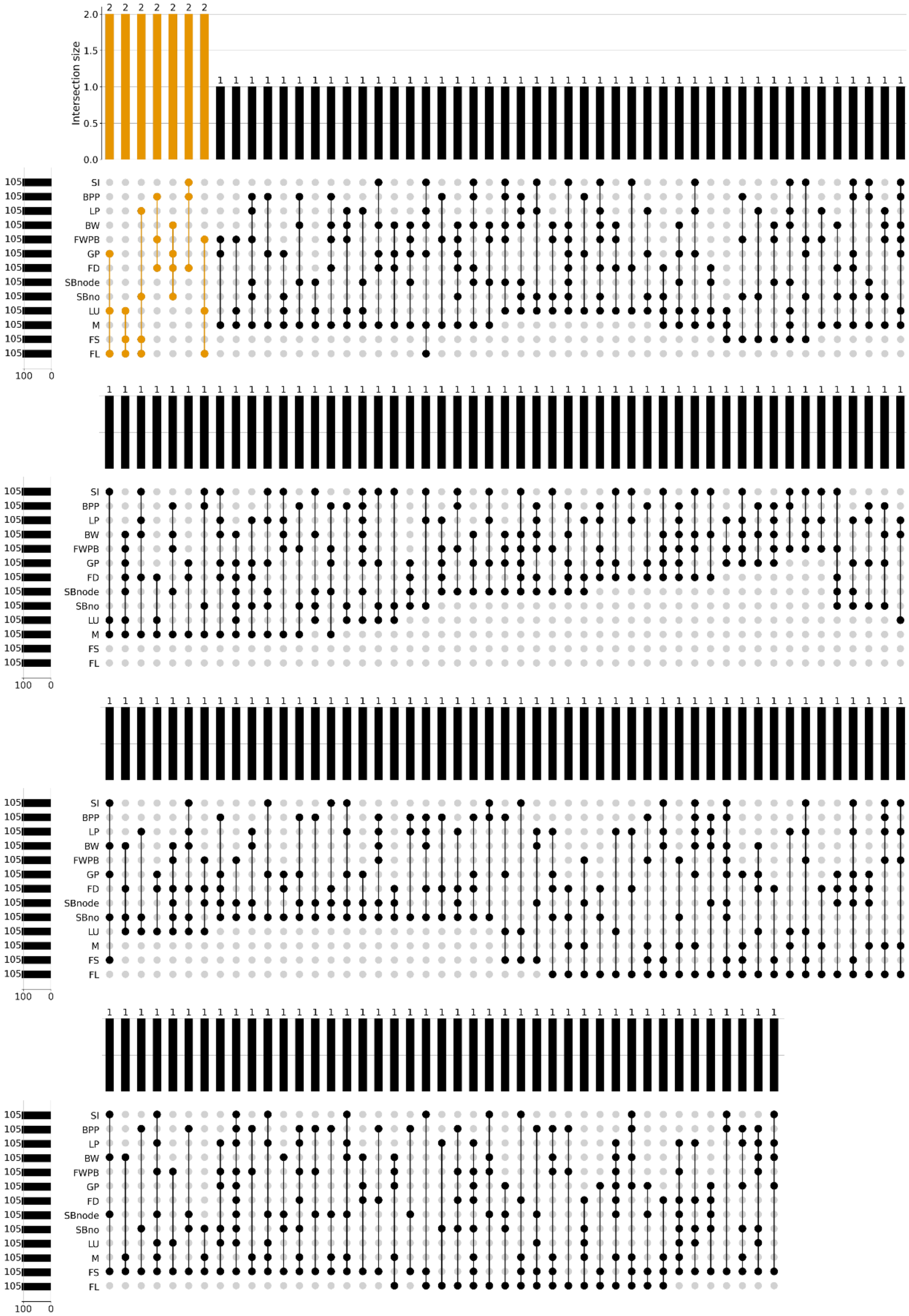
Supplementary Fig. 7. Phenotype plasticity analysis of 13 traits of 419 upland cotton in 6 locations. *P-values<0.05.



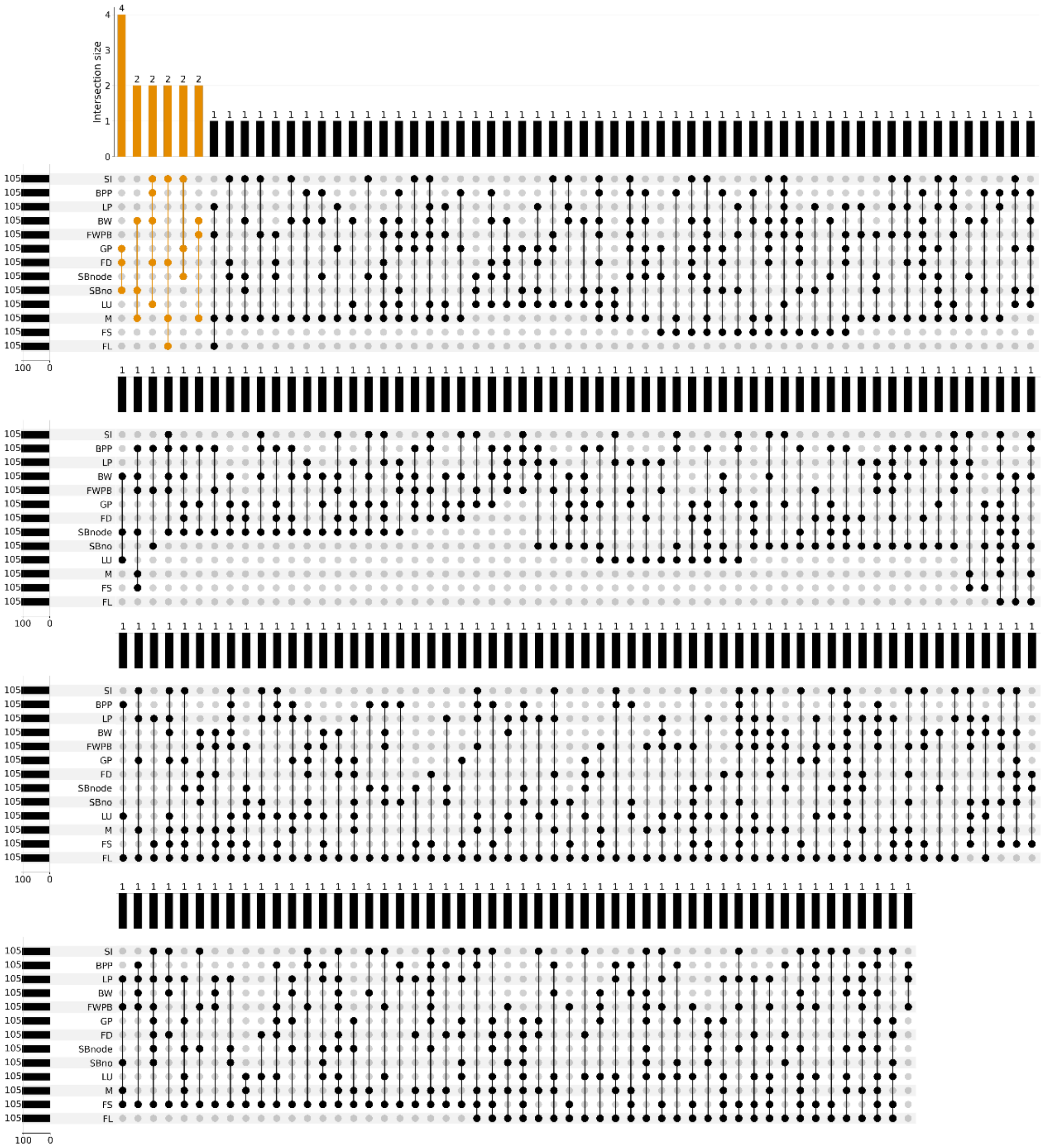
Supplementary Fig. 8. Distribution of wricke's ecovalence of 13 traits of 419 upland cotton. The red dotted line represents the threshold line of wricke's ecovalence, and the green area represents the stable varieties.



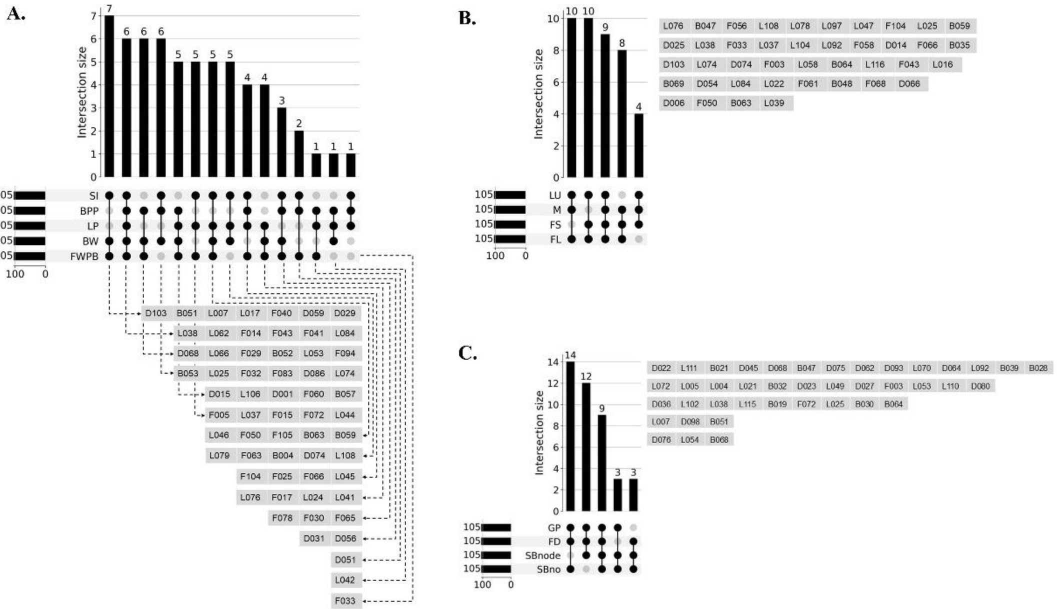
Supplementary Fig. 9. Ranking stable varieties of 13 traits of 419 upland cotton based on wricke's ecovalence. The red dotted line represents the threshold line of wricke's ecovalence, and the green area where it intersects the yellow dotted line represents the stable varieties.



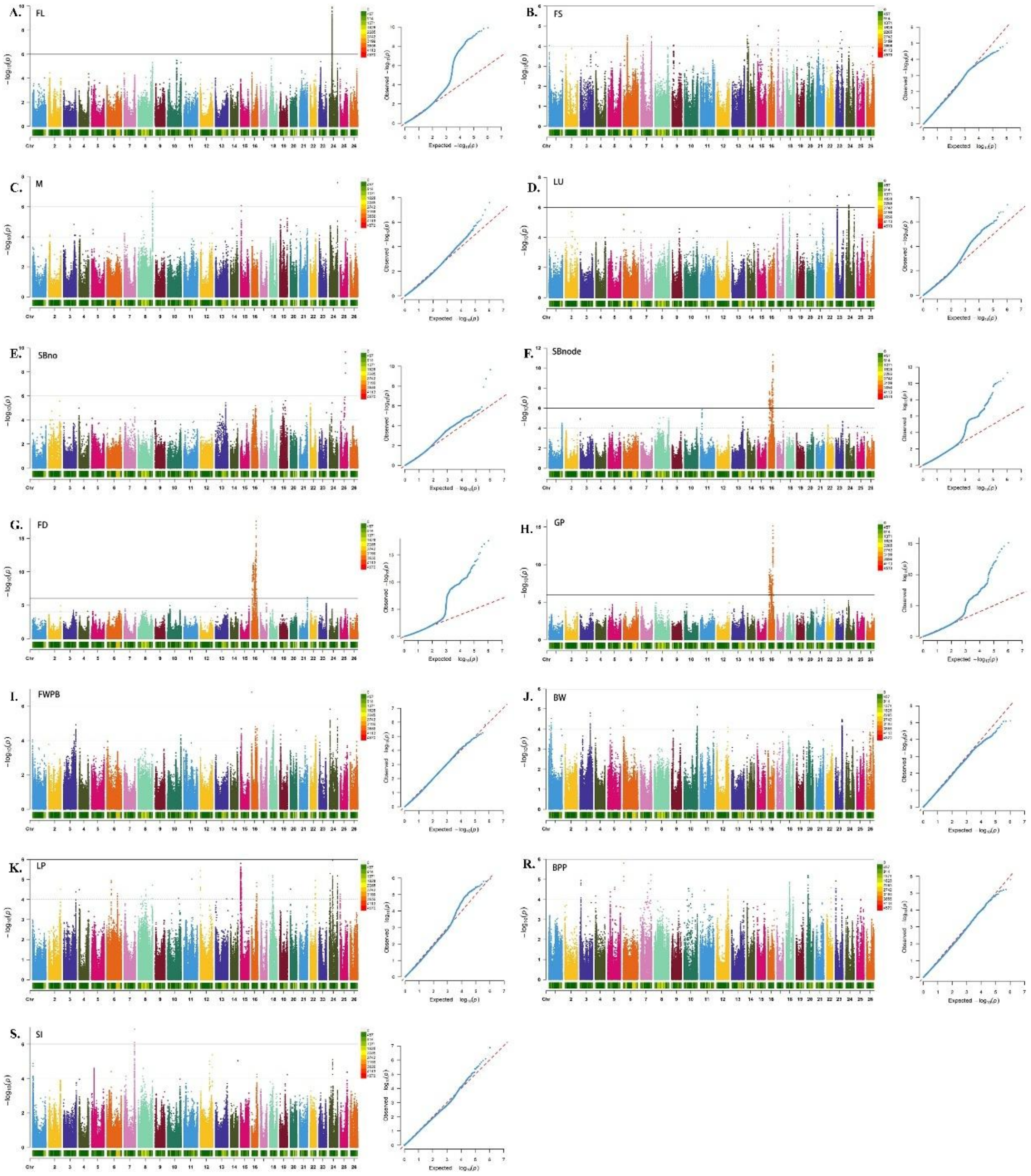
Supplementary Fig. 10. Upset plot of multi-trait, multi-environmentally stable varieties of 13 traits of 419 upland cotton. The yellow areas represent combinations with a high number of stable varieties of multiple trait intersections.



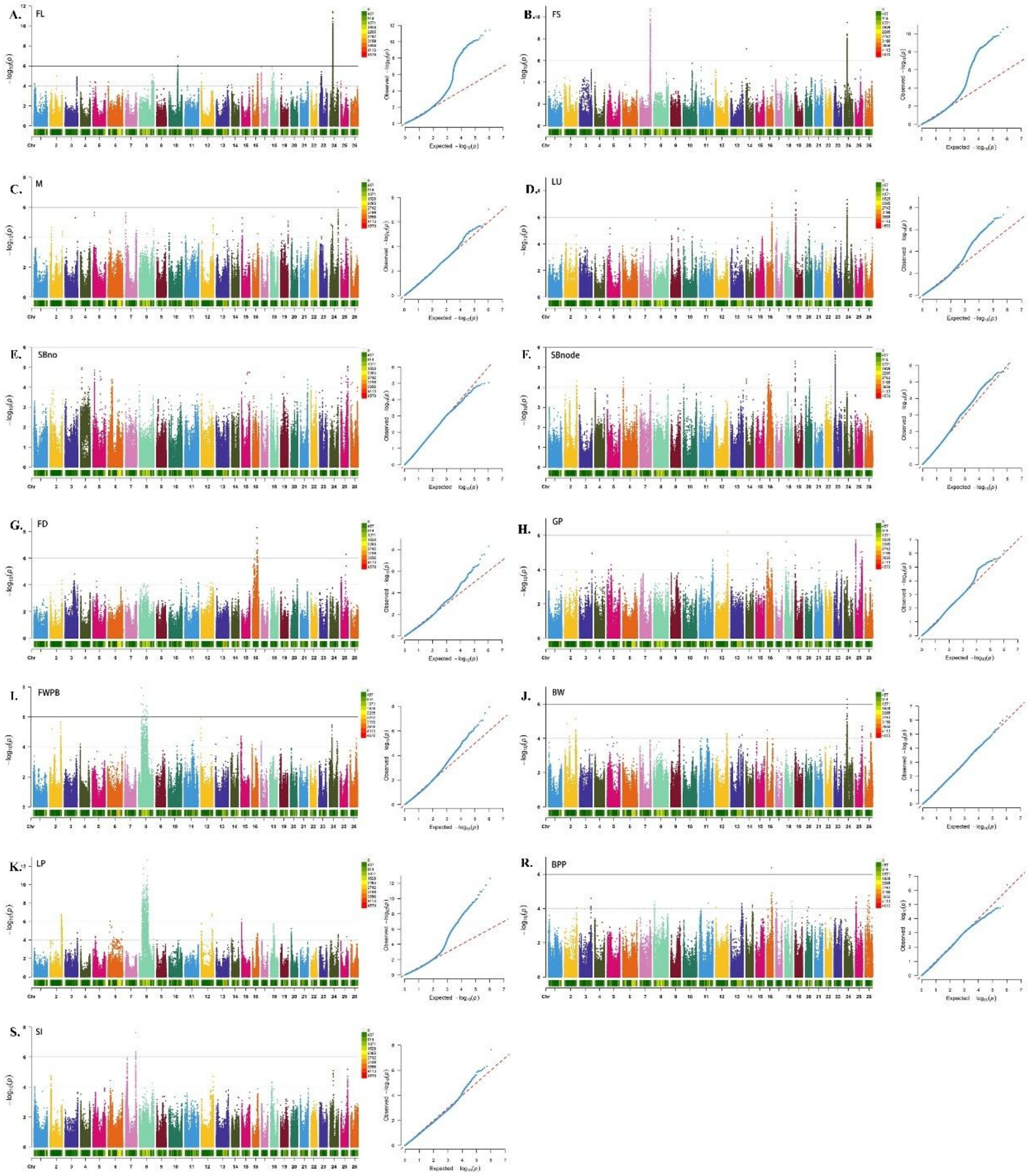
Supplementary Fig. 11. Upset plot of multi-trait, multi-environmentally unstable varieties of 13 traits of 419 upland cotton. The yellow areas represent combinations with a high number of unstable varieties of multiple trait intersection.



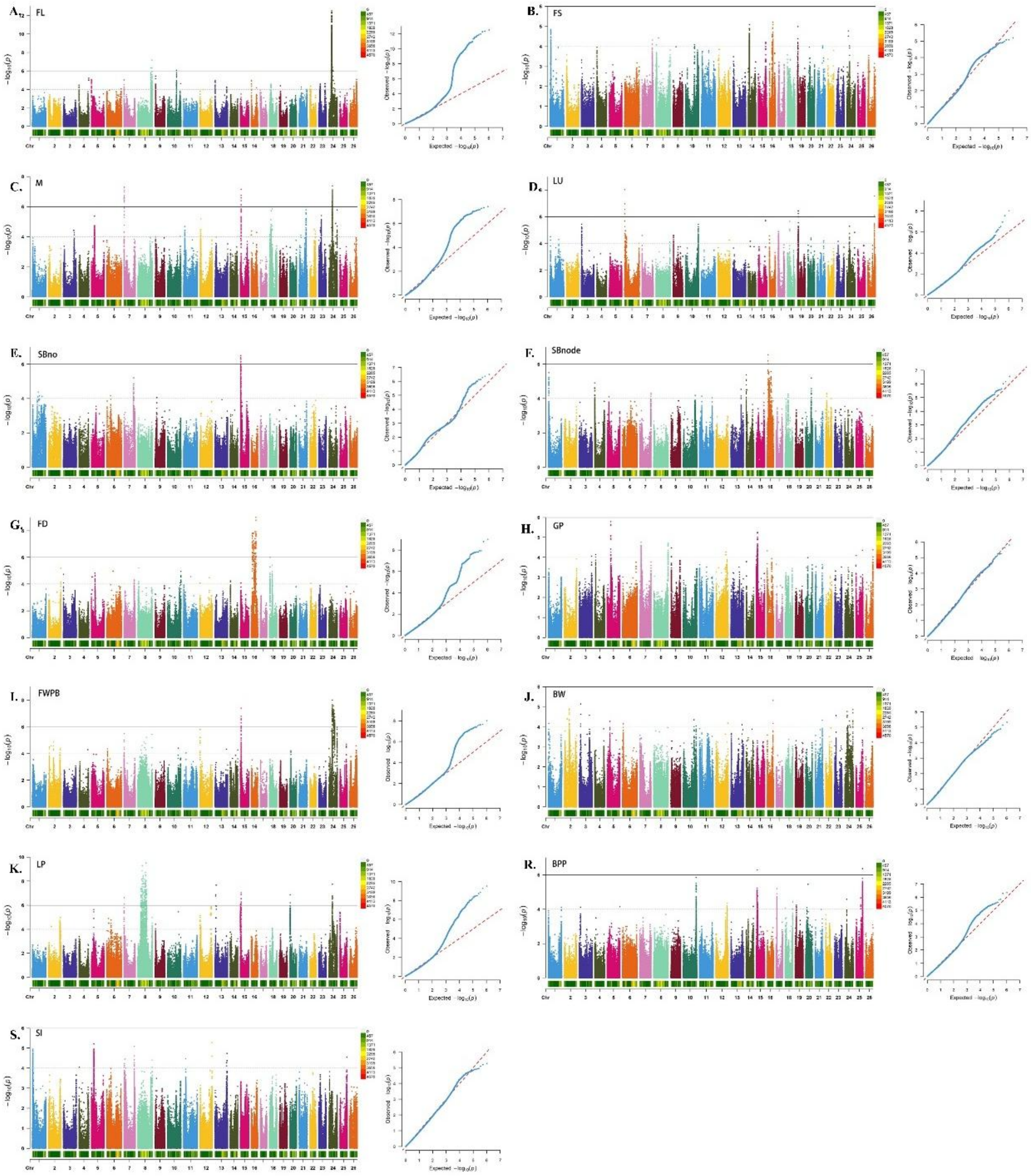
Supplementary Fig. 12. Upset plot of multi-trait, multi-environmentally unstable varieties of 3 trait categories of 419 upland cotton. (A) Multi-environmentally unstable varieties of fiber yield-related traits. Each set of intersections is labelled with the unstable variety of fiber yield-related traits. **(B) Multi-environmentally unstable varieties of fiber quality-related traits.** Each set of intersections is labelled with the unstable variety of fiber quality-related traits. **(C) Multi-environmentally unstable varieties of growth period-related traits.** Each set of intersections is labelled with the unstable variety of growth period-related traits.



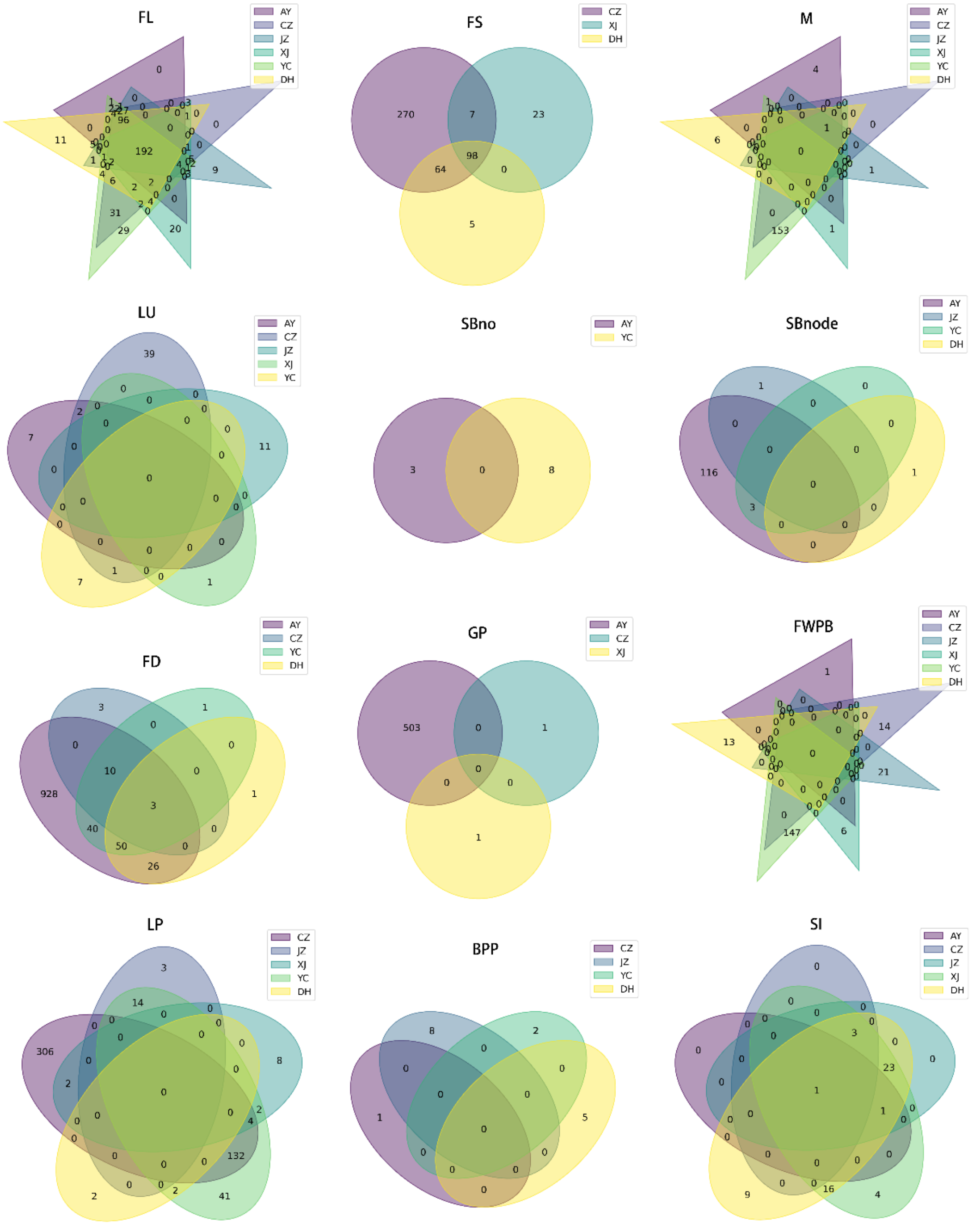
Supplementary Fig. 13. Genome-wide association study of 13 traits of 419 upland cotton using phenotype value from AY. (A~S) Manhattan plots and quantile-quantile plots of 13 traits of 419 upland cotton. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index, AY: Anyang.



Supplementary Fig. 14. Genome-wide association study of 13 traits of 419 upland cotton using phenotype value from CZ. (A~S) Manhattan plots and quantile-quantile plots of 13 traits of 419 upland cotton. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index, CZ: Cangzhou.

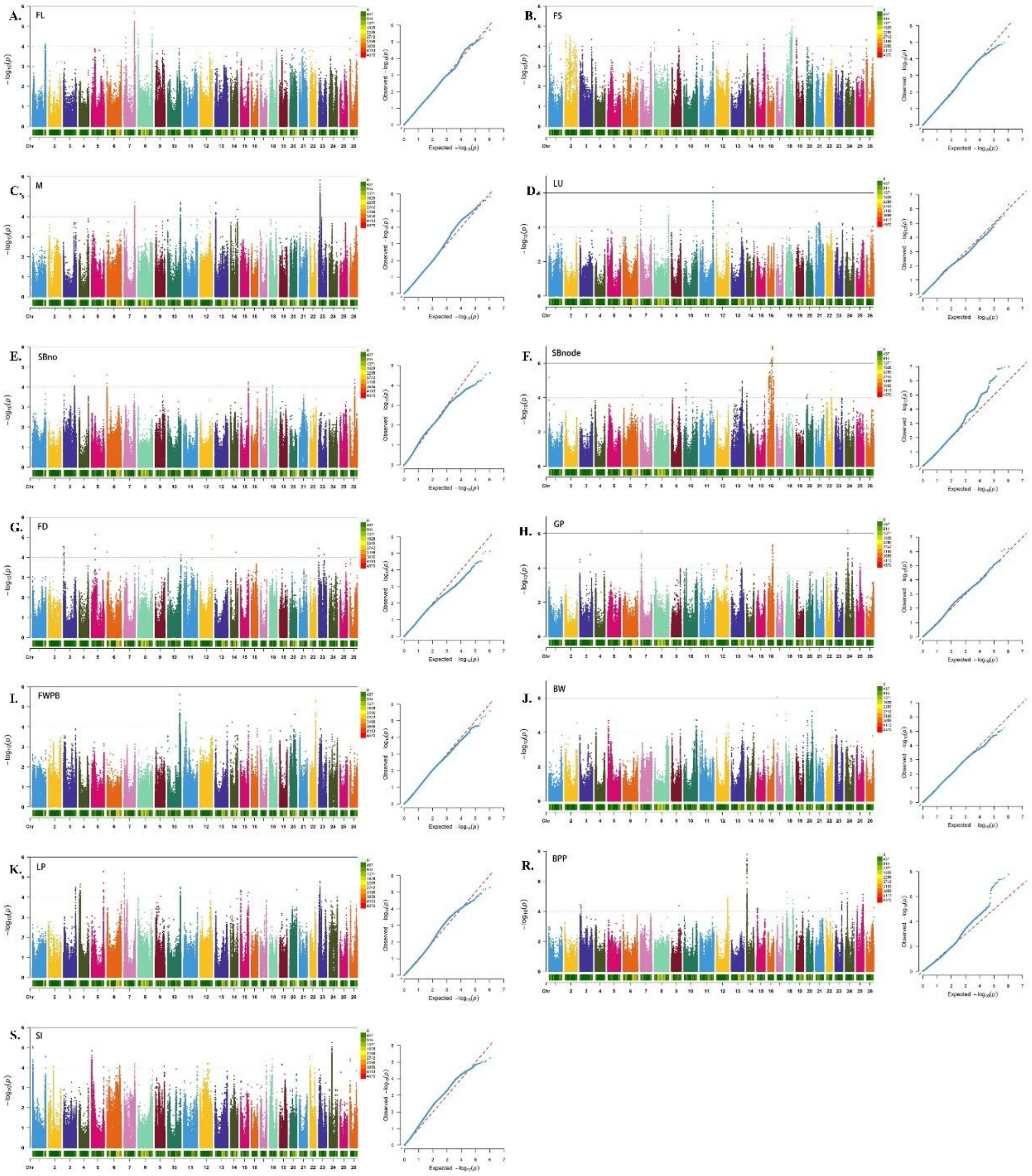


Supplementary Fig. 15. Genome-wide association study of 13 traits of 419 upland cotton using phenotype value from YC. (A~S) Manhattan plots and quantile-quantile plots of 13 traits of 419 upland cotton. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index, YC: Yancheng.

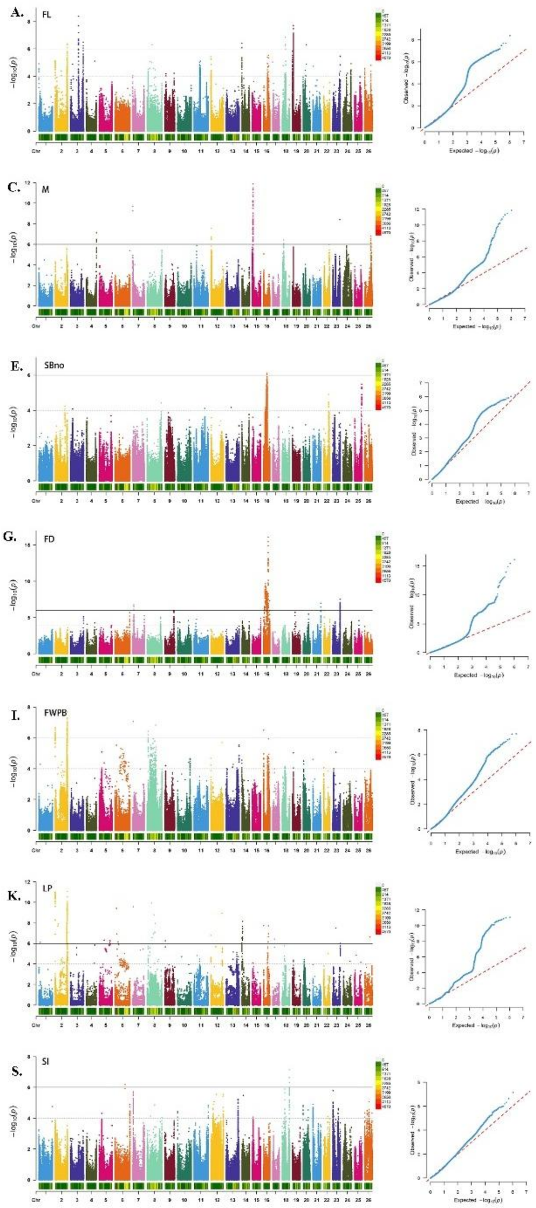


Supplementary Fig. 16. Venn diagram of significantly associated Main-SNPs at six locations identified by GWAS for 13 traits of 419 upland cotton. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index, AY: Anyang, CZ: Cangzhou, JZ: Jingzhou, XJ: Xinjiang, DH: Dunhuang. (BW was only detected significant loci at one location and could not be plotted in a Venn diagram, details of which can be found in Tables

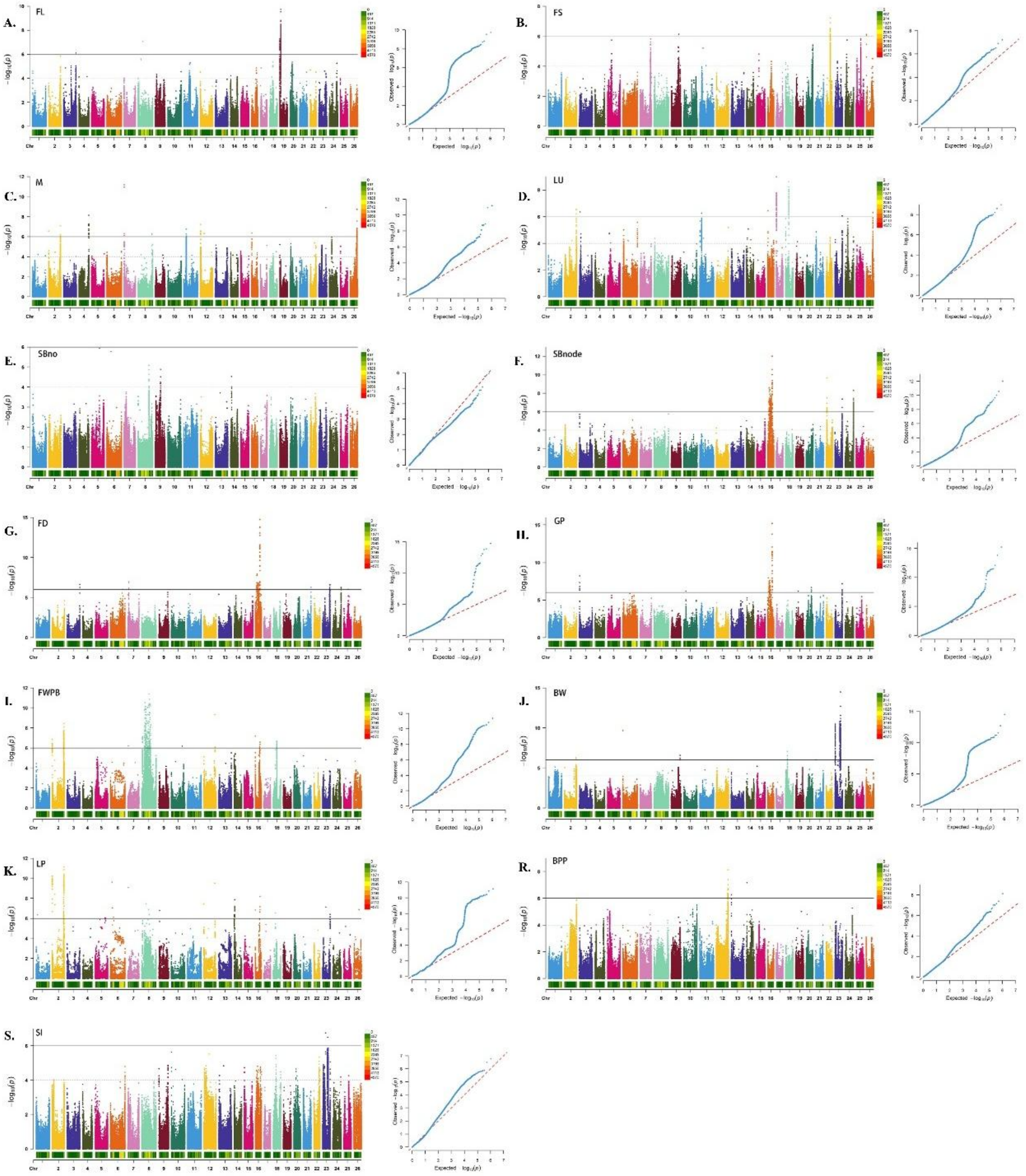
S4, so only Venn diagrams are shown for 12 traits)



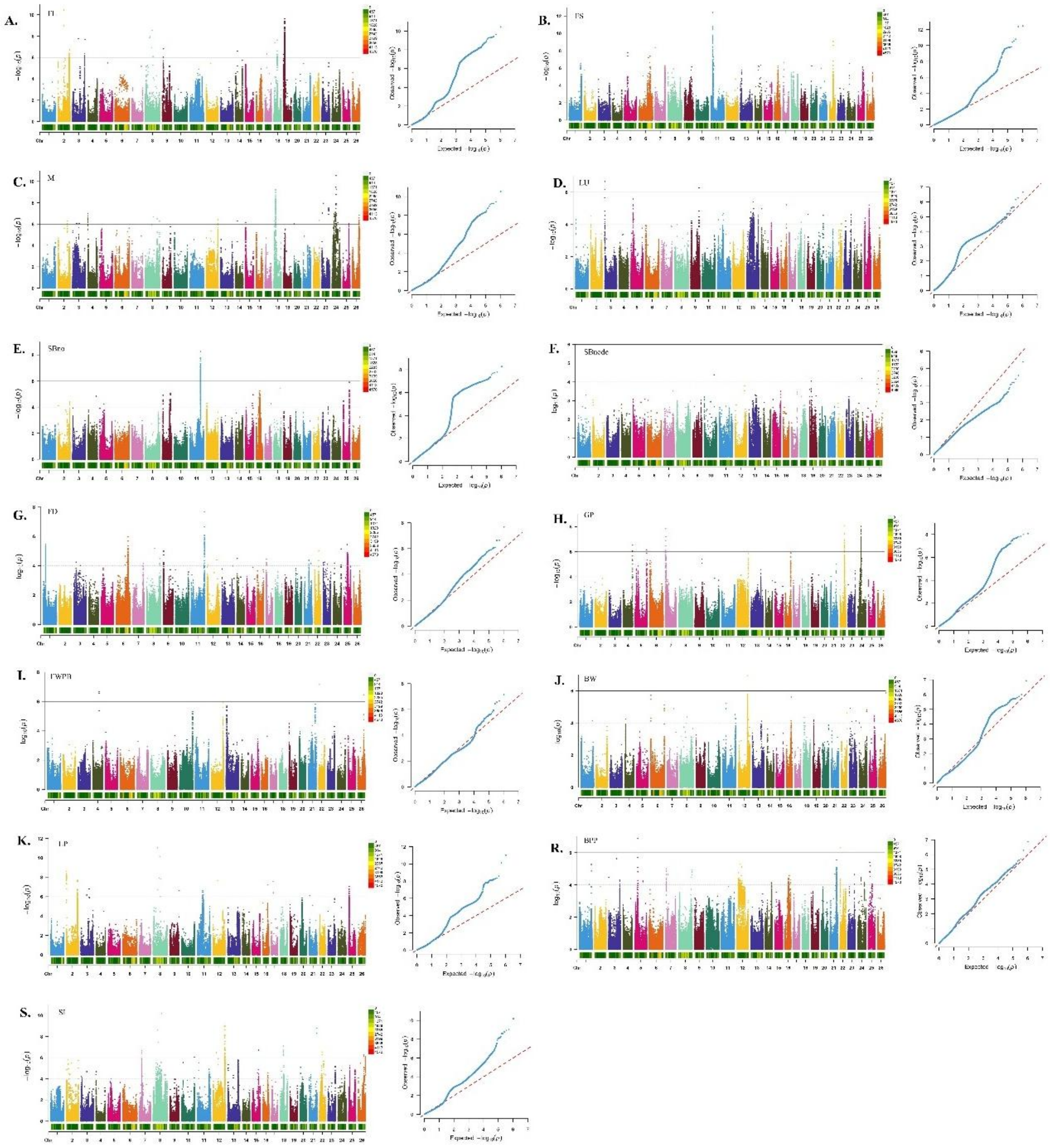
Supplementary Fig. 17. Genome-wide association study of 13 traits in 419 upland cotton using phenotype plasticity parameters. (A~S) Manhattan plots and quantile-quantile plots of different traits. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index.



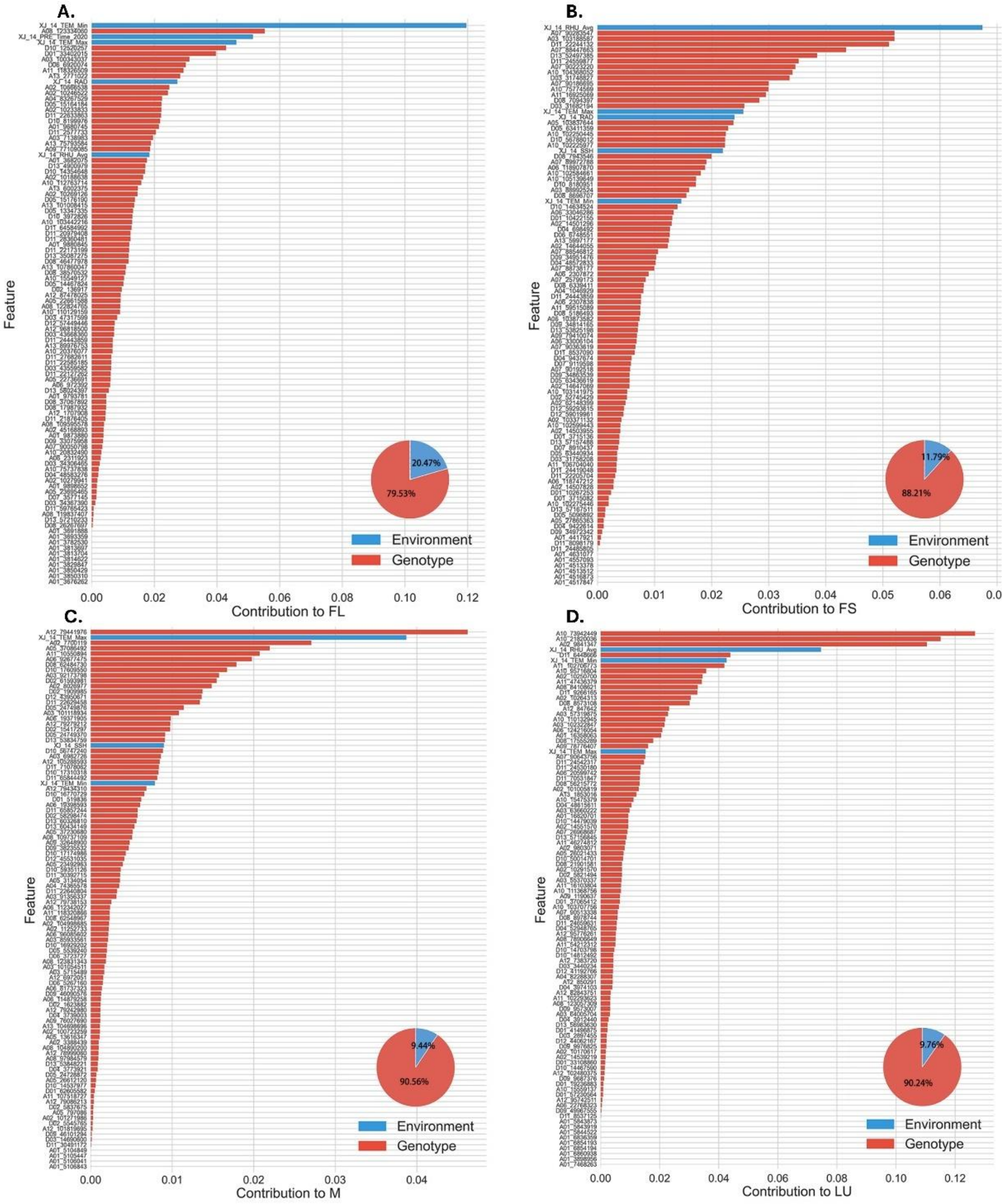
Supplementary Fig. 18. Genome-wide association study of 13 traits in 419 upland cotton using the environmental adaptation parameters of 2014 and 2015. (A~S) Manhattan plots and quantile-quantile plots of different traits. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index.



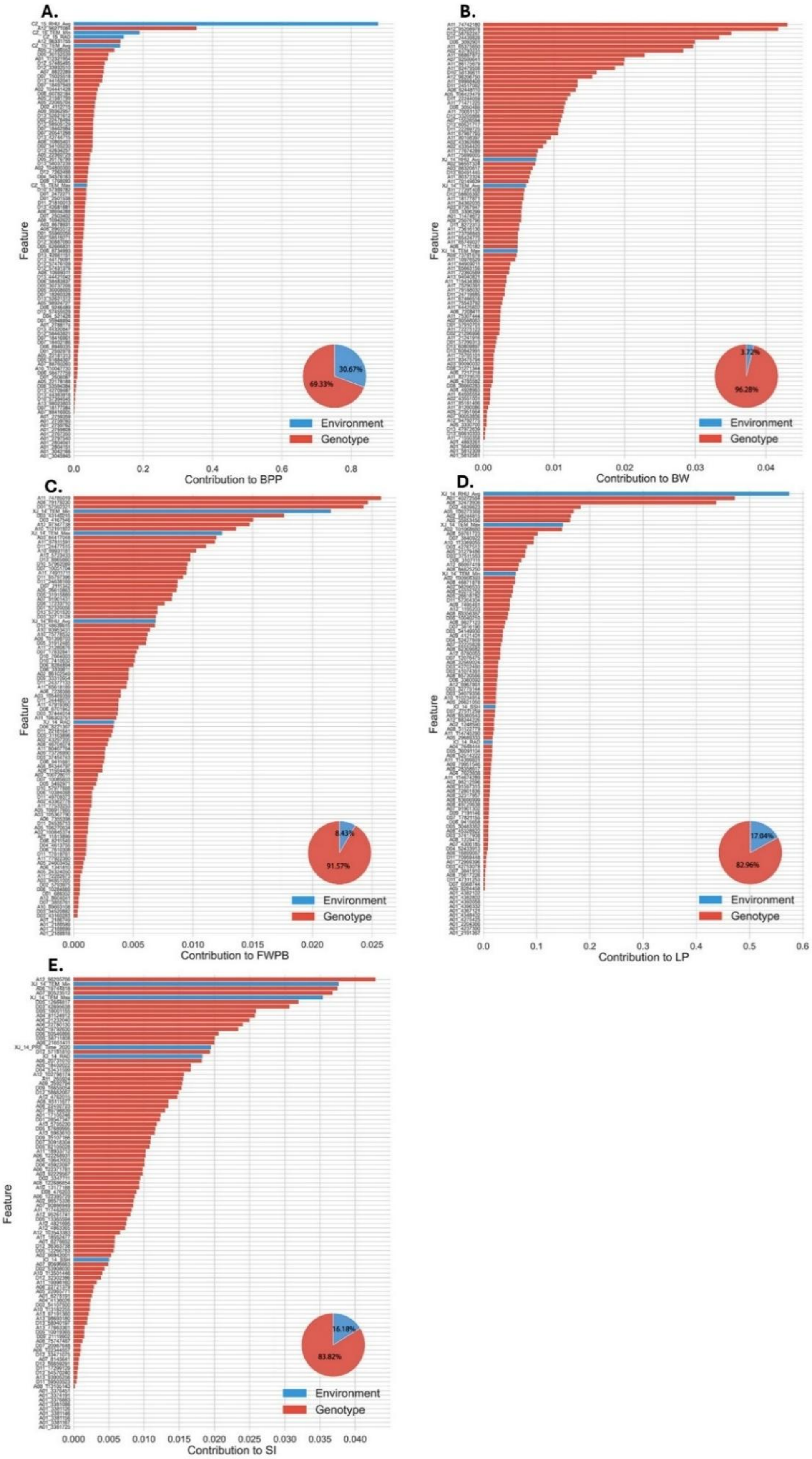
Supplementary Fig. 19. Genome-wide association study of 13 traits in 419 upland cotton using the environmental adaptation parameters of 2015. (A~S) Manhattan plots and quantile-quantile plots of different traits. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index.



Supplementary Fig. 20. Genome-wide association study of 13 traits in 419 upland cotton using the environmental adaptation parameters of YC. (A~S) Manhattan plots and quantile-quantile plots of different traits. The genome-wide significant P-value threshold (10^{-6}) is indicated by the black line. FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity, SBno: sympodial branch number, SBnode: the first Sympodial branch node, FD: flowering date, GP: growth period, FWPB: fiber weight per boll, BW: boll weight, LP: lint percentage, BPP: bolls per plant, SI: seed index. YC: Yancheng.



Supplementary Fig. 21. Feature importance ranking of fiber quality-related traits calculated using the interpretability analysis of the AttGEI-Net model. Horizontal coordinates represent the contribution value of the features to traits, vertical coordinates list the names of the top 100 important features, blue represents the environmental features (PRE, RHU, RAD, SSH, TEM), red represents the genetic features (SNPs loci), and the pie chart in the bottom right corner indicates the percentage of importance of the two features. PRE: Precipitation, RHU: Relative humidity, RAD: Radiation, SSH: Sunshine, TEM: Temperature. **(A~D) Interpretability analysis of AttGEI-Net models for FL, FS, M, LU.** FL: fiber length, FS: fiber strength, M: micronaire, LU: length uniformity.


A.
Feature
30.67%
69.33%
Environment
Genotype
0.0
0.2
0.4
0.6
0.8
Contribution to BPP
B.
Feature
3.72%
96.28%
Environment
Genotype
0.00
0.01
0.02
0.03
0.04
Contribution to BW
C.
Feature
8.43%
91.57%
Environment
Genotype
0.000
0.005
0.010
0.015
0.020
0.025
Contribution to FWPB
D.
Feature
17.04%
82.96%
Environment
Genotype
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Contribution to LP
E.
Feature
16.18%
83.82%
Environment
Genotype
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
Contribution to SI

Supplementary Fig. 22. Feature importance ranking of fiber yield-related traits calculated using the interpretability analysis of the AttGEI-Net model. Horizontal coordinates represent the contribution value of the features to traits, vertical coordinates list the names of the top 100 importance features, blue represents the environmental features (PRE, RHU, RAD, SSH, TEM), red represents the genetic features (SNPs loci), and the pie chart in the bottom right corner indicates the percentage of importance of the two features. PRE: Precipitation, RHU: Relative humidity, RAD: Radiation, SSH: Sunshine, TEM: Temperature. **(A~E) Interpretability analysis of AttGEI-Net models for BPP, BW, FWPB, LP, SI.** BPP: bolls per plant, BW: boll weight, FWPB: fiber weight per boll, LP: lint percentage, SI: seed index.

Supplementary Tables

Supplementary Table 1. The abbreviation and description of 13 traits of upland cotton.

Numbers	Trait Categories	Traits	Abbreviation
1	Fiber quality-related traits	Fiber length(mm)	FL
2		Fiber strength(cN/tex)	FS
3		Micronaire	M
4		Length uniformity(%)	LU
5	Growth period-related traits	Sympodial branch number	SBno
6		The first Sympodial branch node	SBnode
7		Flowering date(d)	FD
8		Growth period(d)	GP
9	Fiber yield-related traits	Fiber weight per boll(g)	FWPB
10		Boll weight(g)	BW
11		Lint percentage(%)	LP
12		Bolls per plant	BPP
13		Seed index(%)	SI

Supplementary Table 2. Mean phenotype value and cross-environment coefficients of variation (CV) of 13 traits of 419 upland cotton in 12 environments.

Trait categories	Traits	Cross-environment CV (%)	Mean phenotype value	Min phenotype value	Max phenotype value	Number of environments
Fiber quality-related traits	FL(mm)	6.77±0.27	29.18±1.97	18.84	35.75	12
	FS(cN/tex)	8.43±0.90	27.06±2.28	22.47	41.26	12
	M	14.32±1.48	4.63±0.66	2.06	6.58	12
	LU(%)	1.83±0.22	84.74±1.55	77.15	89.20	12
Growth period-related traits	FD(d)	11.27±1.53	84.81±9.56	60.00	108.00	12
	GP(d)	10.26±2.11	143.71±14.75	111.00	178.00	12
	SBno	25.68±4.24	9.63±2.47	3.30	16.80	10
	SBnode	35.48±3.80	5.79±2.05	2.00	13.70	10
Fiber yield-related traits	FWPB(g)	22.24±1.97	2.05±0.46	0.08	3.73	12
	BPP	44.09±2.77	14.75±6.50	2.40	35.70	10
	BW(g)	16.70±1.36	5.36±0.90	2.40	9.00	12
	LP(%)	13.881.20	38.12±5.29	2.50	55.00	12
	SI(%)	13.19±0.76	11.31±1.49	6.70	20.30	12

Supplementary Table 3. The prediction accuracy of the AttGEI-Net model for 13 traits of 419 upland cotton in three scenarios.

Traits	Scenario		
	Scenario1	Scenario2	Scenario3
FL	0.76±0.01	0.84±0.02	0.72±0.03
FS	0.55±0.01	0.52±0.02	0.41±0.03
M	0.99±0.00	0.70±0.04	0.67±0.07
LU	0.59±0.02	0.55±0.03	0.52±0.04
SBno	0.98±0.00	0.32±0.08	0.37±0.06
SBnode	0.99±0.00	0.50±0.05	0.56±0.04
FD	0.95±0.00	0.46±0.07	0.43±0.06
GP	0.94±0.00	0.33±0.07	0.44±0.09
FWPB	0.78±0.01	0.74±0.03	0.70±0.03
BW	0.72±0.01	0.60±0.06	0.57±0.06
LP	0.80±0.02	0.87±0.01	0.72±0.01
BPP	0.94±0.00	0.33±0.07	0.39±0.05
SI	0.73±0.02	0.86±0.01	0.61±0.03

Supplementary Table 4. The numbers of Main-SNPs associated with 13 traits of 419 upland cotton using phenotype values from six locations for GWAS. Six locations are Anyang(AY), Cangzhou(CZ), Dunhuang(DH), Jinzhou(JZ), Xinjiang(XJ), Yancheng(YC), respectively.

Trait categories	Traits	Numbers		AY	CZ	DH	JZ	XJ	YC
Fiber quality related-traits	FL	902	492	349	400	331	345	243	438
	FS		467	0	439	167	0	128	0
	M		167	6	1	7	2	2	154
	LU		68	9	42	0	11	1	8
Growth period related-traits	SBno	1,076	11	3	0	0	0	/	8
	SBnode		121	119	0	1	1	/	3
	FD		1,062	1,057	16	80	0	0	104
	GP		505	503	1	0	0	1	0
Fiber yield related-traits	FWPB	752	202	1	14	13	21	6	147
	BW		1	0	1	0	0	0	0
	LP		516	0	444	4	17	16	195
	BPP		16	0	1	5	8	/	2
	SI		57	2	4	53	28	48	0
Summary		2,705							