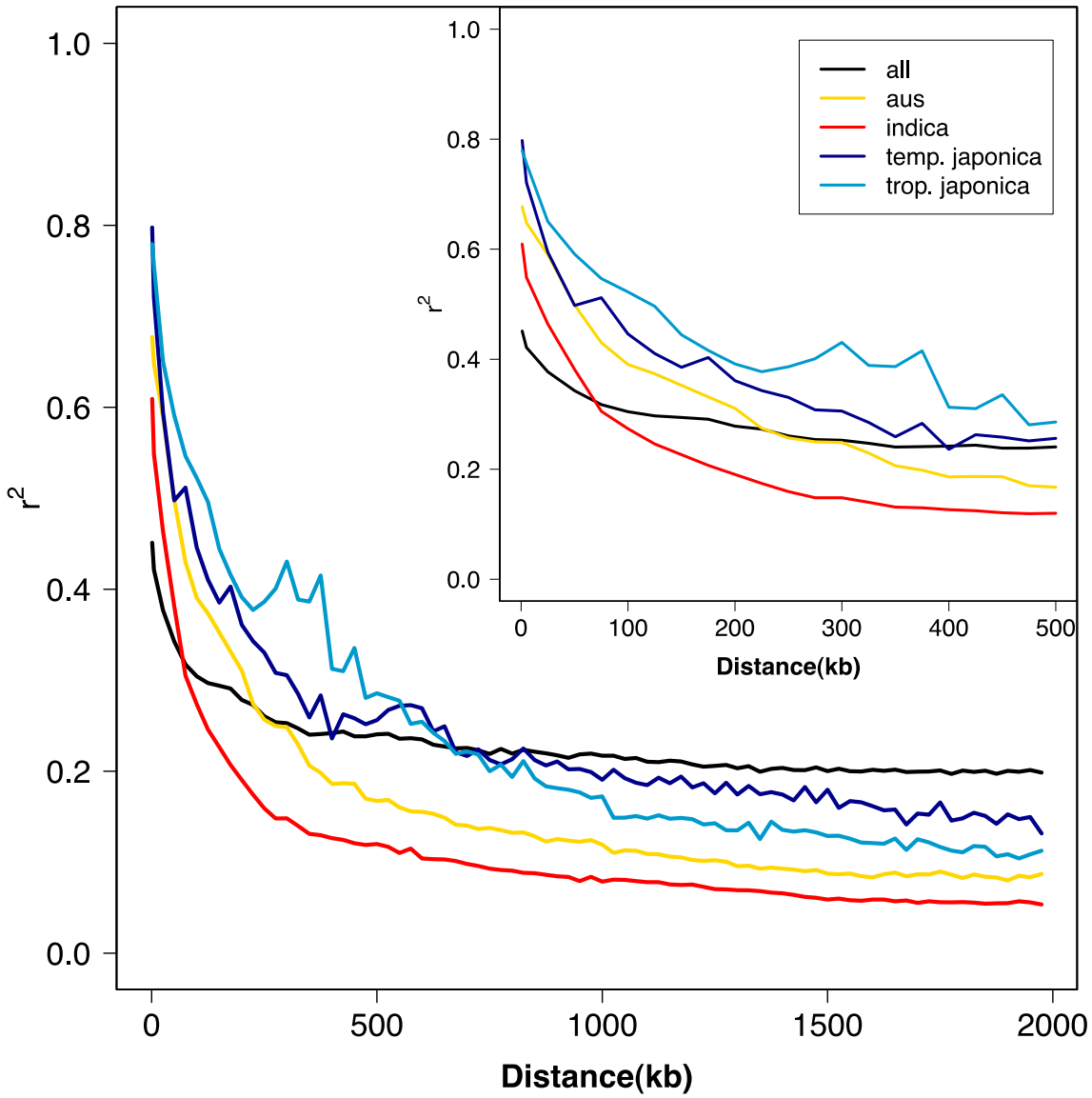
Genome-wide association mapping reveals rich genetic architecture of complex traits in *Oryza sativa*

Keyan Zhao, Chih-Wei Tung, Georgia C. Eizenga, Mark H. Wright, Md. Liakat Ali, Adam H. Price, Gareth J. Norton, M. Rafiqul Islam, Andy Reynolds, Jason Mezey, Anna M. McClung, Carlos D. Bustamante, Susan R. McCouch

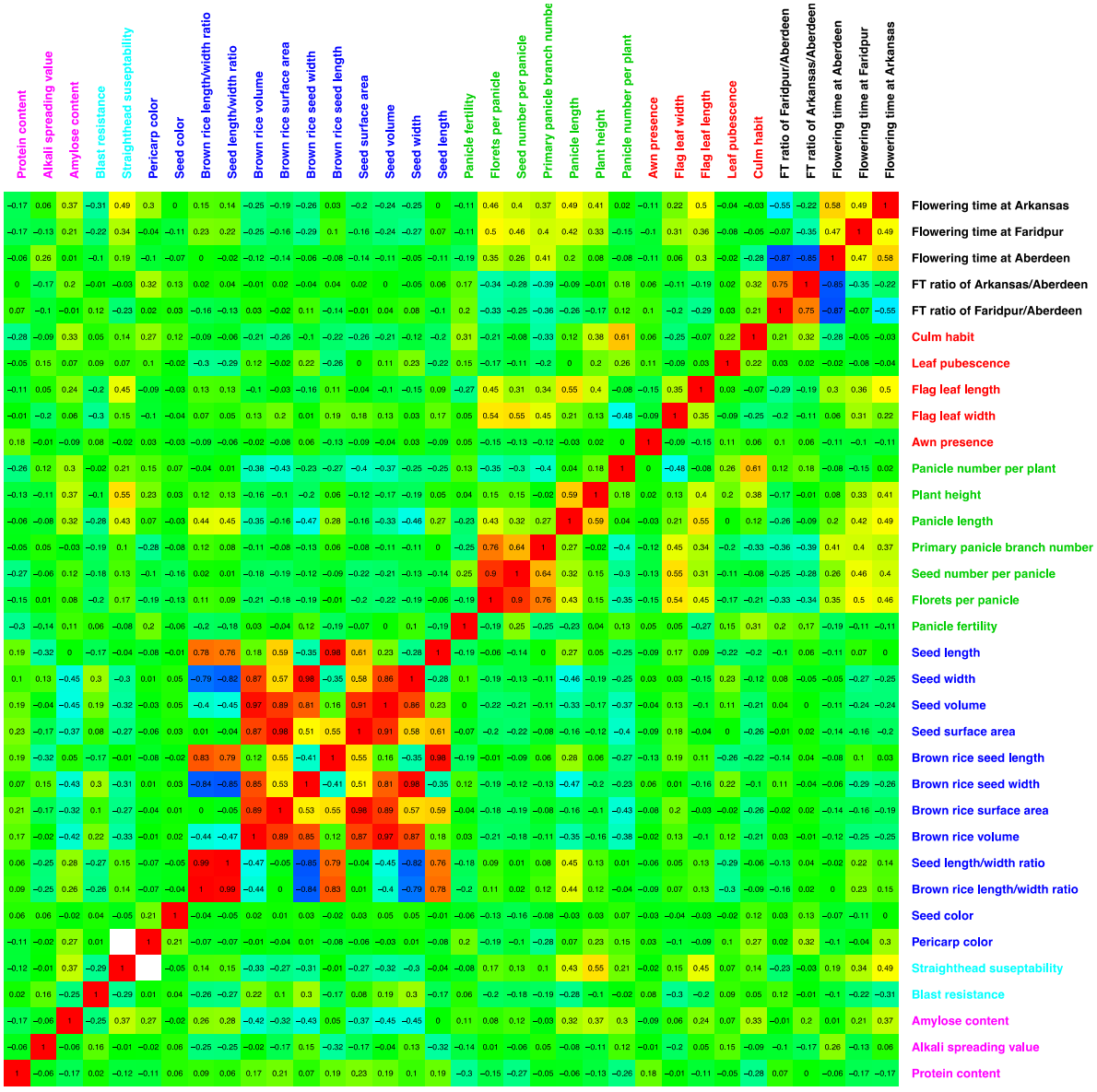
Supplementary Figures (S1-S41)

Supplementary Table (S1)

Supplementary Method

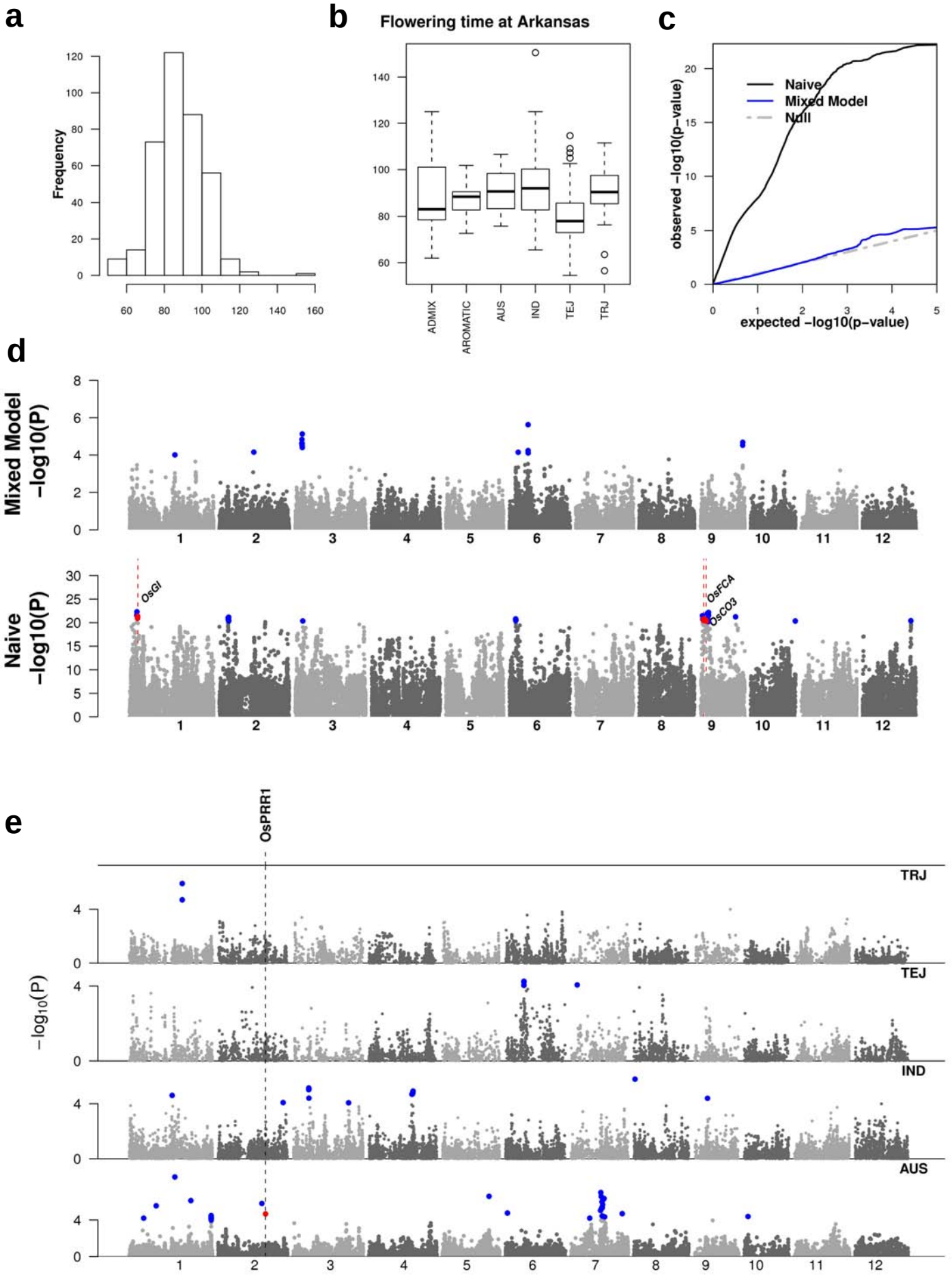


Supplementary Figure S1 Decay of LD in various sample populations. Decay of LD (measured as genotypic r^2) as a function of distance between SNPs.

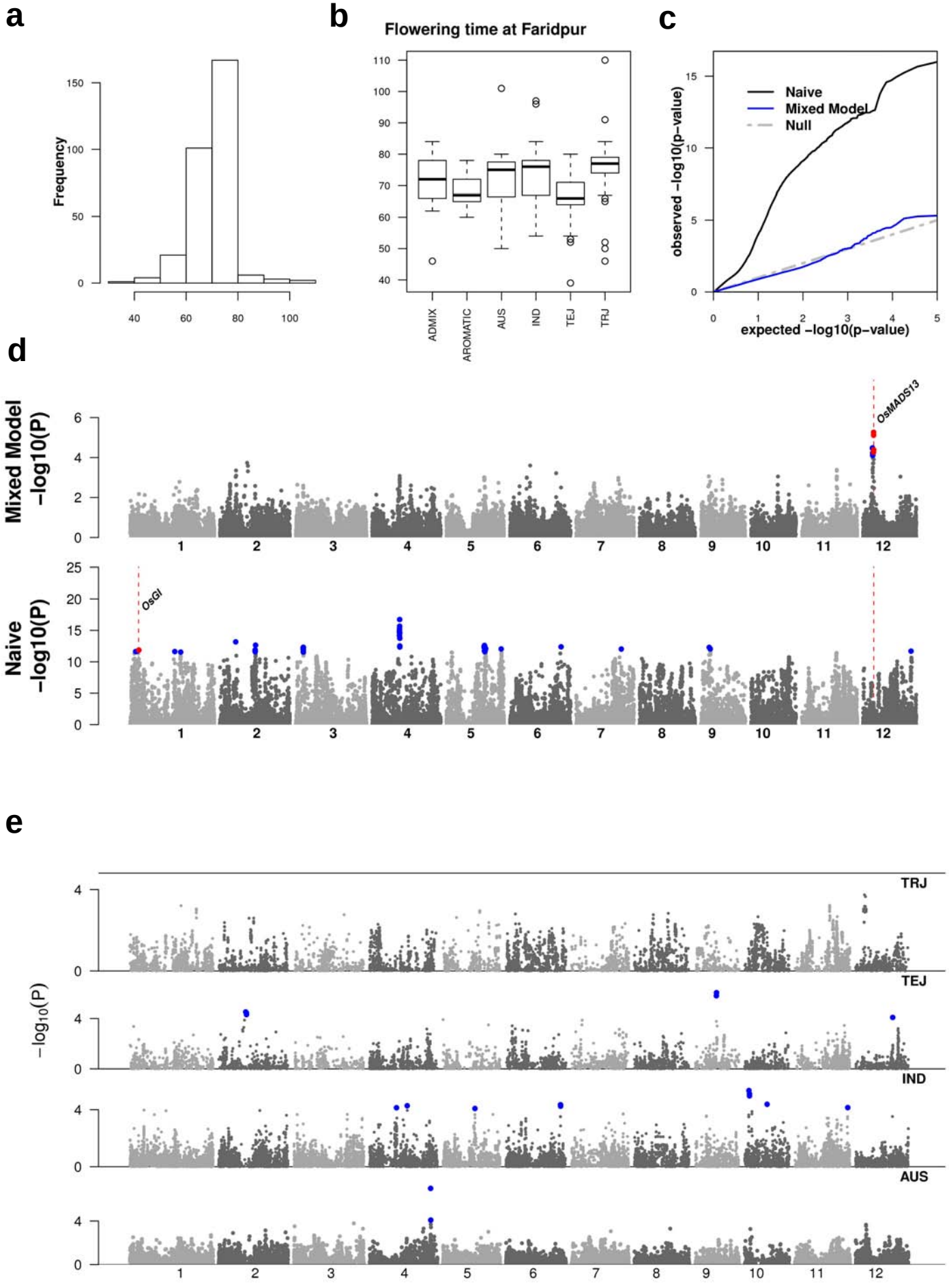


Supplementary Figure S2 Pairwise correlations of phenotypes across accessions, measured as the Pearson correlation.

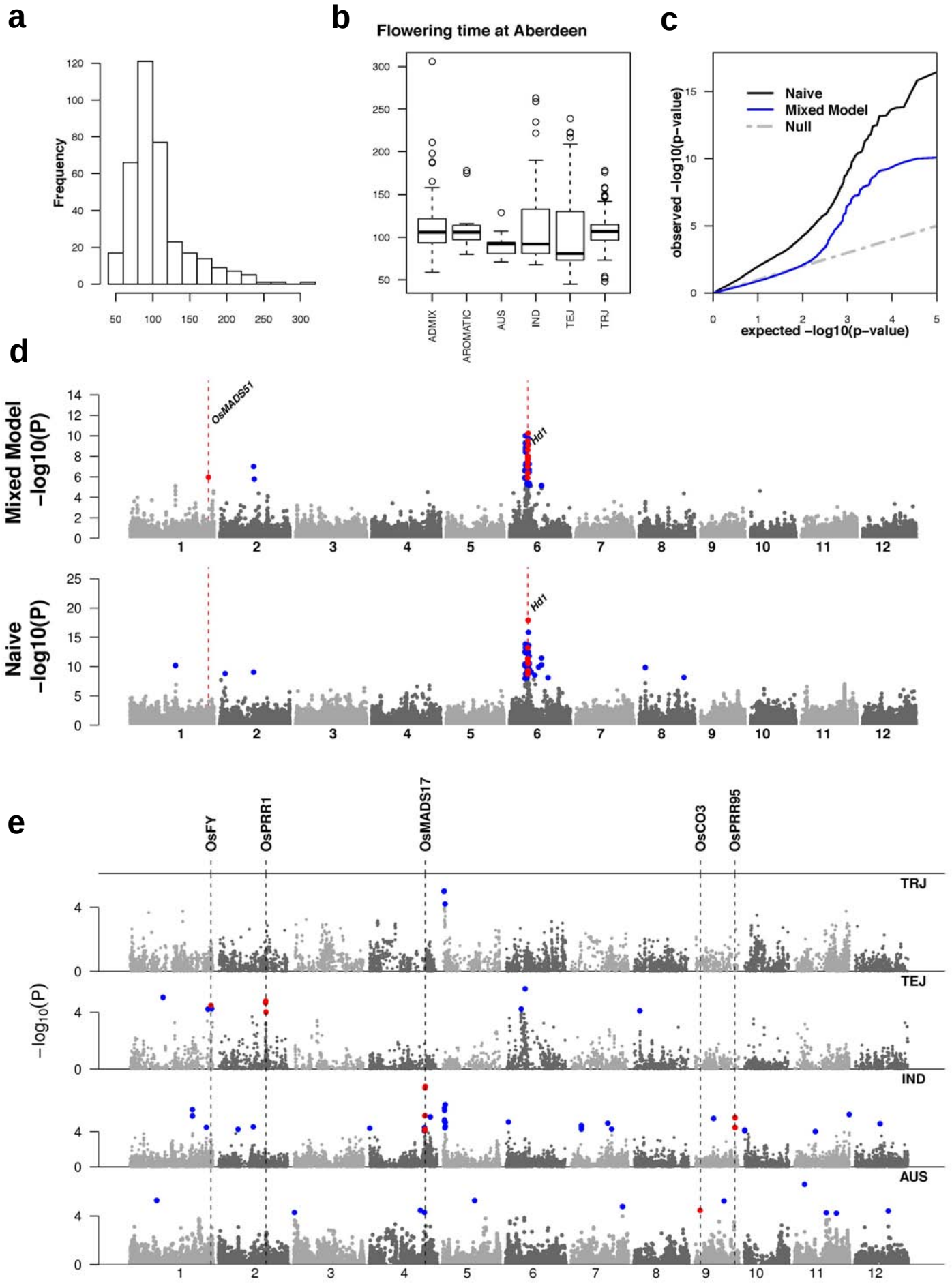
Supplementary Figure S3 to S36. The p-value of the SNPs and Quantile-Quantile plot of p-values for all traits studied**.** (a) A histogram showing the distribution of each phenotype in the rice diversity panel. (b) Box plot showing the mean, median and range of phenotypic variation for each *O. sativa* subpopulation independently. (c) Q-Q plot showing the expected null distribution of p-values, assuming no associations, represented as a broken grey line; distribution of p-values observed using the naïve model represented as a solid black line; distribution of p-values observed using the mixed model represented as a solid blue line. For simplicity, we only display those with p-values < 0.01. (d) P-values along the genome from the mixed model and naïve approaches. For the naïve approach, the top 50 SNPs are indicated; red dots indicates position of a previously discovered candidate gene within 200kb; blue dots indicates a newly discovered region containing no obvious candidate genes. For the mixed model, SNPs with p-values < 10^{-4} are colored red indicating candidate gene(s) within 200kb, or blue to represent new regions. (e) P-values along the genome from the mixed model in the *tropical japonica, temperate japonica, indica* and *aus* subpopulations independently. SNPs indicated in red and blue color as described in (d) above.



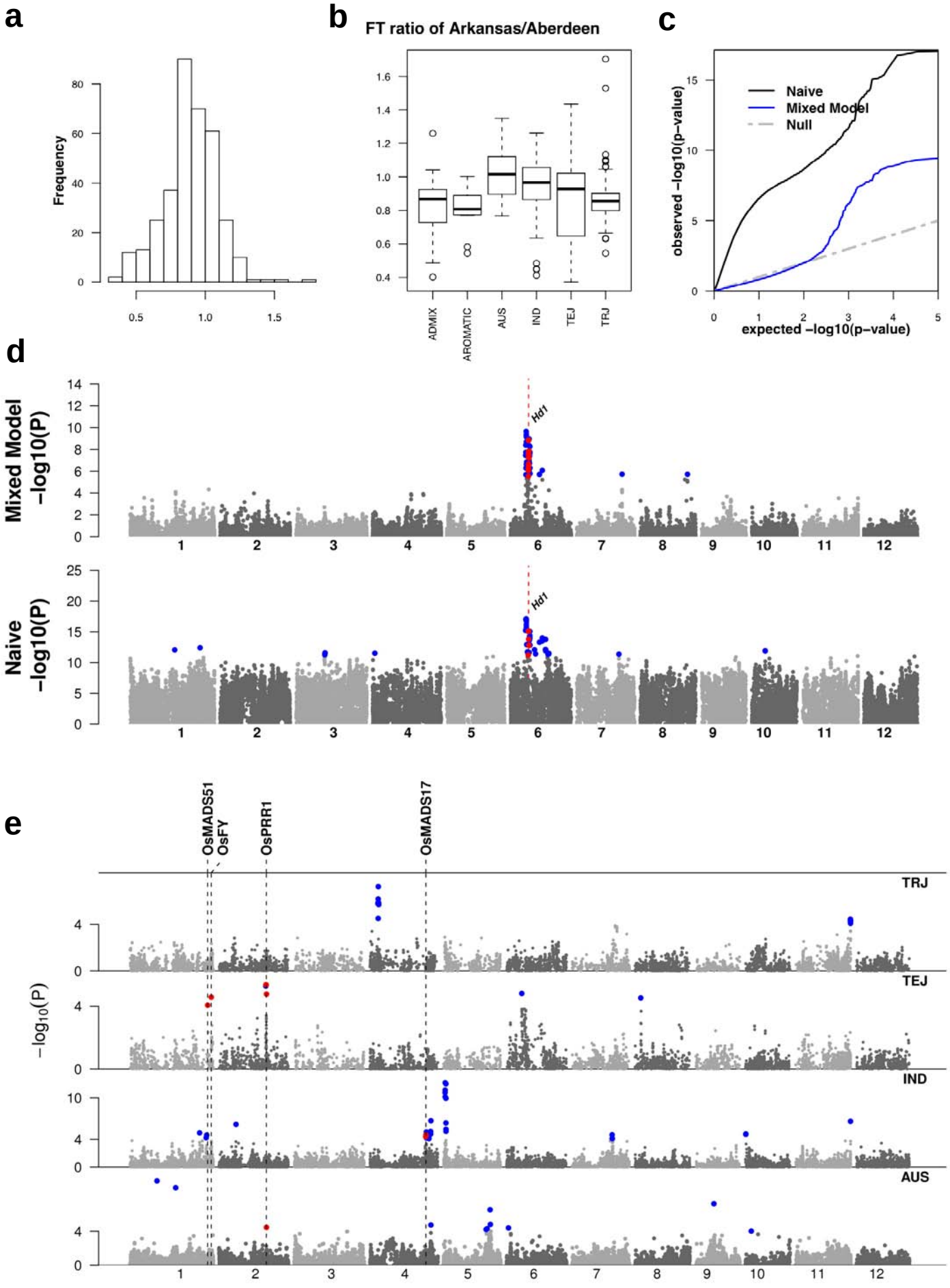
Supplementary Figure S3: Summary of GWAS results for Flowering time at Arkansas. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



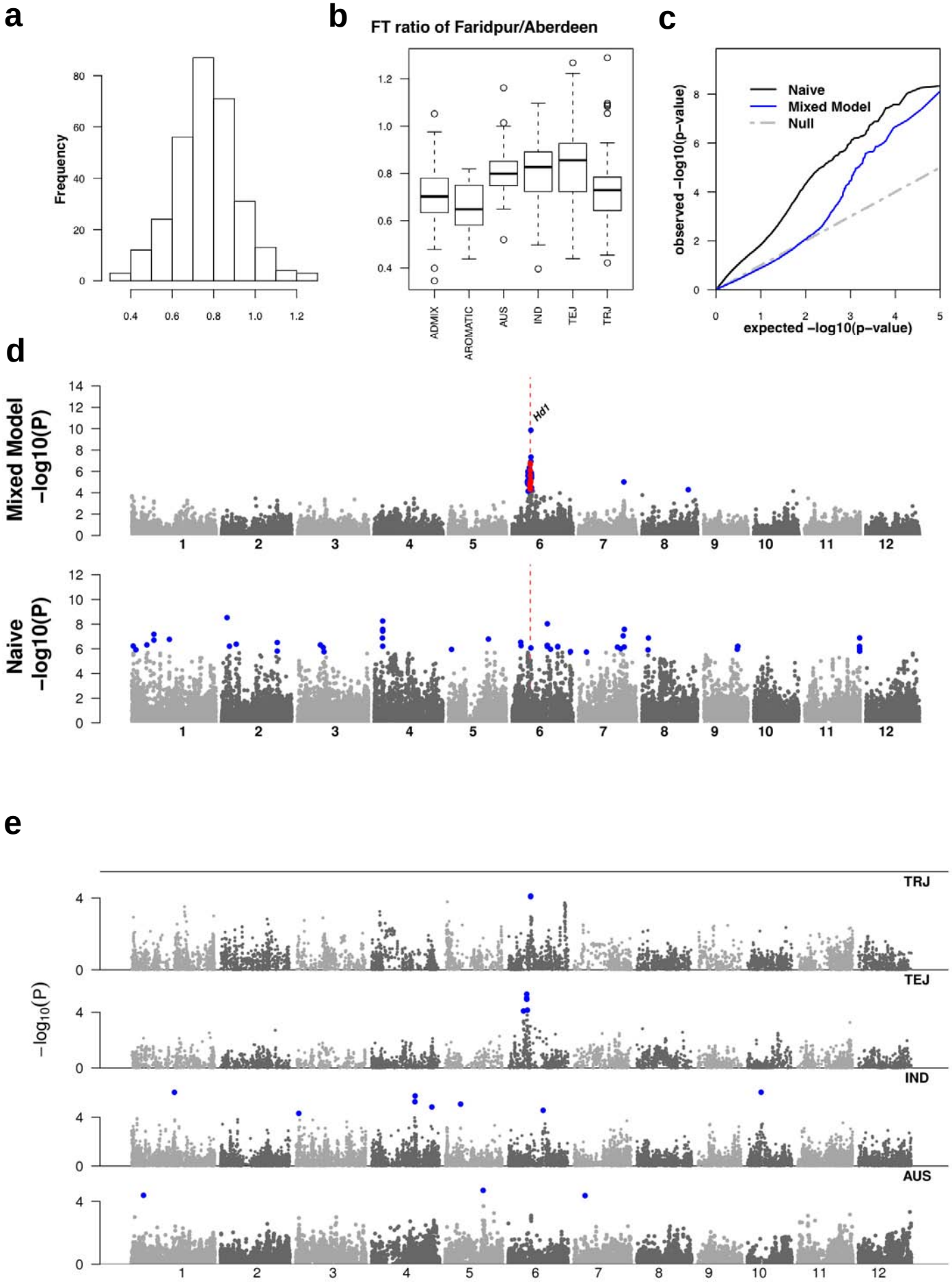
Supplementary Figure S4: Summary of GWAS results for Flowering time at Faridpur. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



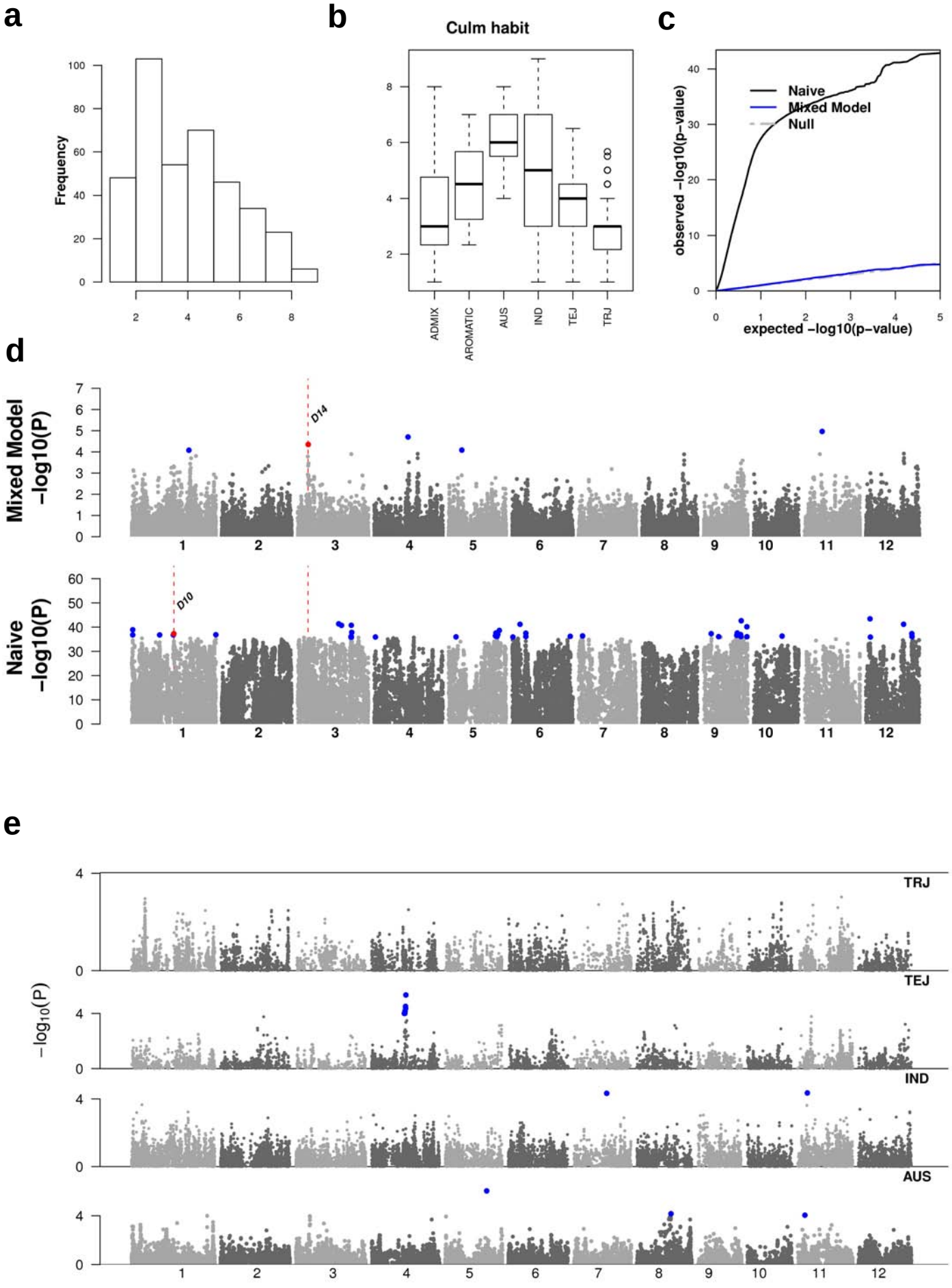
Supplementary Figure S5: Summary of GWAS results for Flowering time at Aberdeen. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



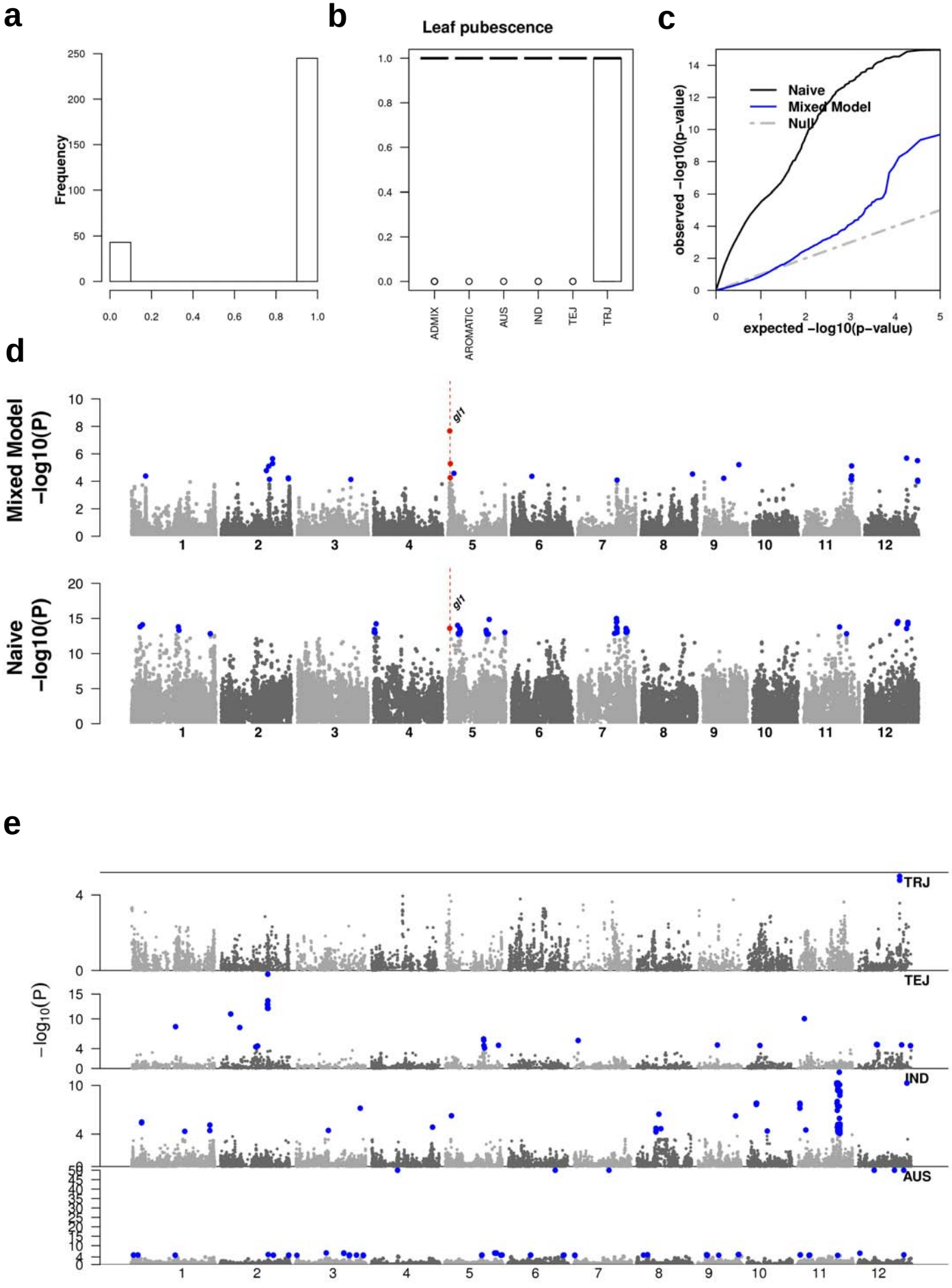
Supplementary Figure S6: Summary of GWAS results for FT ratio of Arkansas/Aberdeen. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



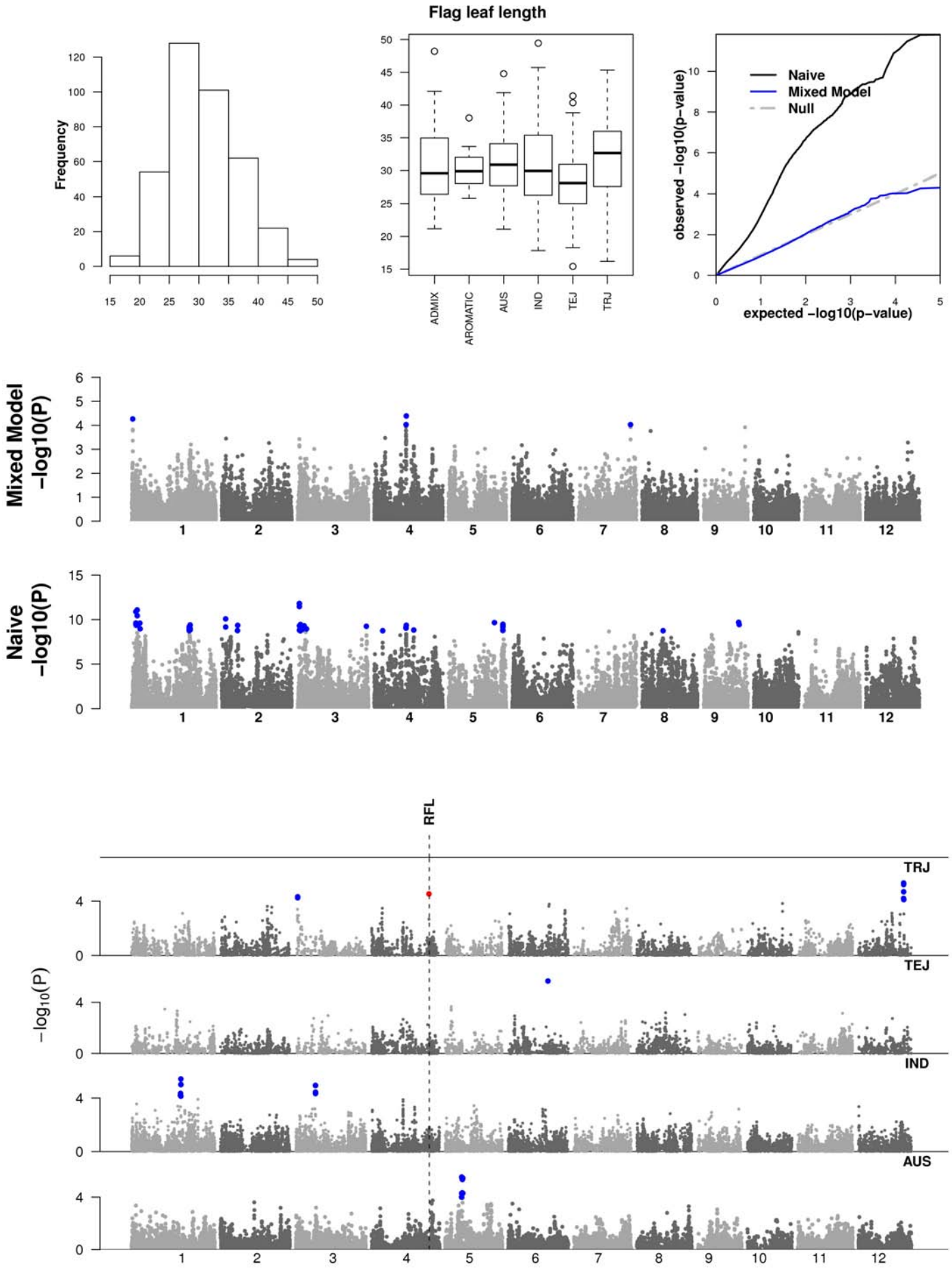
Supplementary Figure S7: Summary of GWAS results for FT ratio of Faridpur/Aberdeen. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



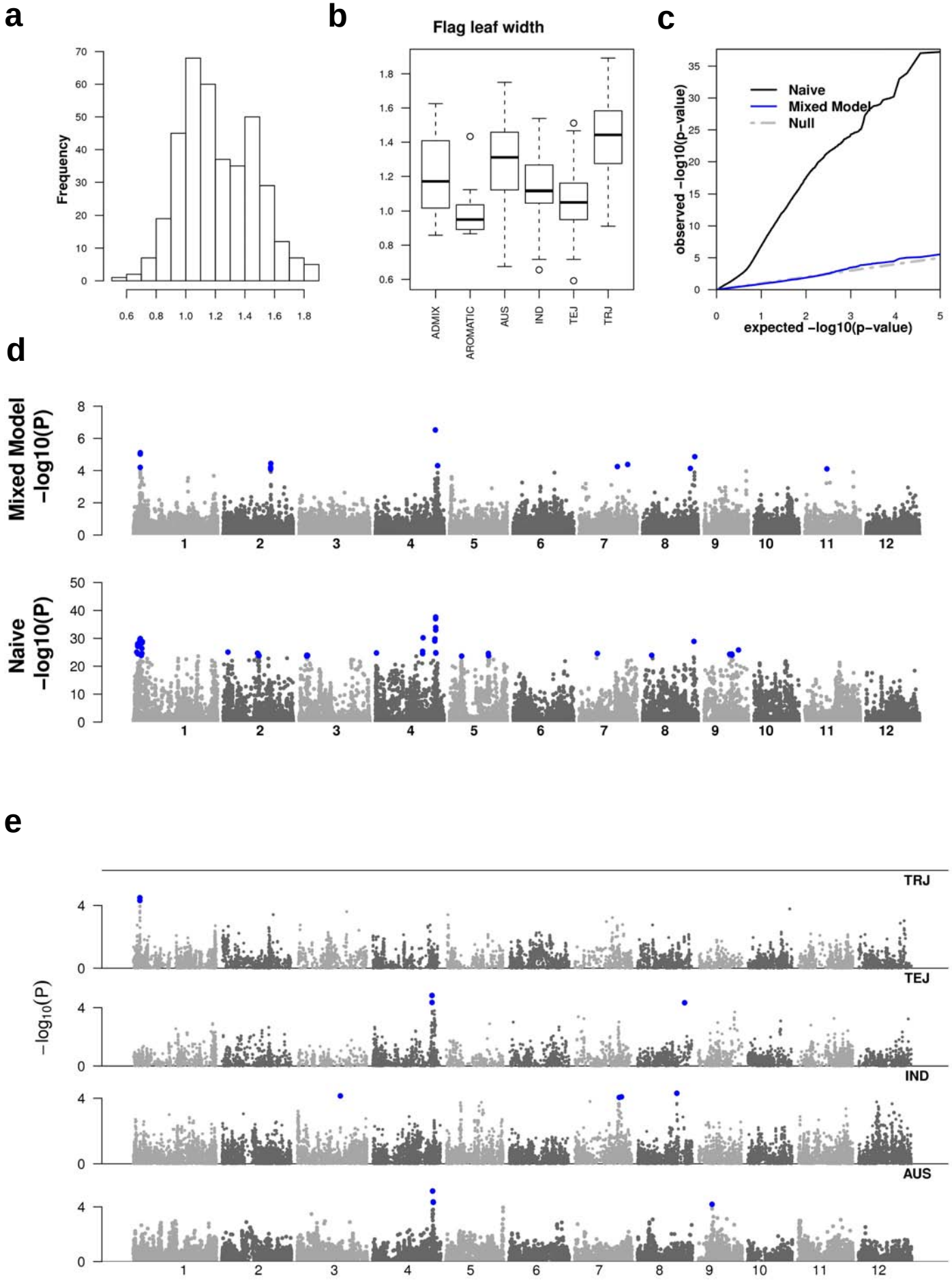
Supplementary Figure S8: Summary of GWAS results for Culm habit. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



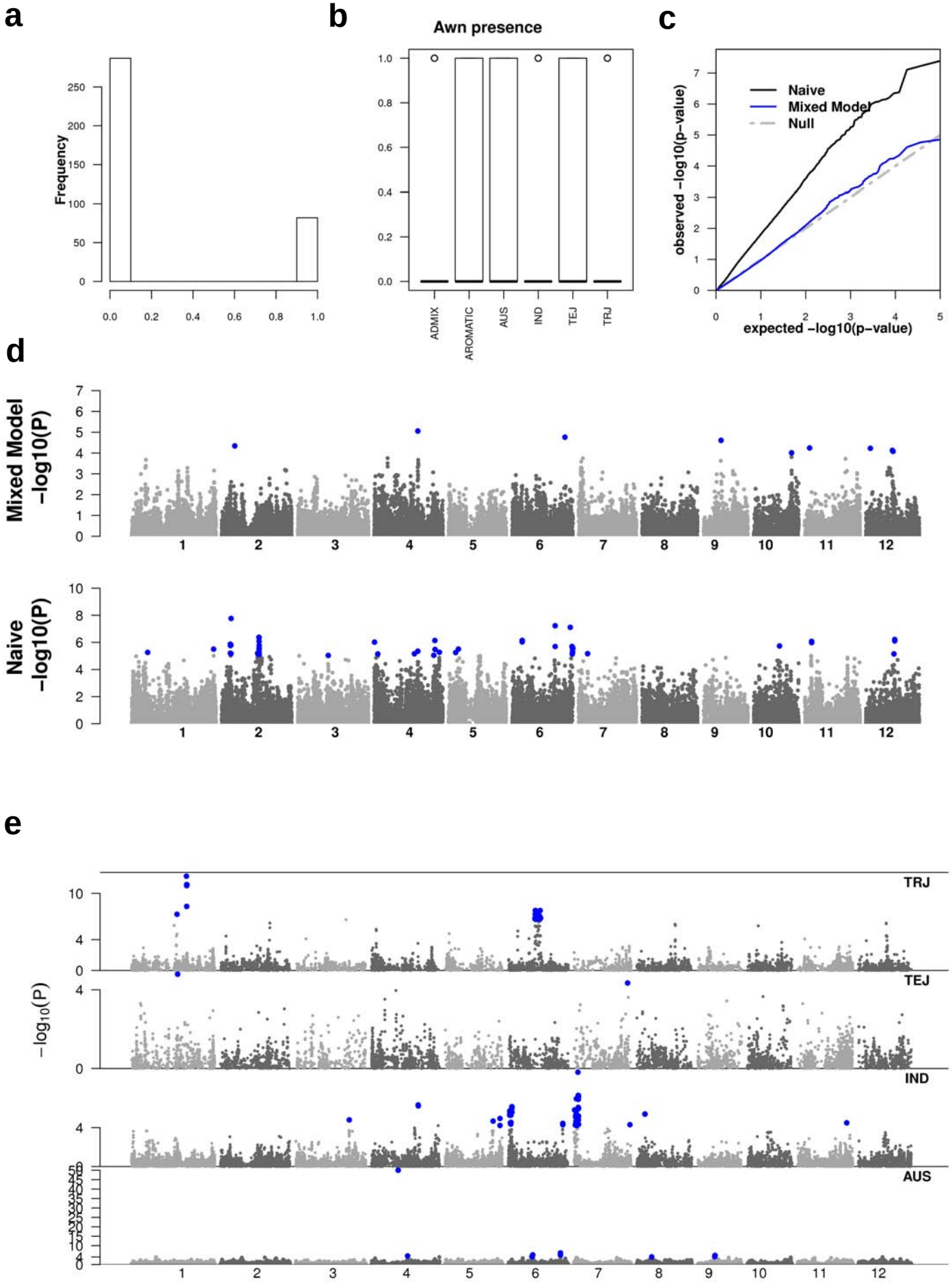
Supplementary Figure S9: Summary of GWAS results for Leaf pubescence. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



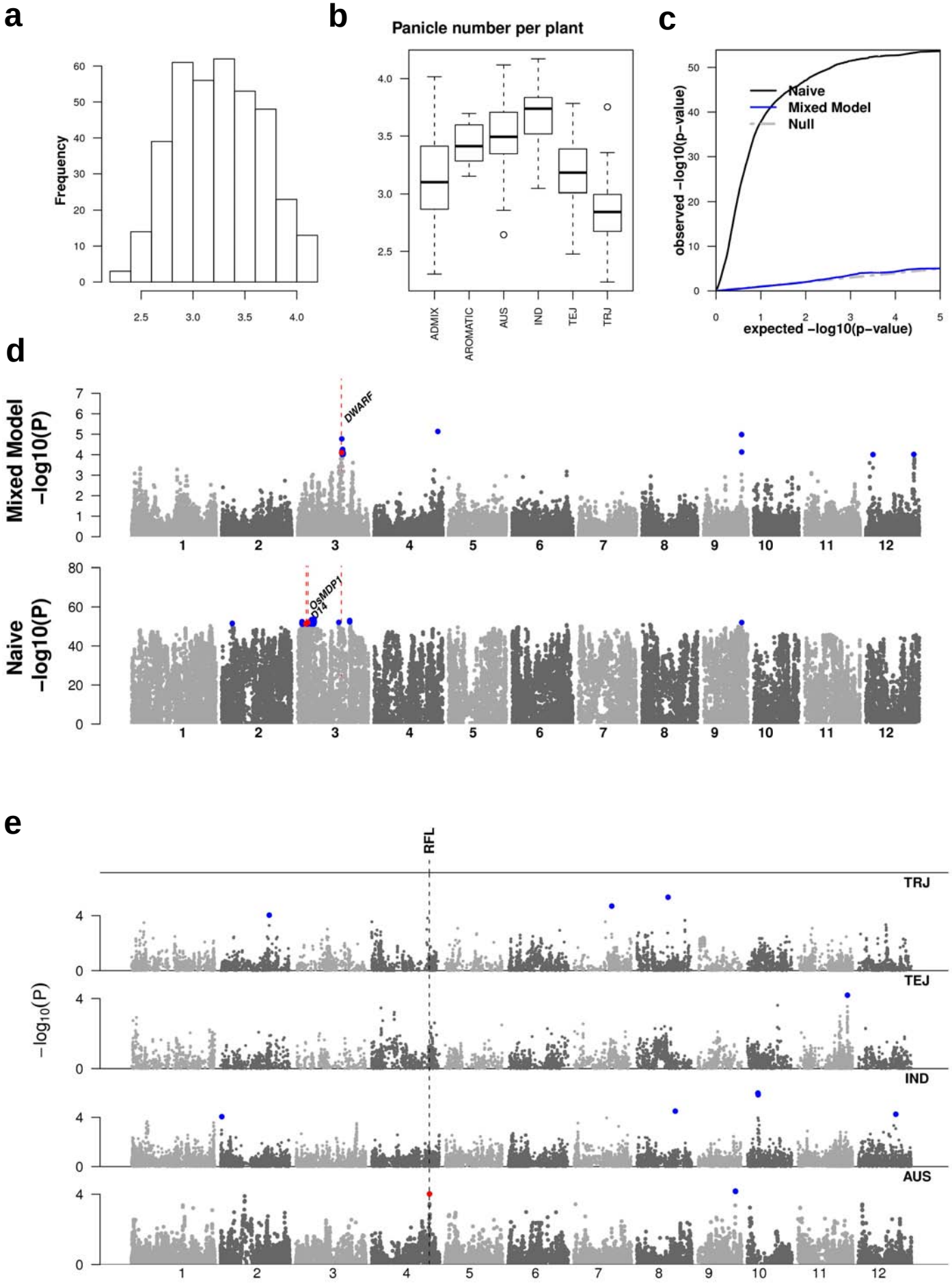
Supplementary Figure S10: Summary of GWAS results for Flag leaf length. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



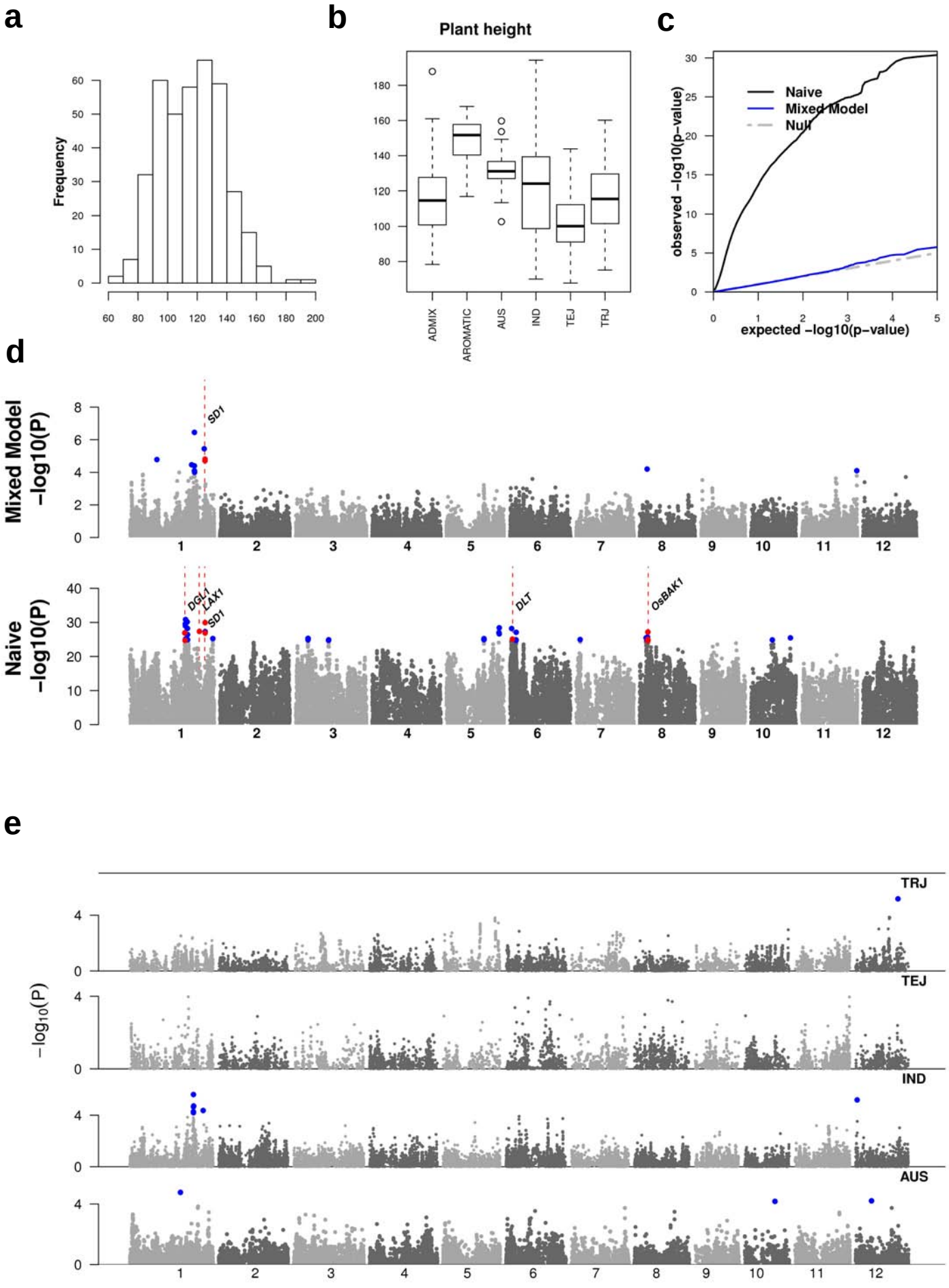
Supplementary Figure S11: Summary of GWAS results for Flag leaf width. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



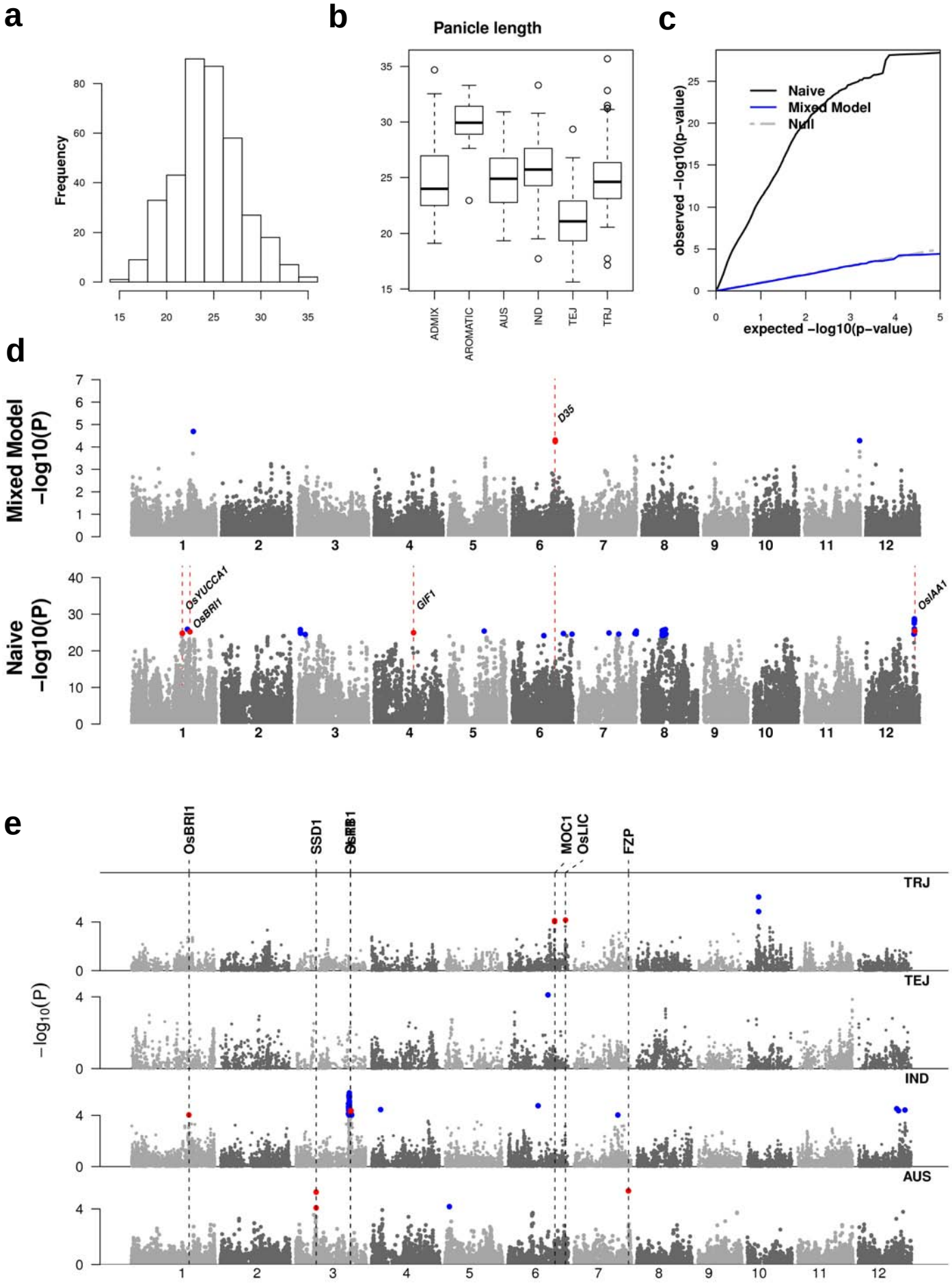
Supplementary Figure S12: Summary of GWAS results for Awn presence. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



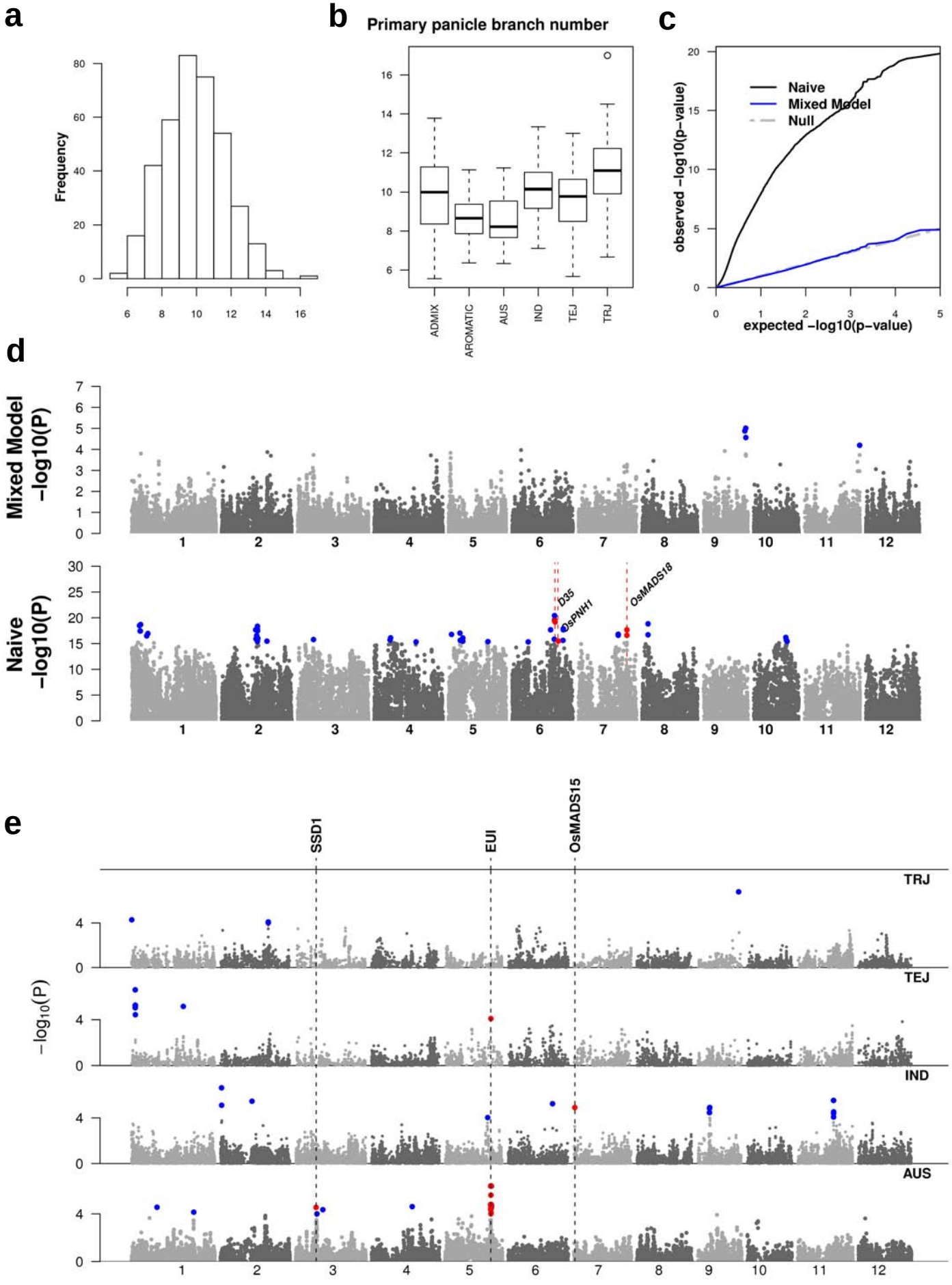
Supplementary Figure S13: Summary of GWAS results for Panicle number per plant. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



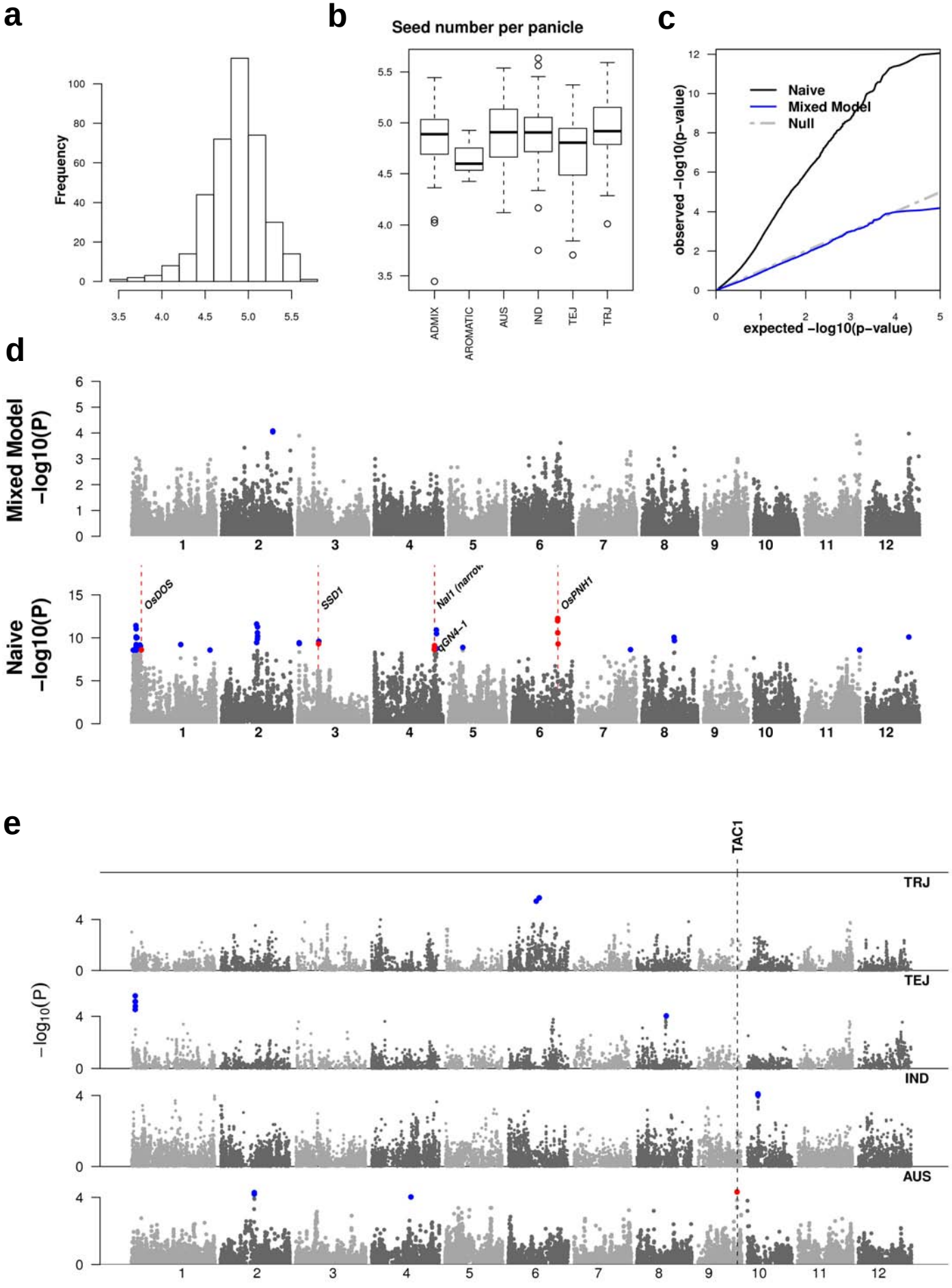
Supplementary Figure S14: Summary of GWAS results for Plant height. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



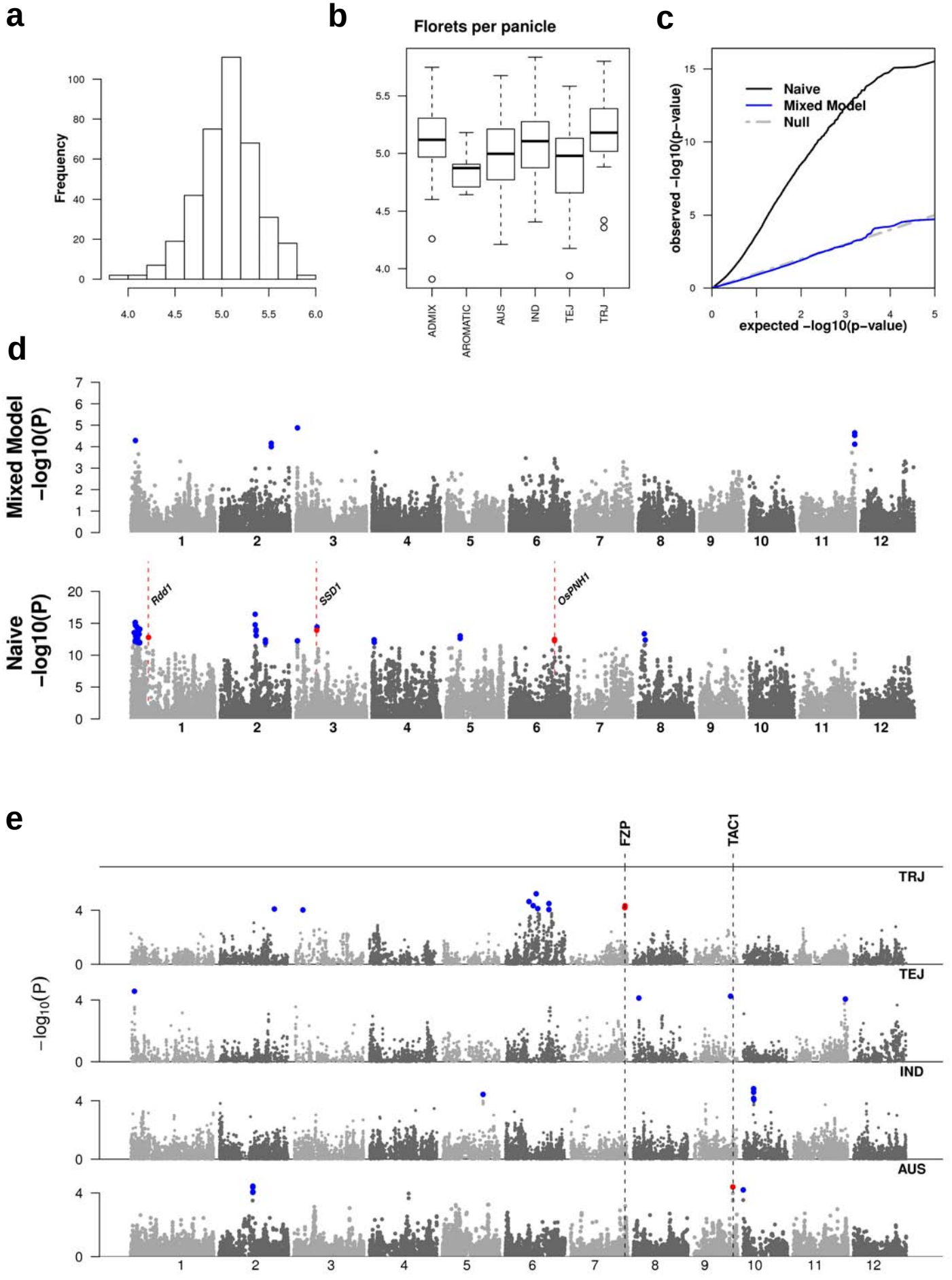
Supplementary Figure S15: Summary of GWAS results for Panicle length. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



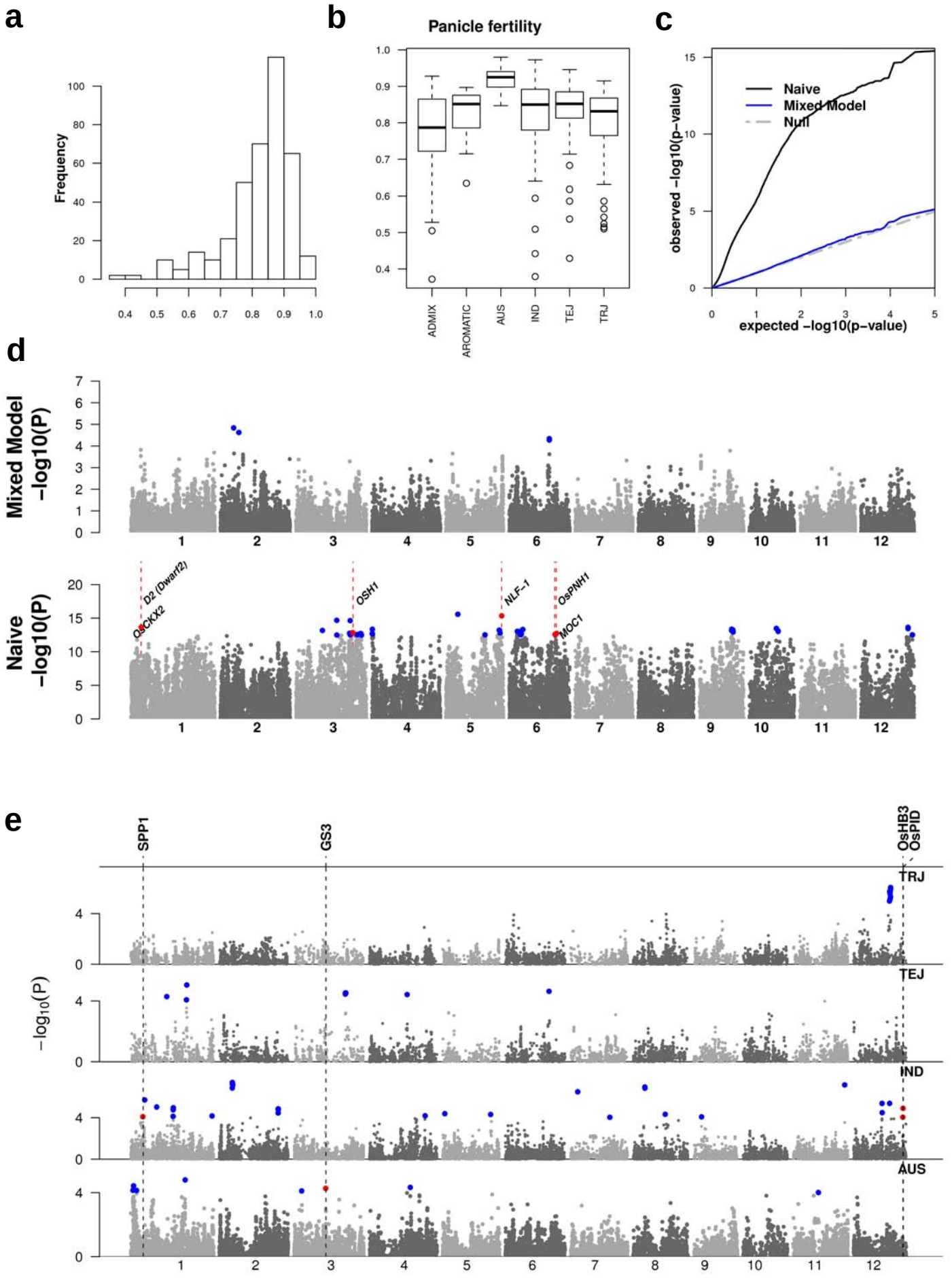
Supplementary Figure S16: Summary of GWAS results for Primary panicle branch number. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



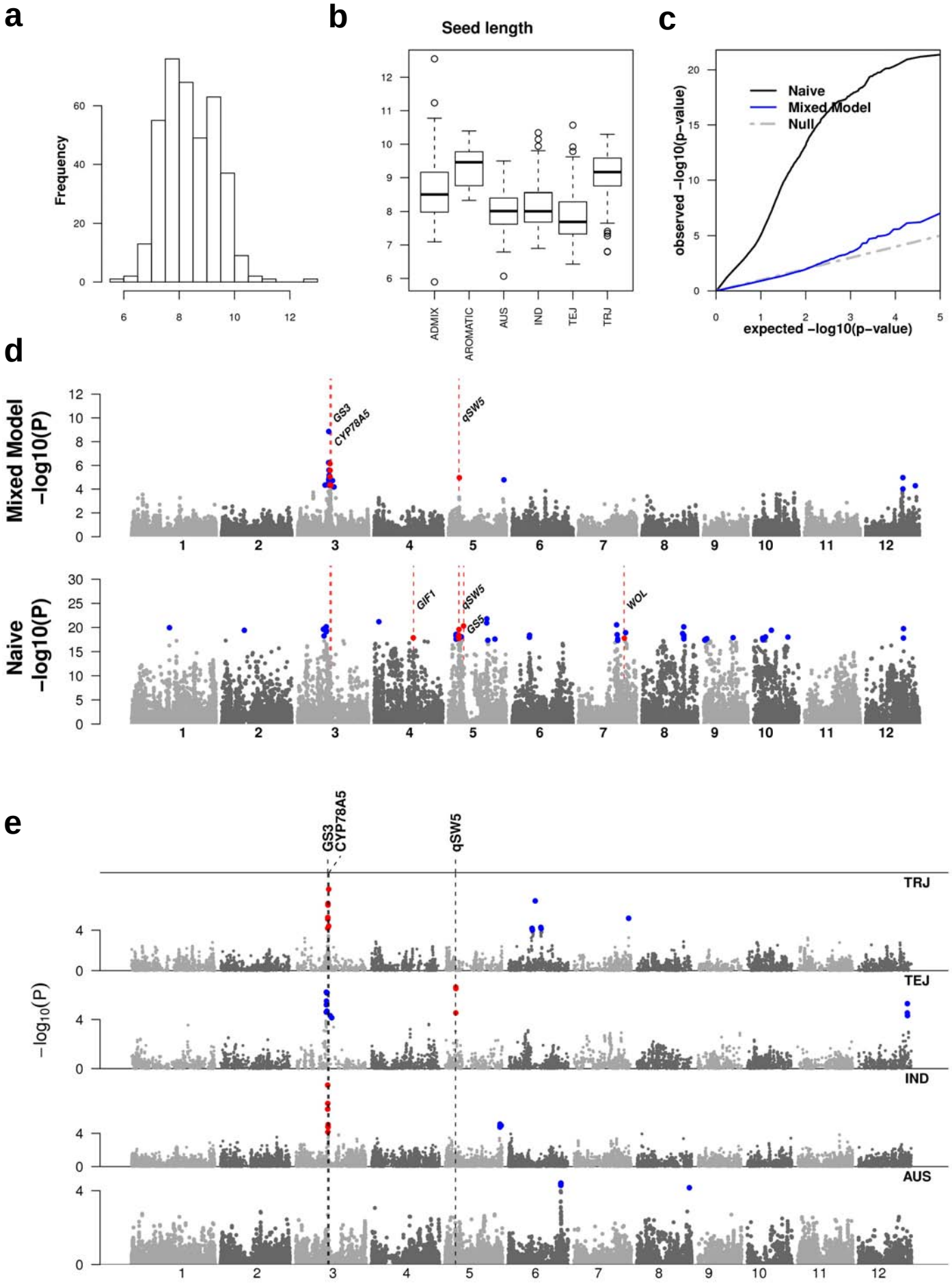
Supplementary Figure S17: Summary of GWAS results for Seed number per panicle. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



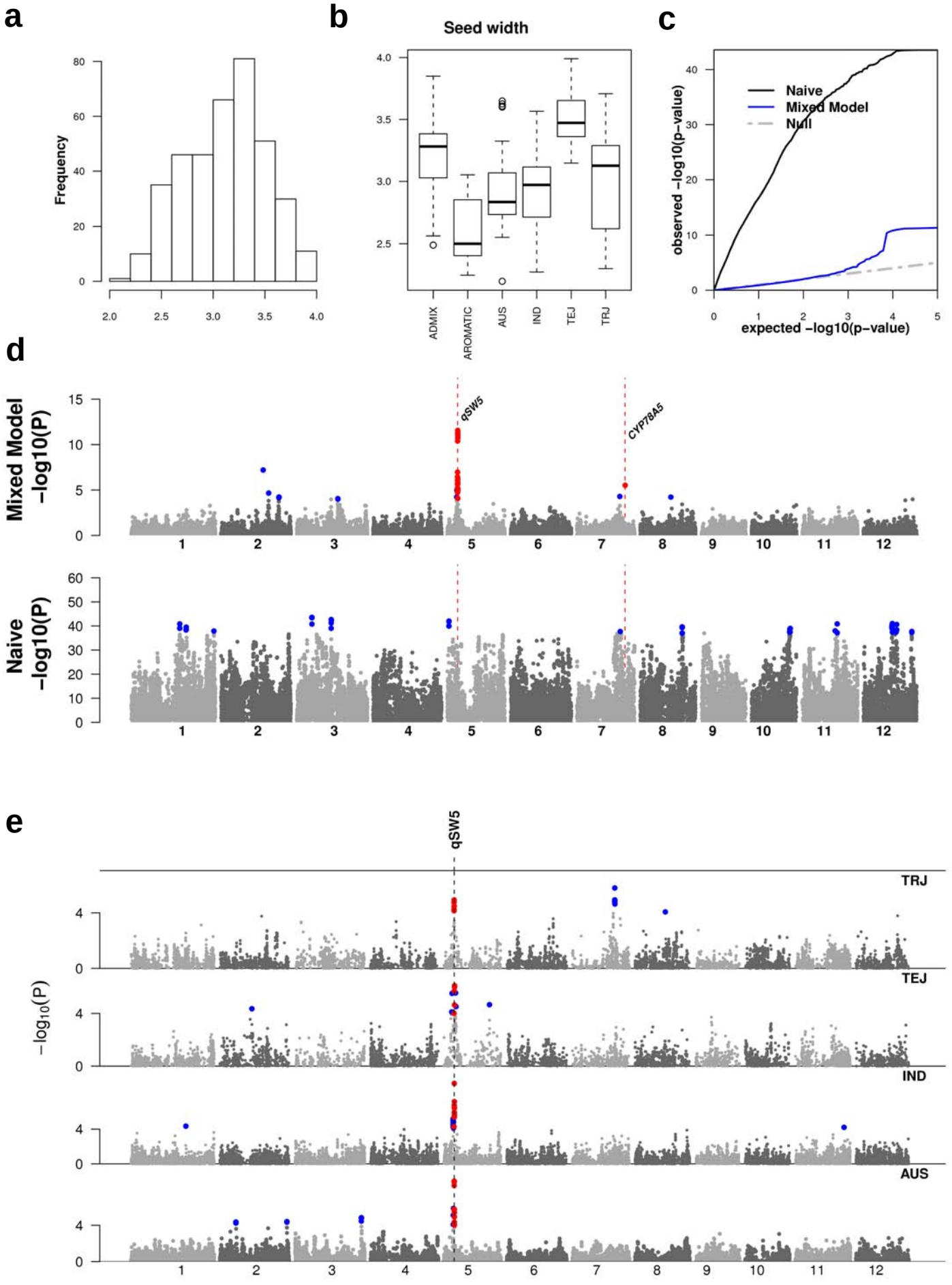
Supplementary Figure S18: Summary of GWAS results for Florets per panicle. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



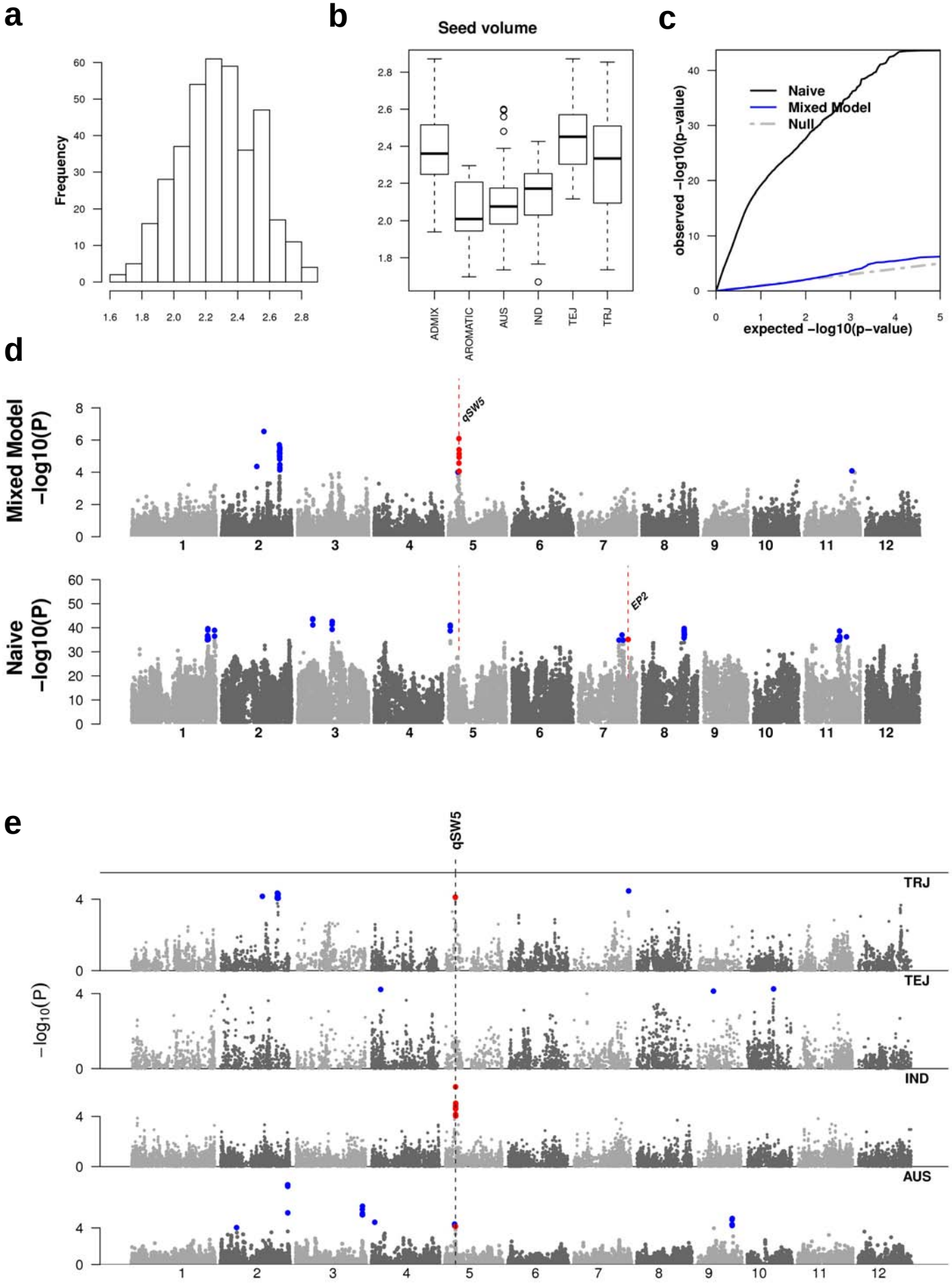
Supplementary Figure S19: Summary of GWAS results for Panicle fertility. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



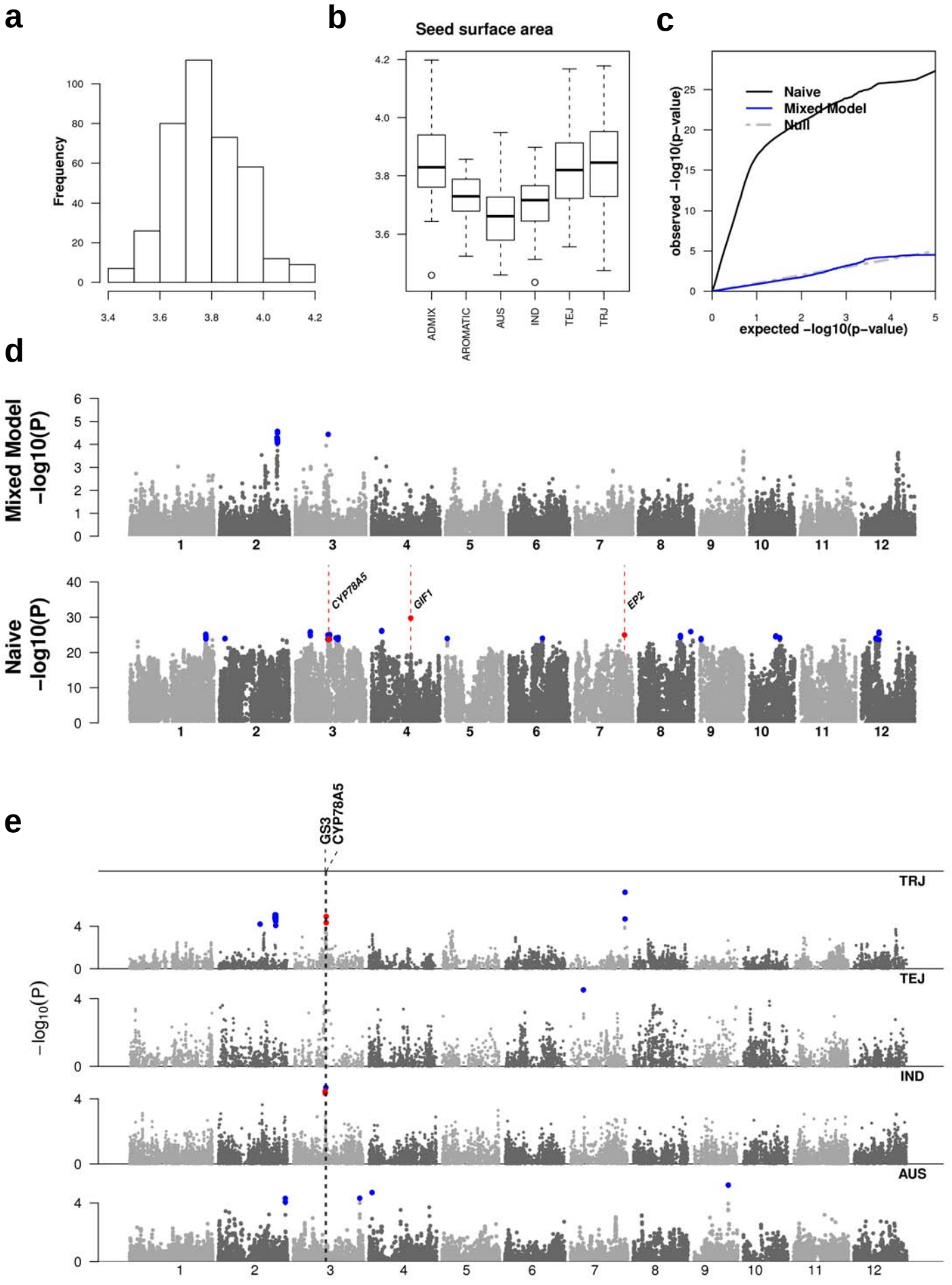
Supplementary Figure S20: Summary of GWAS results for Seed length. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



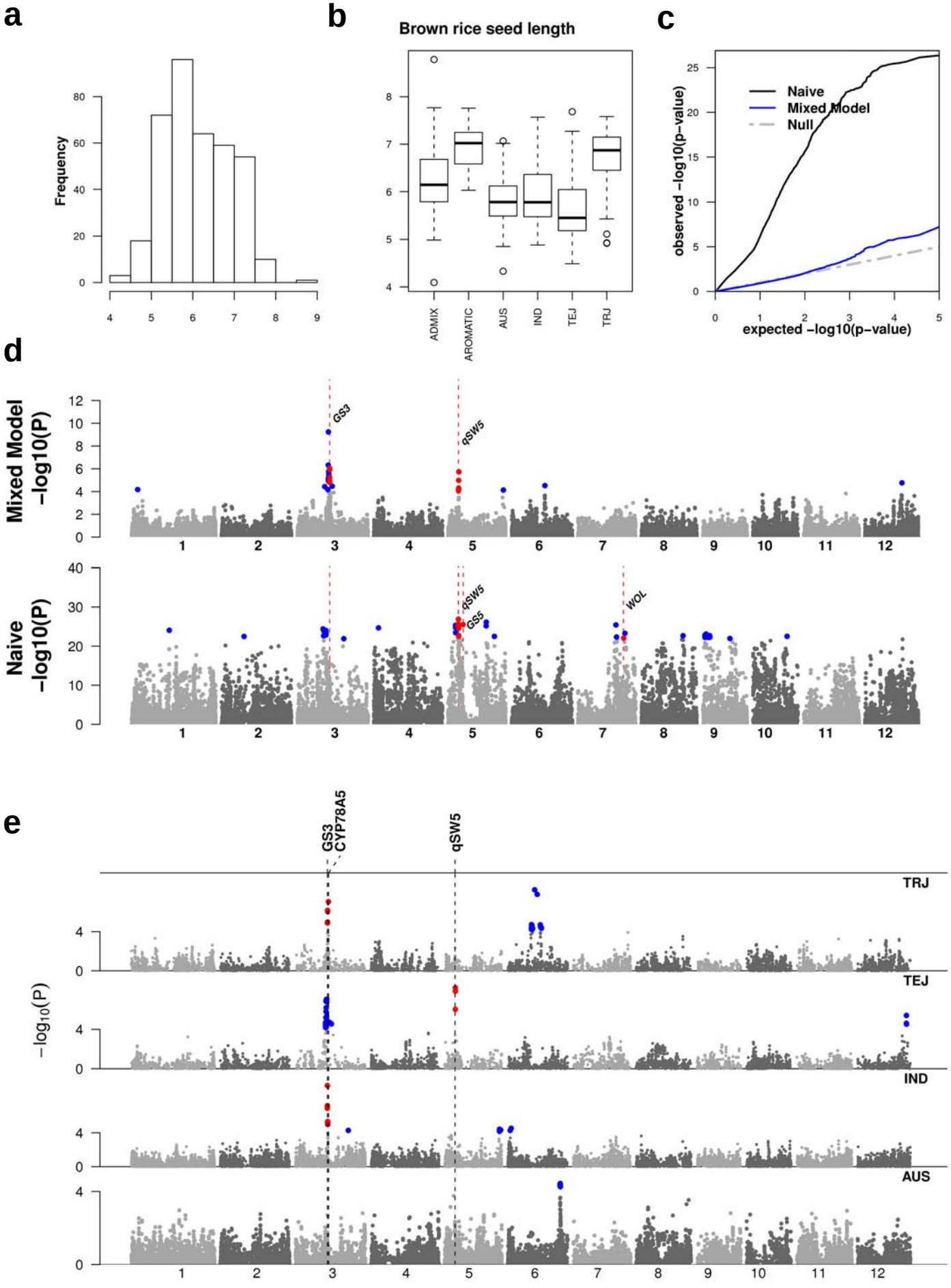
Supplementary Figure S21: Summary of GWAS results for Seed width. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



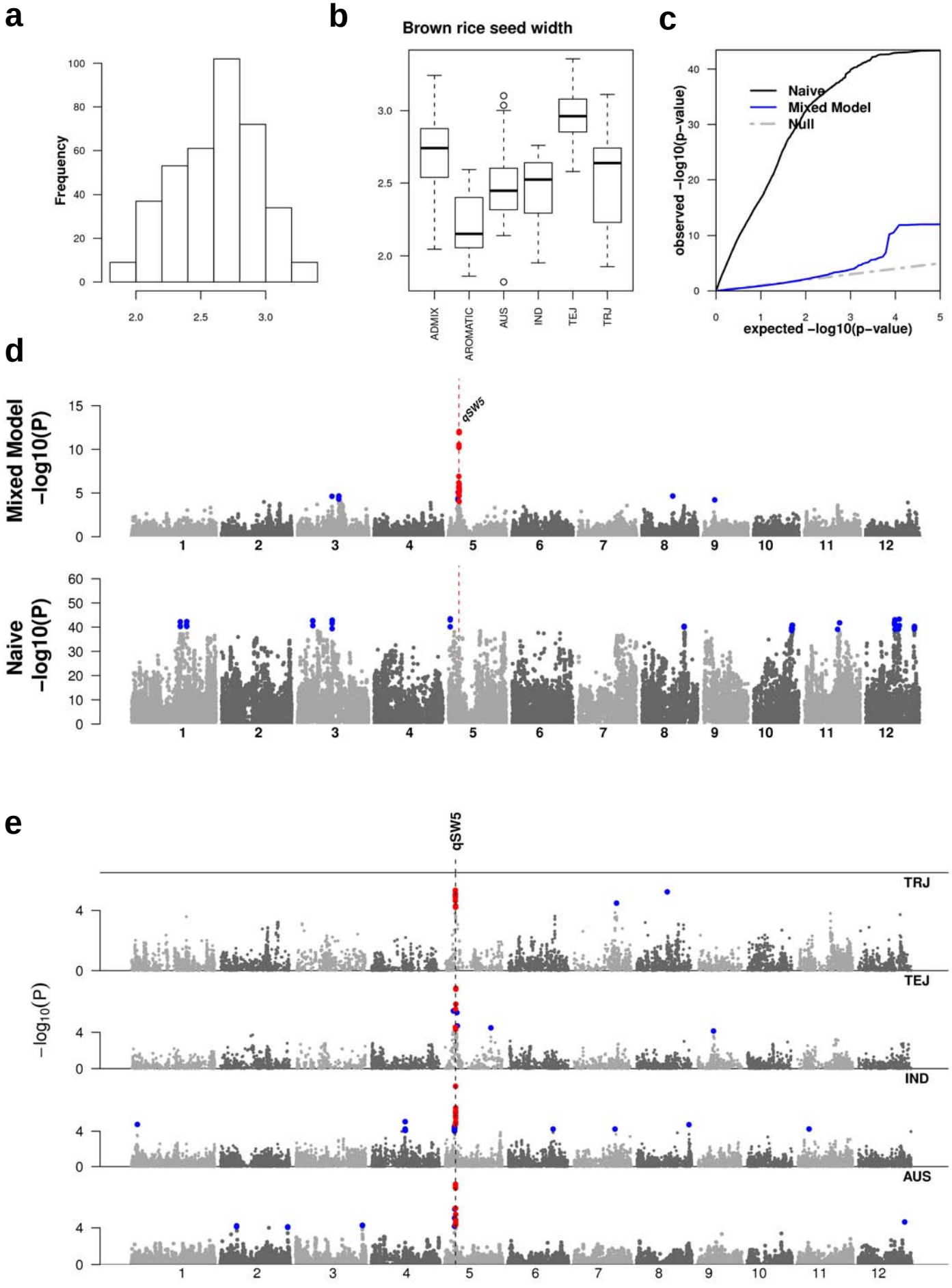
Supplementary Figure S22: Summary of GWAS results for Seed volume. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



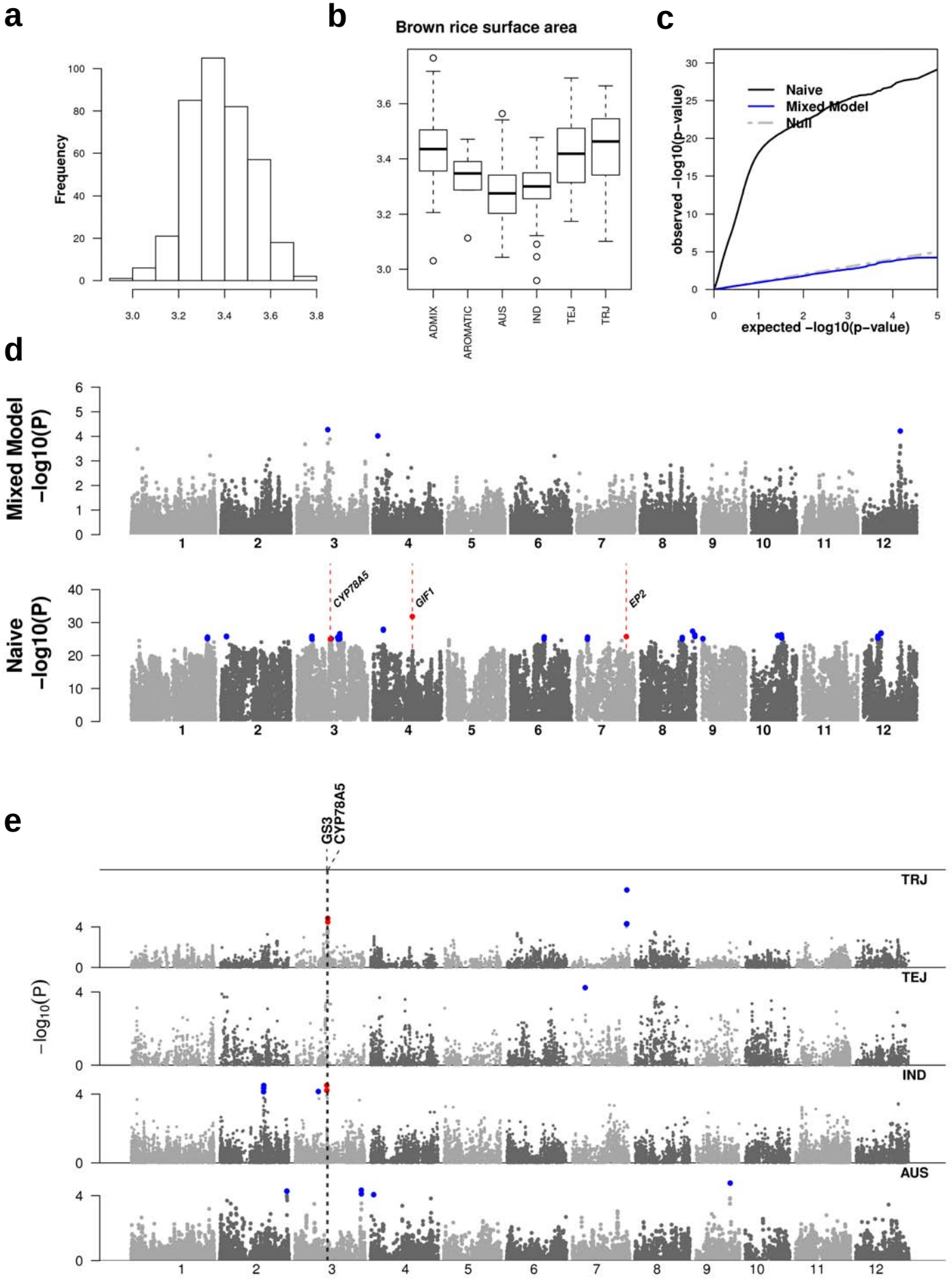
Supplementary Figure S23: Summary of GWAS results for Seed surface area. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



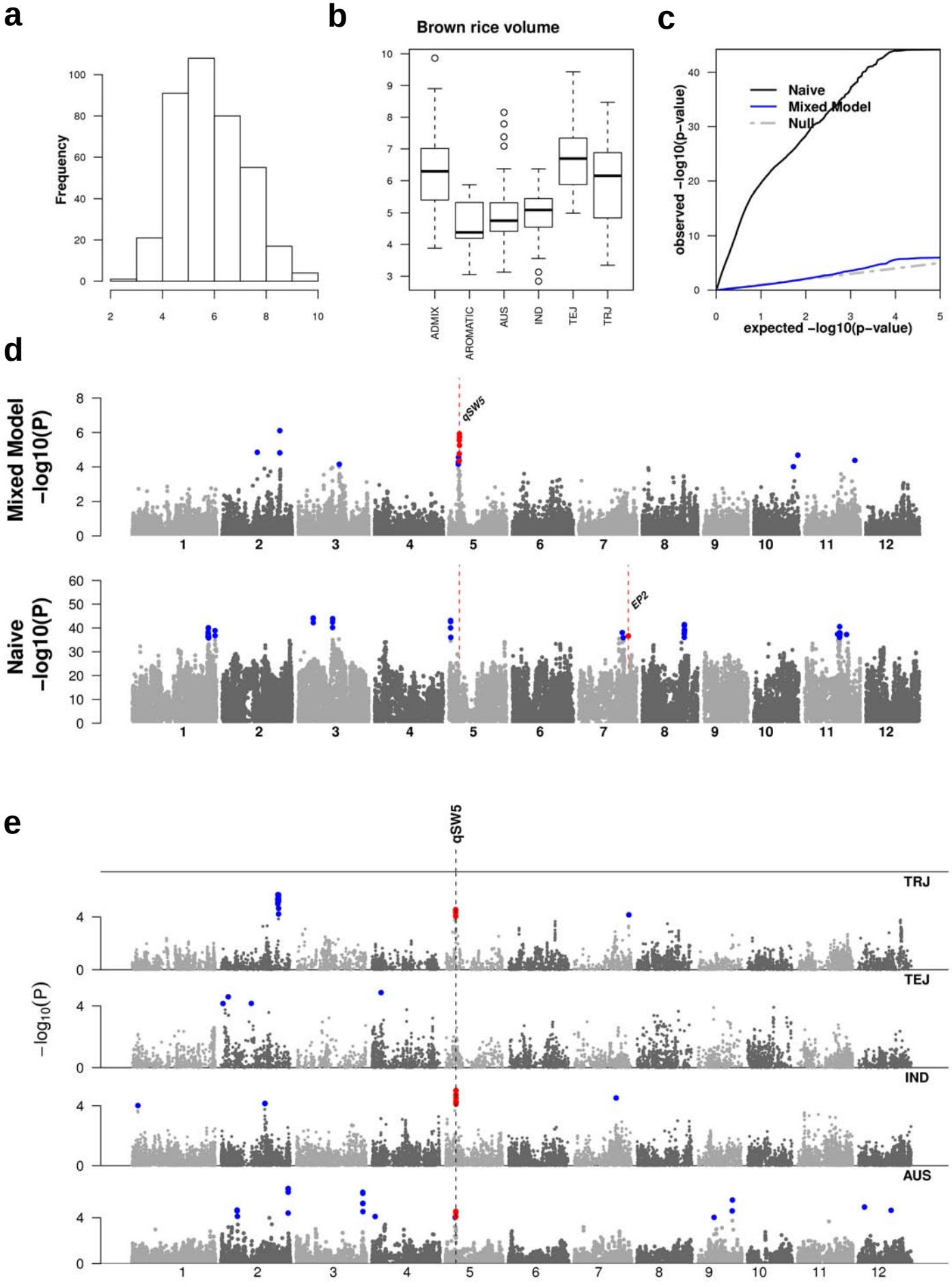
Supplementary Figure S24: Summary of GWAS results for Brown rice seed length. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



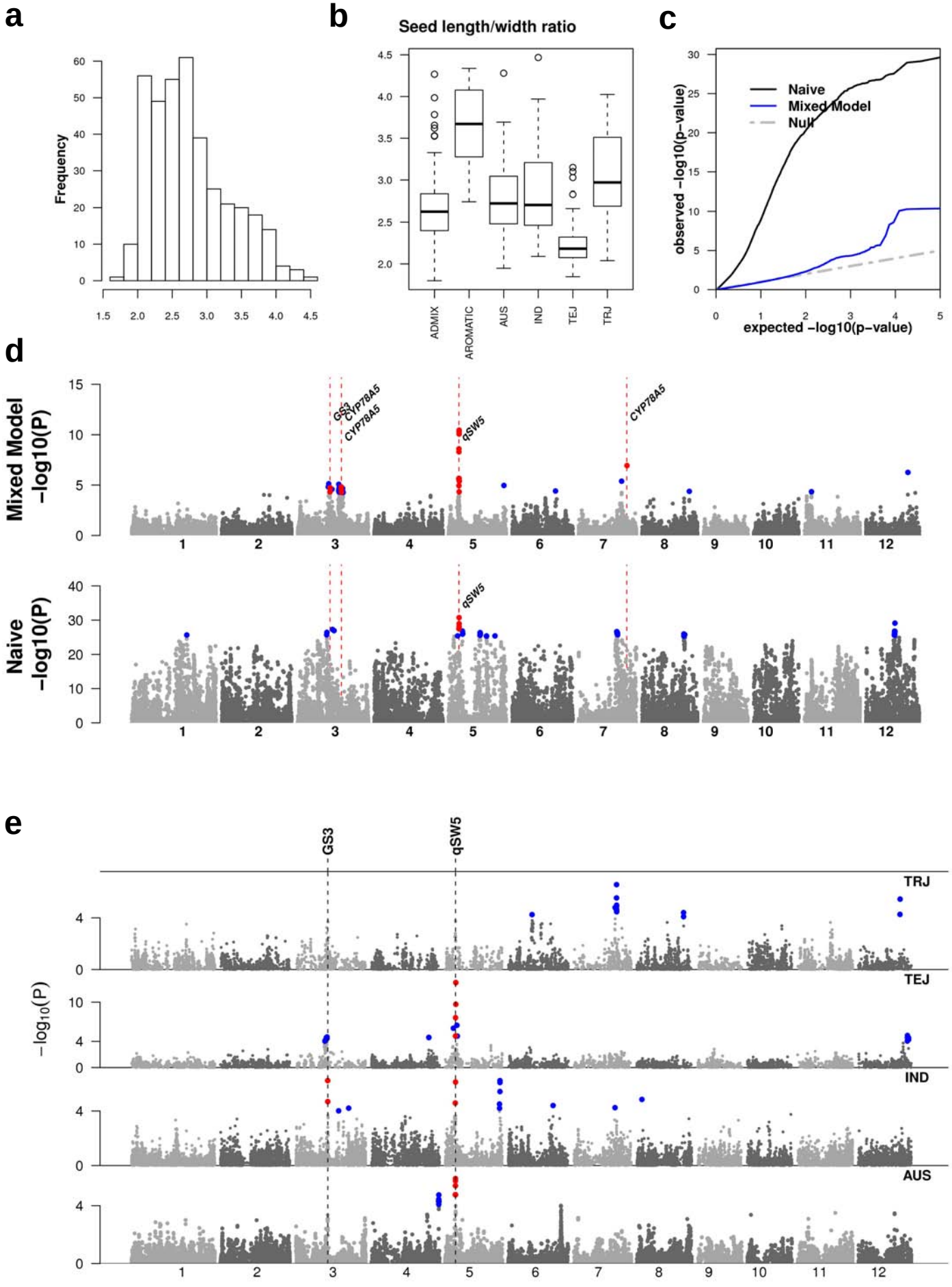
Supplementary Figure S25: Summary of GWAS results for Brown rice seed width. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



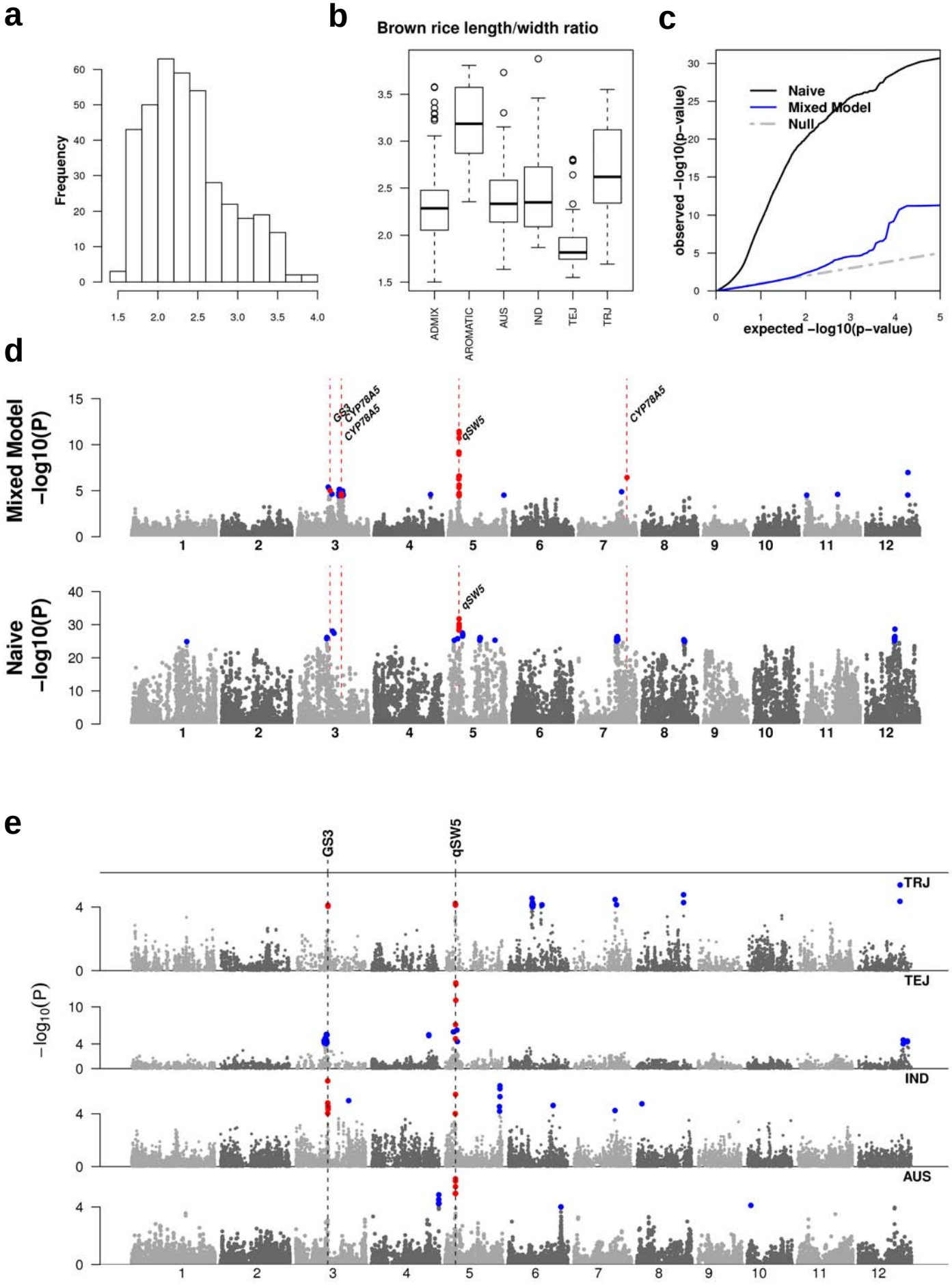
Supplementary Figure S26: Summary of GWAS results for Brown rice surface area. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



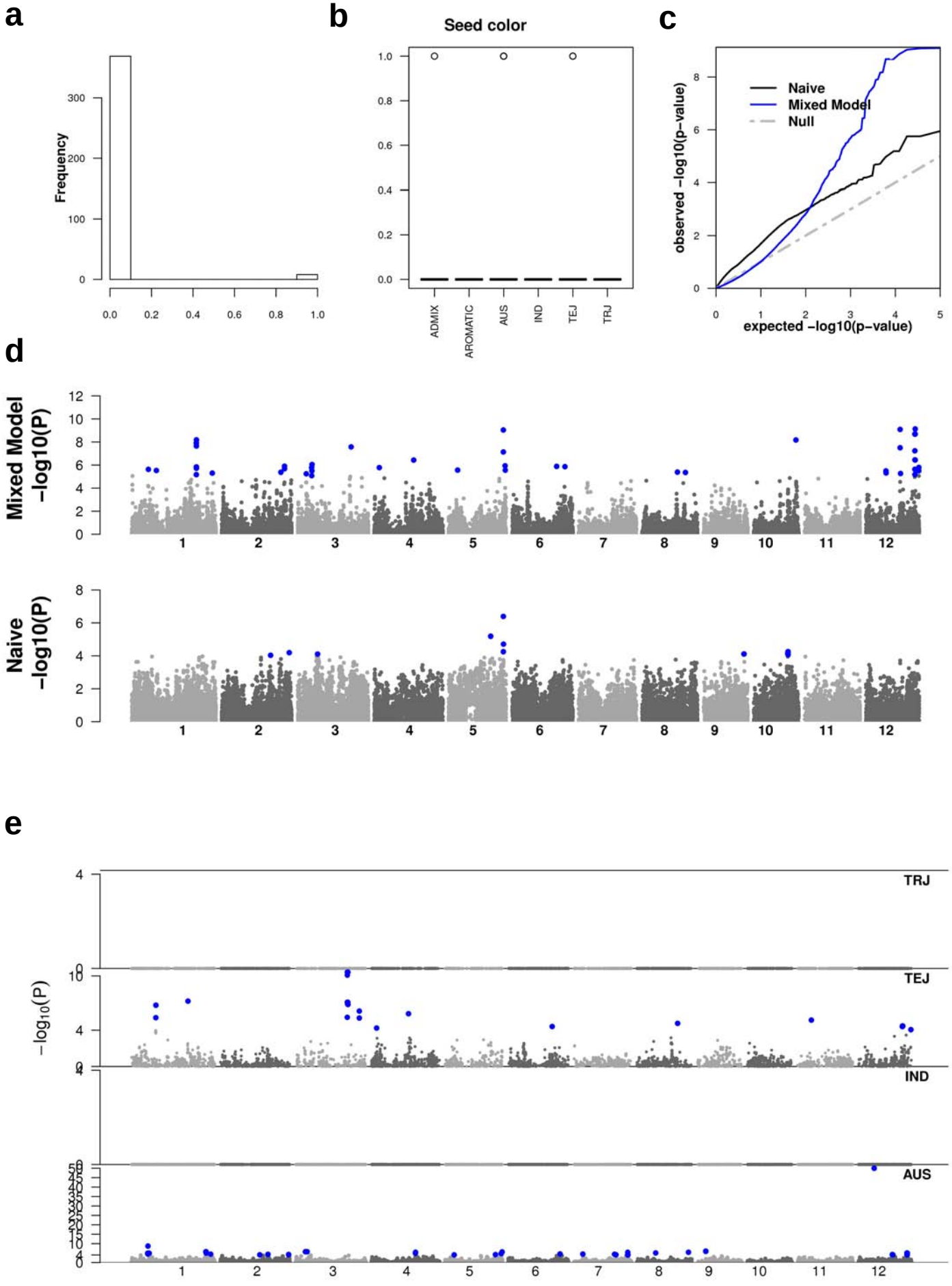
Supplementary Figure S27: Summary of GWAS results for Brown rice volume. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



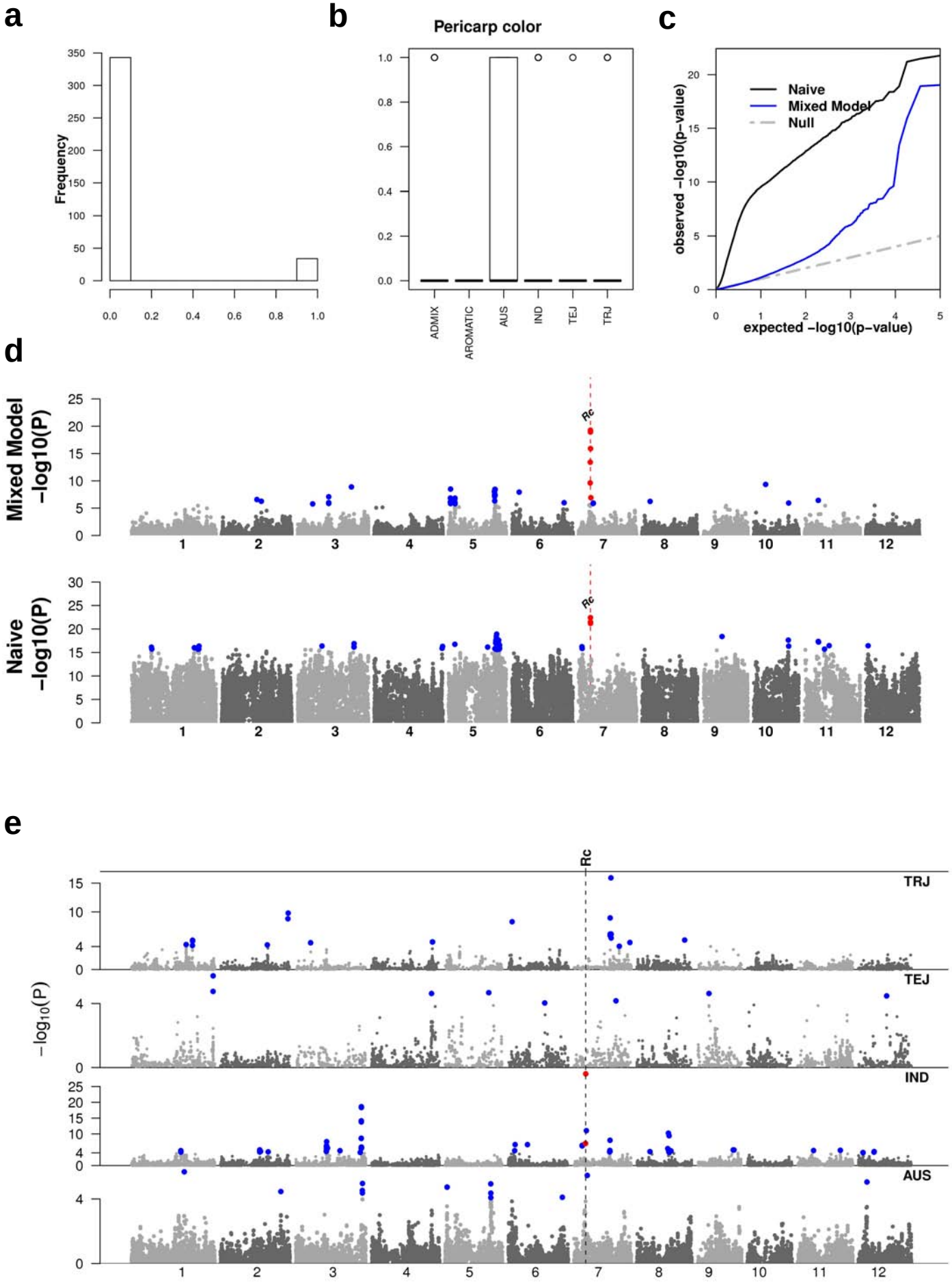
Supplementary Figure S28: Summary of GWAS results for Seed length/width ratio. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



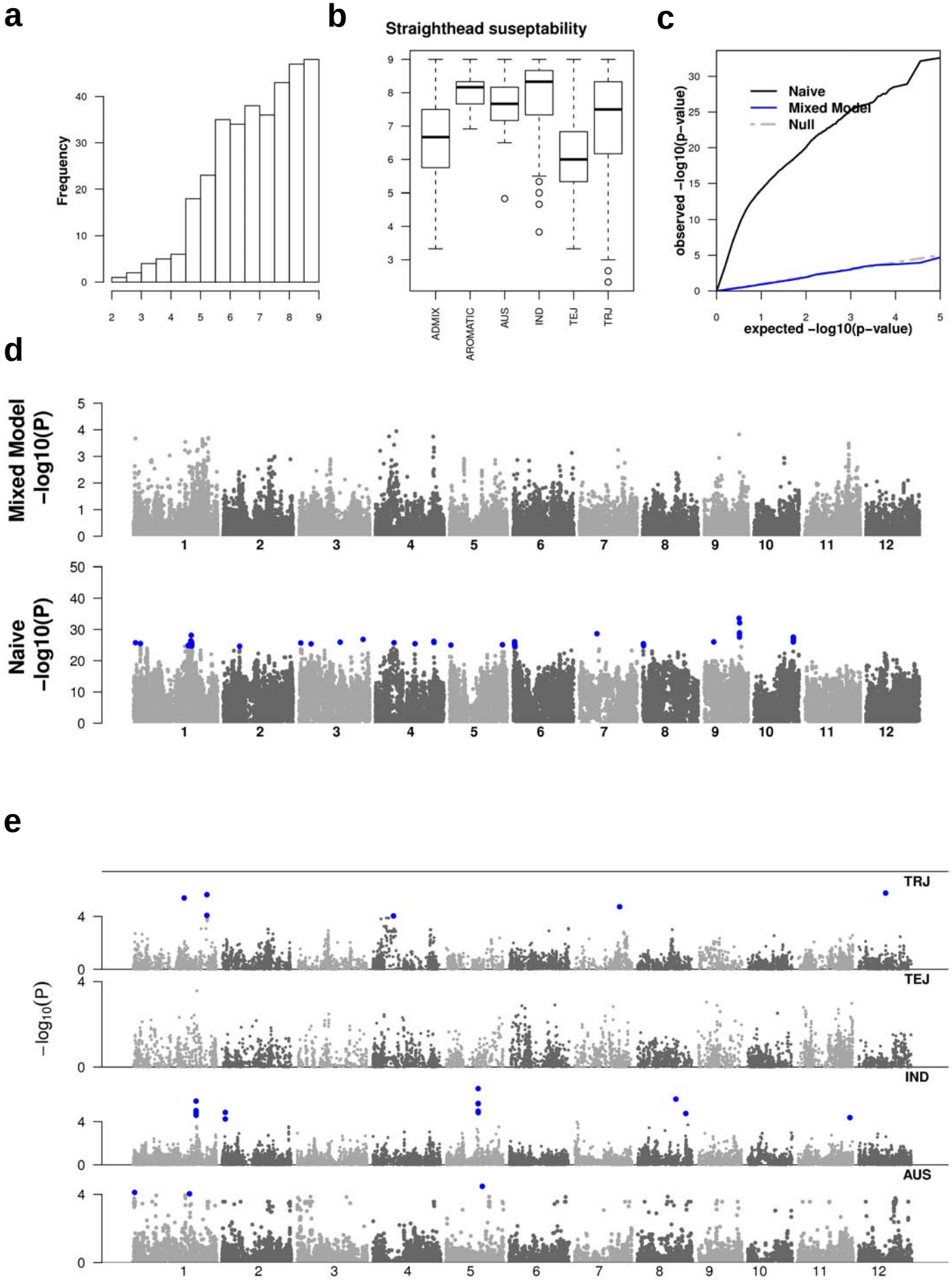
Supplementary Figure S29: Summary of GWAS results for Brown rice length/width ratio. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



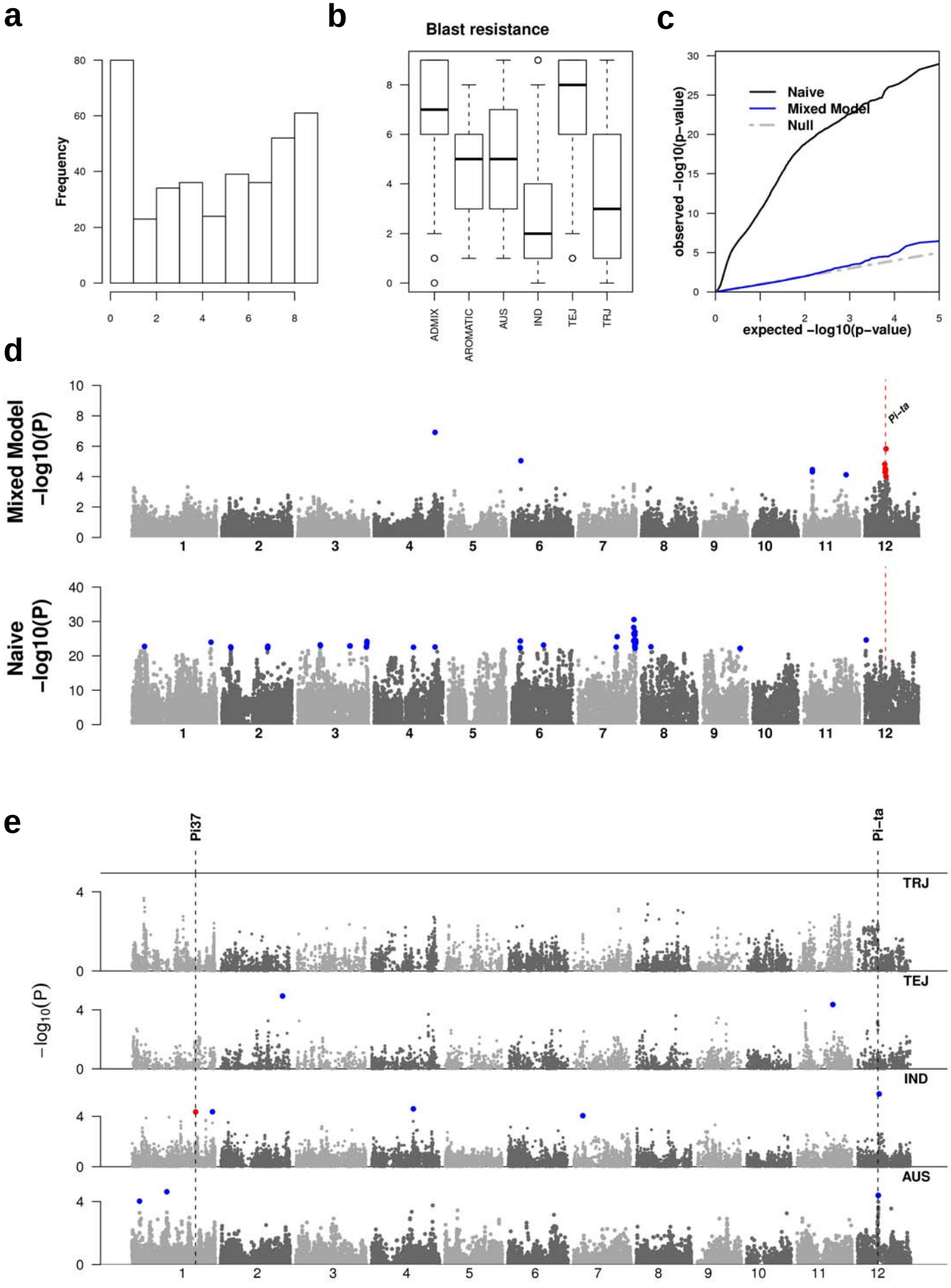
Supplementary Figure S30: Summary of GWAS results for Seed color . **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



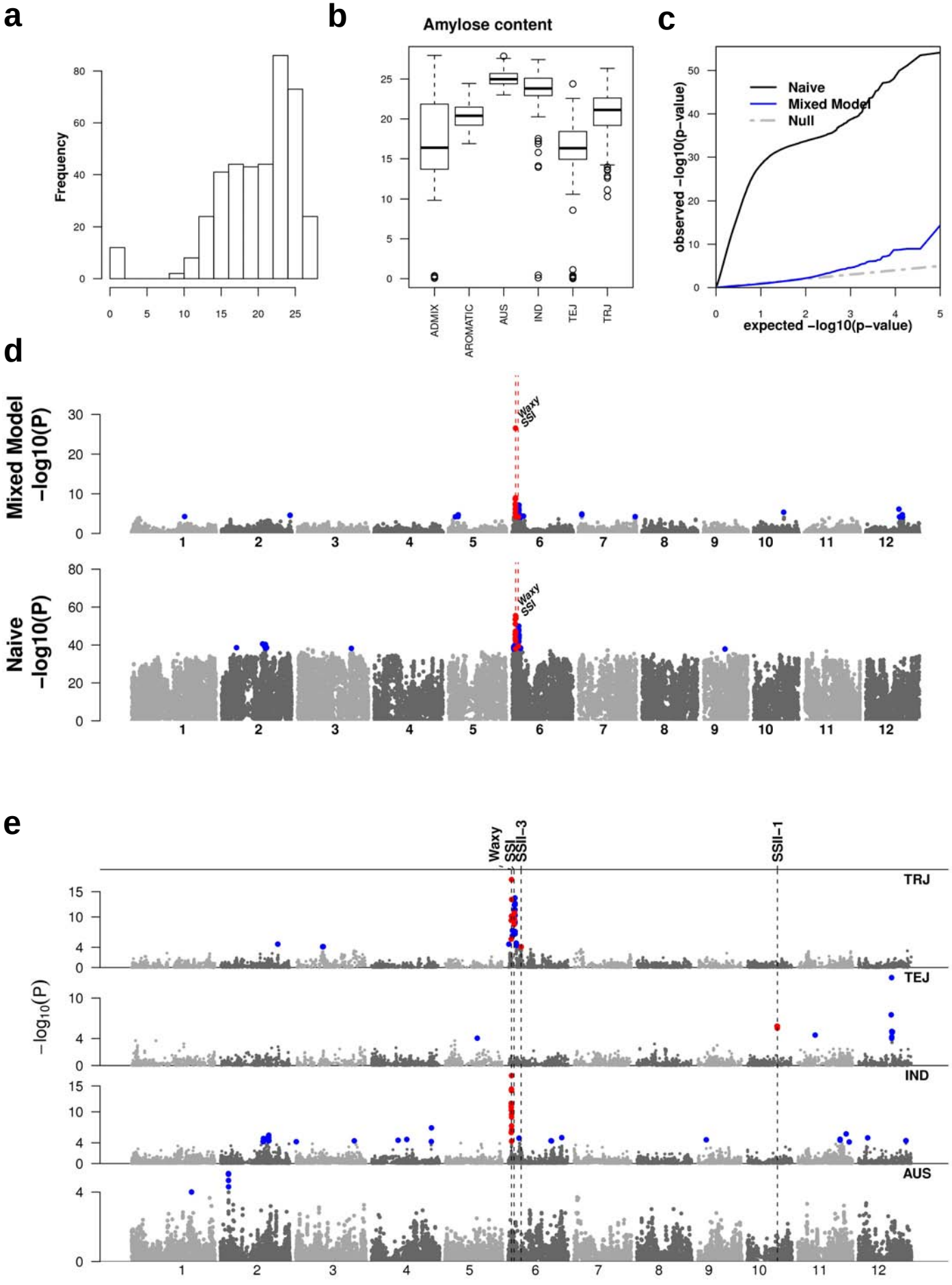
Supplementary Figure S31: Summary of GWAS results for Pericarp color. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



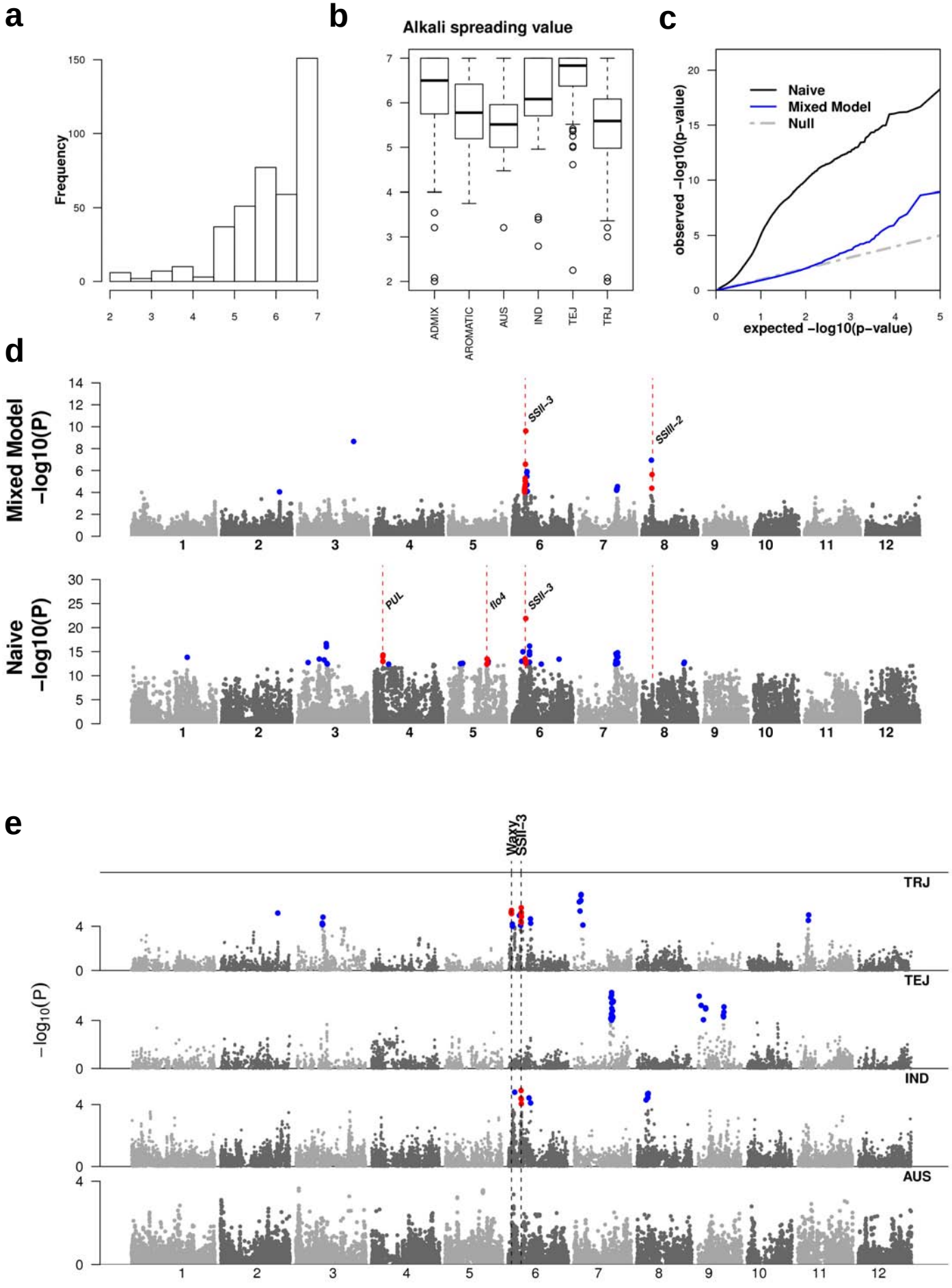
Supplementary Figure S32: Summary of GWAS results for Straighthead suseptability . (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.



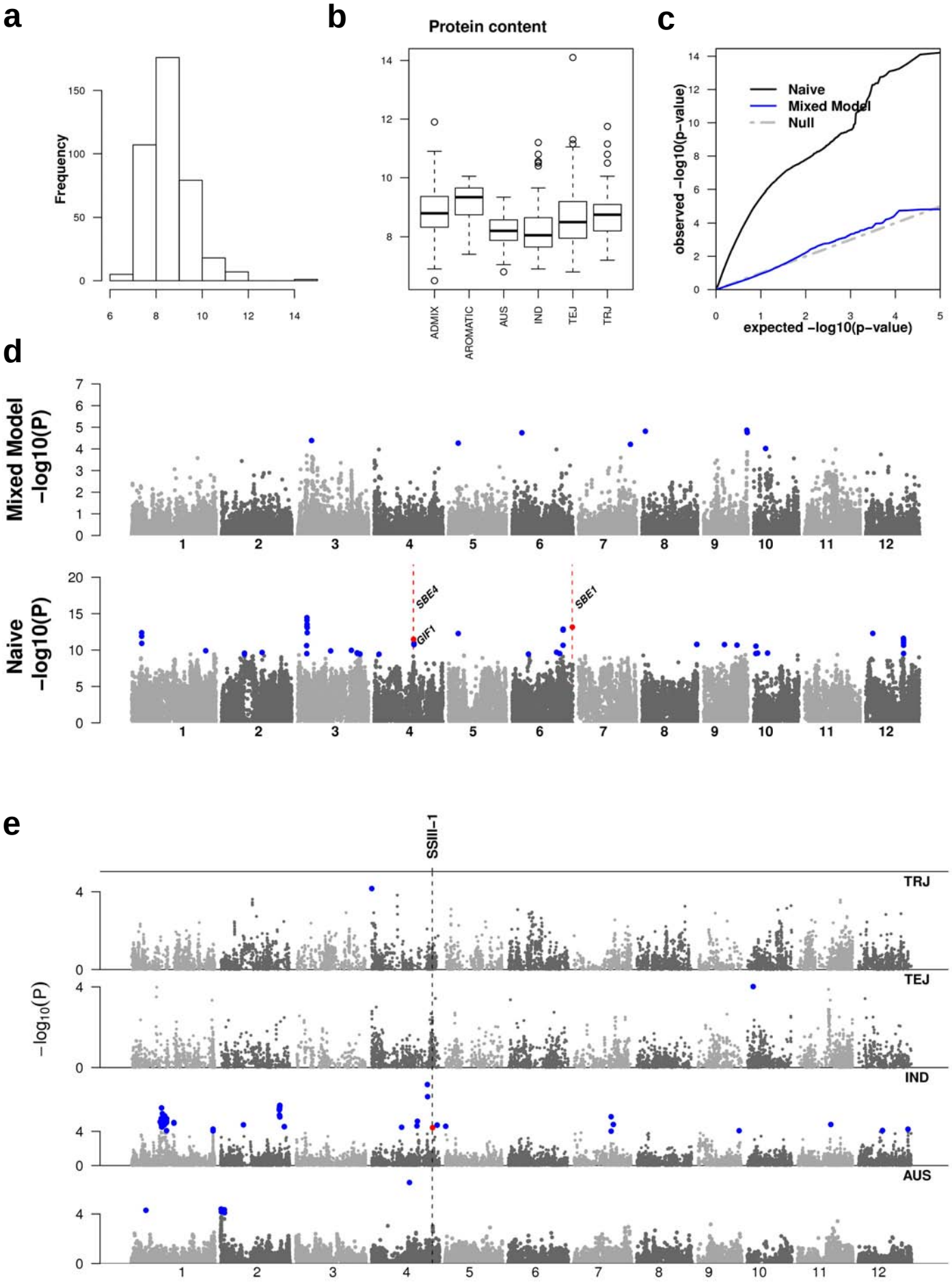
Supplementary Figure S33: Summary of GWAS results for Blast resistance. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.



Supplementary Figure S34: Summary of GWAS results for Amylose content. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.

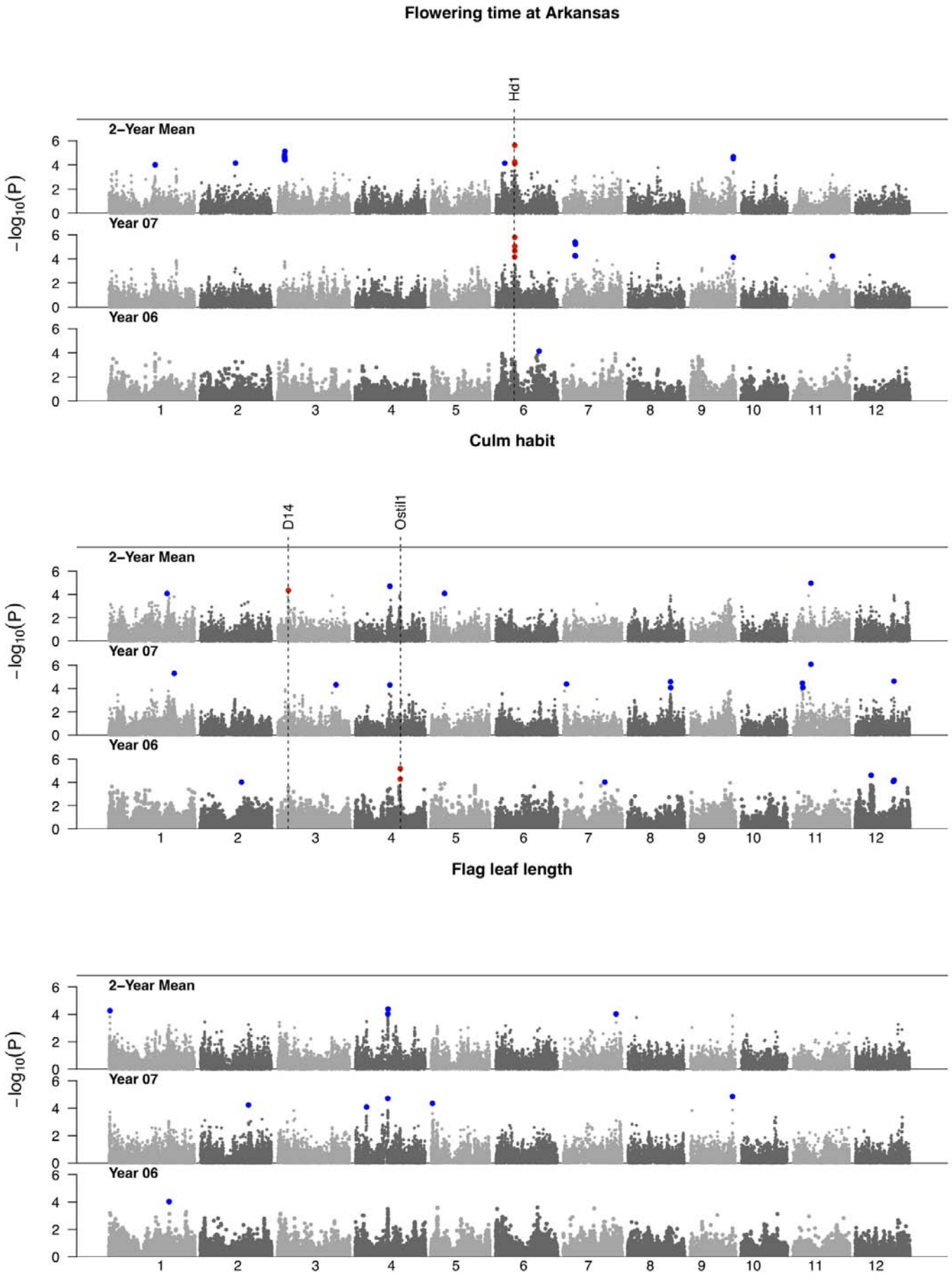


Supplementary Figure S35: Summary of GWAS results for Alkali spreading value. **(a,b,c)** Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. **(d)** Naive and Mixed Model results. **(e)** Mixed Model results for association within each subpopulation.

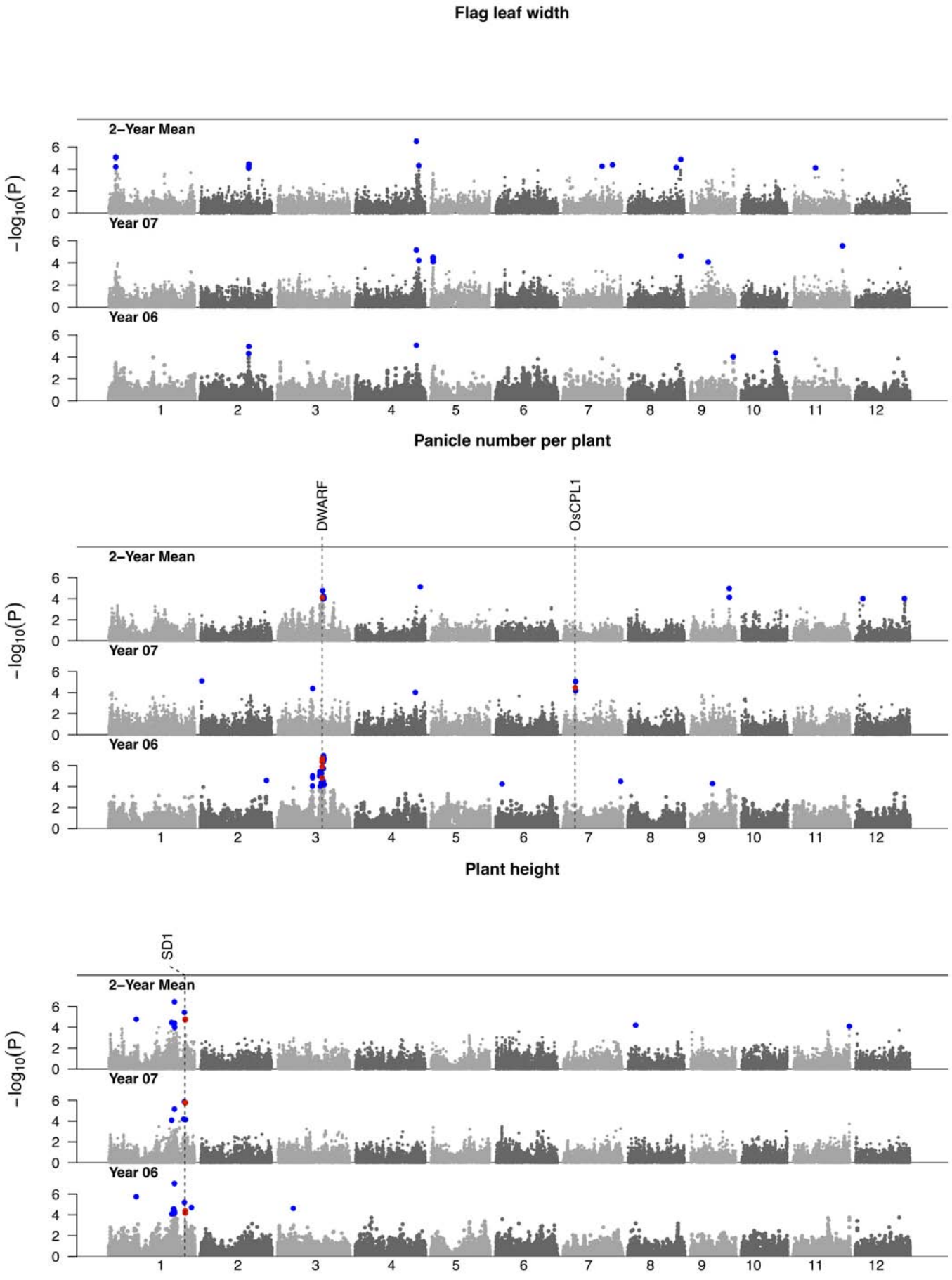


Supplementary Figure S36: Summary of GWAS results for Protein content. (a,b,c) Phenotype histogram, distribution of subpopulations and quantile-quantile plots of p-values. (d) Naive and Mixed Model results. (e) Mixed Model results for association within each subpopulation.

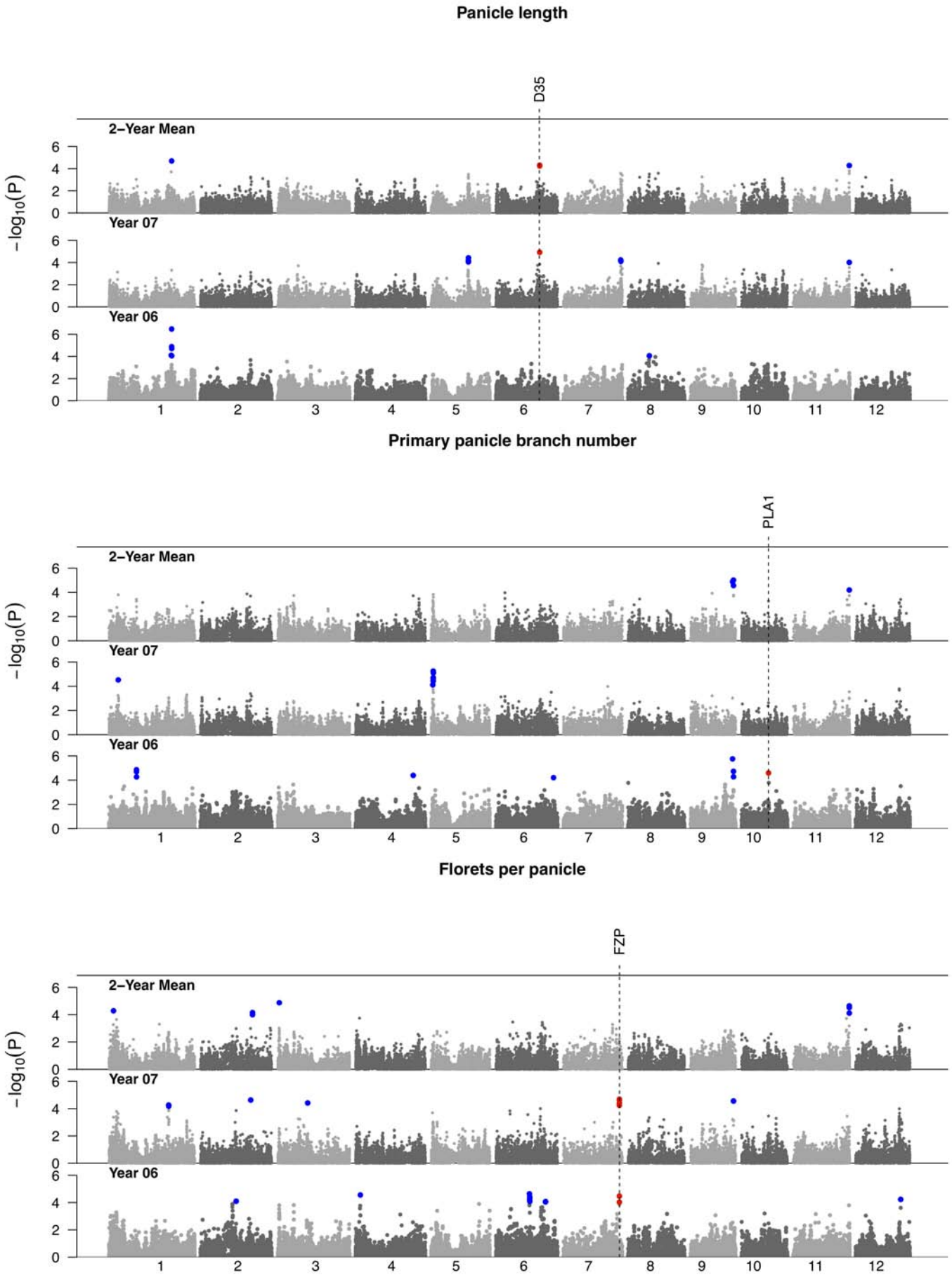
Supplementary Figure S37 to S40. Genome-wide association scan for eleven traits with broad sense heritability <0.8 using data from 2006, 2007 and the two year-mean. Genome-wide p-values from the mixed model for eleven traits evaluated on field-grown plants in Arkansas during 2006 and 2007. X-axis shows the SNPs along each chromosome; Y-axis is the $-\log_{10}$ (p-value) for each phenotype. Blue and red dots indicate SNPs with p-values $< 1 \times 10^{-4}$ in the mixed model and the top 50 SNPs in the naïve method; SNPs within 200kb range of known genes are in red; other significant SNPs are in blue. Candidate gene locations shown as red vertical dashed lines with names on top.



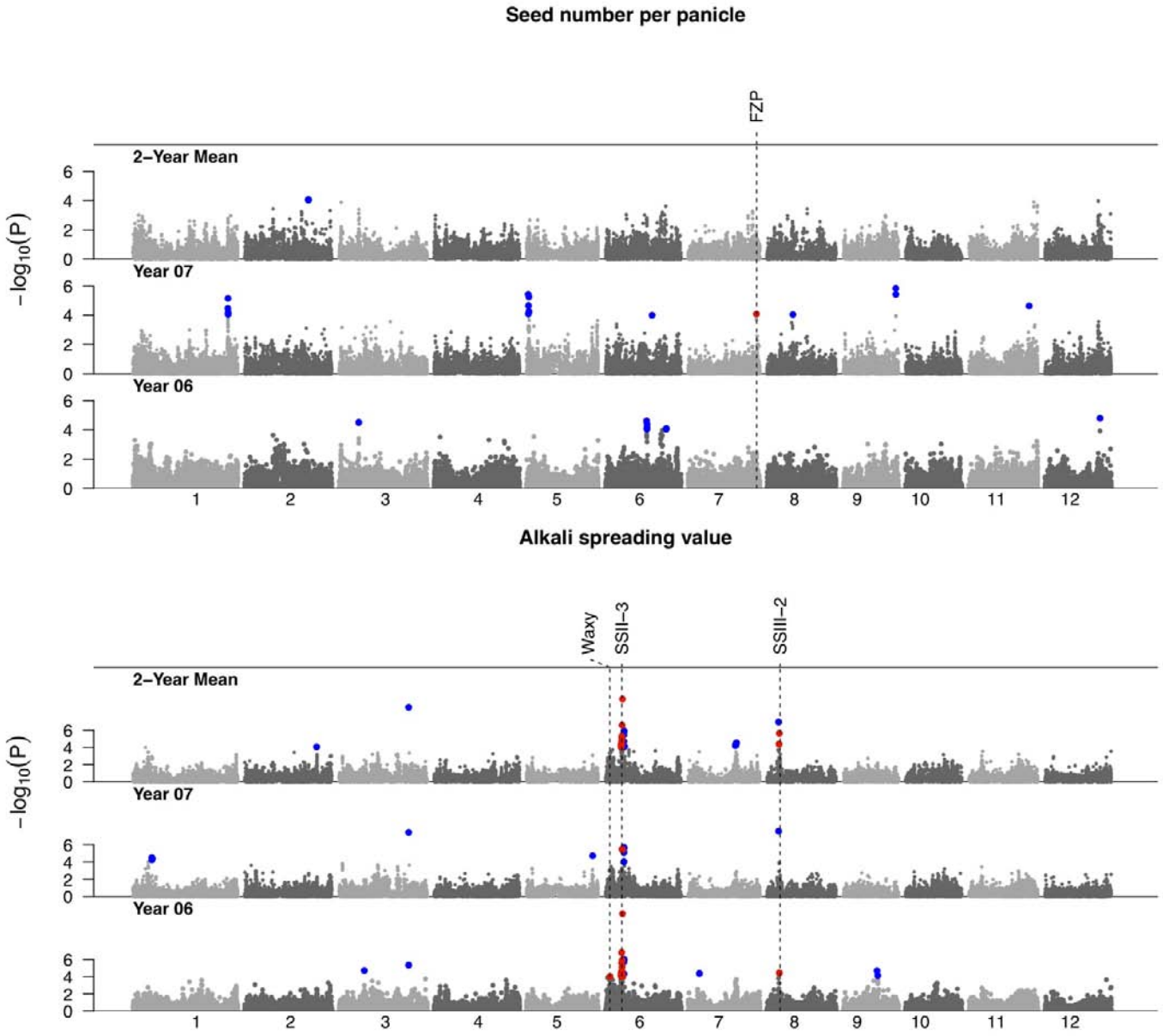
Supplementary Figure S37: GWAS of flowering time in Arkansas, culm habit and flag leaf length in year 2006, 2007 and using the average across two years.



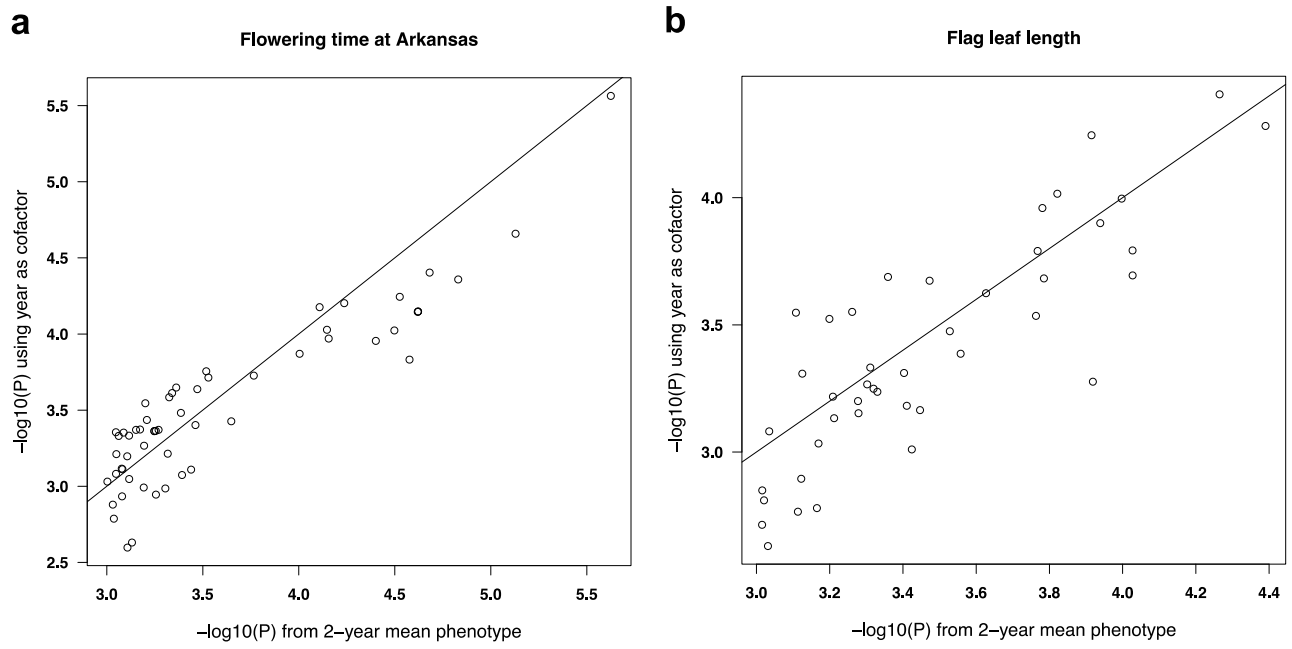
Supplementary Figure S38: GWAS of flag leaf width, panicle number per plant and plant height in year 2006, 2007 and using the average across two years.



Supplementary Figure S39: GWAS of panicle length, primary panicle branch number and floret per panicle in year 2006, 2007 and using the average across two years.



Supplementary Figure S40: GWAS of seed number per panicle and alkaline spreading value in year 2006, 2007 and using the average across two years.



Supplementary Figure S41. Correlation between significant SNPs for flowering time and flag leaf length identified using "year" as a cofactor in the regression model and p-values obtained using the 2-year phenotypic mean for traits: Flowering time at Arkansas (a) and Flag leaf length (b).

Supplementary Table S1. List of phenotypes used in this study and summary of year-effect and population structure-effect on phenotypes. Year effect is estimated as the correlation coefficient of the two-year mean values for each accession, when two-year data are available. V(Q)/V(P) is the percentage of phenotypic variance explained by the subpopulation structure.

Table S1. List of the traits used for genome-wide association mapping					
Trait category	**Trait name**	**Acronym**[(1)]	**Trait description**	**Correlation coefficient across years**[(2)]	**V(Q)/V(P): Phenotypic variance explained by the top 4 genotypic PCs**
Flowering time/Heading date	Days to heading (in Arkansas, USA)	DTHD	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of planting, measured at Stuttgart, Arkansas, US	0.63	19.1%
	Days to heading (in Faridpur, Bangladesh)	DTHD	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of transplanting, measured at Faridpur, Bangladesh	NA	2.4%
	Days to heading (in Aberdeen, Scotland)	DTHD	Number of days until the inflorescence is 50% emerged from the flag leaf counted from the day of planting, measured at Aberdeen, UK	NA	17.2%
	Ratio of heading date (Arkansas/Aberdeen)	PPDSN	Days to heading in Arkansas/flowering time in Aberdeen, UK	NA	10.3%
	Ratio of heading date (Faridpur/Aberdeen)	PPDSN	Days to heading in Faridpur/flowering time in Aberdeen, UK	NA	6.8%
Plant morphology	Culm habit	CULMHAB	Average culm angle of plants at maturity	0.52	38.7%
	Flag leaf length	FLFLG	Length of the flag leaf measured from leaf base to leaf tip (cm)	0.60	7.7%
	Flag leaf width	FLFWD	Width of the flag leaf measured at the widest portion of flag leaf lamina (cm)	0.73	37.8%
	Leaf pubescence	LFLPUBES	Presence or absence of hairs on the leaf blade surface	NA	14.6%
	Awn presence	AWNPLU	Presence or absence of awns; (whiskers on lemma) frequency and length of awns on seeds along a panicle	NA	5.1%
Yield-related	Panicle number per plant	PNNB	Average number of panicles (inflorescences) per plant	0.67	59.1%
	Plant height	PTHT	Height of plant from soil surface to tip of main panicle (inflorescence) (cm)	0.78	31.8%
	Panicle length	PNLG	Length of panicle (inflorescence) from the base to the tip (cm)	0.77	33.2%
	Primary panicle branch number	PBRNB	Number of primary branches along the panicle (inflorescence)	0.67	26.7%
	Florets per panicle	FLNBPPN	Average number of flowers (florets or spikelets) on main panicle (inflorescence)	0.50	9.9%

	Seeds per panicle	FILGRNB	Number of seeds per panicle (inflorescence), determined by counting the number of filled spikelets along the main panicle	0.69	15.6%
	Panicle fertility	PCENTST	Percent of spikelets that filled and produced seeds determined as the ratio of seeds per panicle/spikelets per panicle	NA	14.6%
Grain morphology-related [3]	Seed length	HULGRLG	Length of the seed with hull (palea and lemma)[4]	0.95	26.6%
	Seed width	HULGRWD	Width of the seed with hull (palea and lemma)	0.94	39.3%
	Seed volume	HULGRVOL	Volume of the seed with hull (palea and lemma)	0.94	29.5%
	Seed surface area	HULGRSURFAR	Surface area of the seed with hull (palea and lemma)	0.94	22.4%
	Seed length/width ratio	HULGRLGWDRO	Ratio of seed length/ seed width (with hull)	NA	35.5%
	Seed color	HULCL	Describes the color of the seed hull (lemma and palea); transformed to pigmented vs non-pigmented (binary trait)	NA	4.0%
	Grain length	DHULGRLG	Length of the unpolished rice grain (dehulled seed)[5]	0.96	32.7%
	Grain width	DHULGRWD	Width of the unpolished rice grain (dehulled seed)	0.94	41.5%
	Grain volume	DHULGRVOL	Volume of the unpolished rice grain (dehulled seed)	0.93	30.3%
	Grain surface area	DHULGRSURFAR	Surface area of the unpolished rice grain (dehulled seed)	0.93	24.1%
	Grain length/width ratio	DHULGRLGWDRO	Ratio of unpolished rice grain length/grain width (dehulled seed)	NA	36.9%
	Grain color	DHULGRCL	Color of unpolished rice grain, i.e., color of pericarp or fruit wall; transformed to pigmented vs non-pigmented (binary trait)	NA	22.5%
Cooking, eating and nutritional quality	Alkali spreading value	ASV	Observed by placing six milled-rice kernels in 10ml 1.7% KOH in a shallow container for 23hrs at 30 degree (°C) temperature and scoring for the extent of digestion of the starch based on its level of intactness. Measure for alkali digestion is inversely proportional to the gelatinization temperature, e.g. if alkali digestion is low, the gelatinization temperature is high. Estimation of ASV of each accession was done in the Rice Quality Lab, USDA-ARS, Stuttgart, Arkansas as described by Little et al. 1958[6]	0.60	18.2%
	Amylose content	AMYCN	Amount of amylose present in the milled grains. Estimation of amylose content of each accession was done in the Rice Quality Lab, USDA-ARS, Stuttgart, Arkansas as described by Gealy and Bryant 2009[7]	0.95	41.3%
	Brown rice protein content	DHULPROTCN	Protein content in brown rice (dehulled grain with pericarp). Estimation of protein content of each accession was done in the Rice Quality Lab, USDA-ARS, Stuttgart, Arkansas as described by Gealy and Bryant 2009 [7]	0.58	9.1%

Stress tolerance	Blast resistance	LFBLRS	Disease severity on rice leaf caused by the fungus *Pyricularia oryzae*. For blast resistance evaluation, the accessions were inoculated with a mixture of the U.S. blast races, IB-49, IC-17 and IE-1K (Jia et al. 2009) in a blast screening nursery at Beaumont, TX during 2009 and 2010 as described by Lee et al. 2003. The disease severity was scored on a "0" (no disease) to "9" (dead) scale when the plants were three to four weeks old as described by Marchetti et al. 1987[8]	0.66	30.3%
	Straighthead susceptibility	STRTHEAD	Physiological disorder symptoms on panicles caused mainly due to arsenic toxicity in the soil.The straighthead symptom was evaluated during 2007 and 2009 on the 353 accessions grown in a field pretreated with arsenic at a rate of 6.7kg/ha of monosodium methanearsonate (MSMA) and rated visually based on floret sterility and panicle emergence on a scale from 1 (normal panicle emergence and less than 20% sterility) to 9 (stubby plants with no panicle emergence) as described by Agrama and Yan (2009)[9] and Dilday et al. (2000)[10]	NA	29.0%

NOTES:

[1] The trait ontology acronyms according to Liang et al. (2008) as listed in Gramene (www.gramene.org/trait_ontology).

[2] The correlation between years is not available for traits measured in only one year, nor for derived traits.

[3] Grain-morphology related traits were measured on both seeds (grain + hull) and unpolished (brown) rice grains using the WinSeedle Pro V2007 software and STD4800 scanner (Regent Instruments Inc., Quebec, Canada).

[4] Seed = Mature spikelet filled with grain and surrounded by hull (palea and lemma)

[5] Unpolished rice grain = rice grain with pericarp (dehulled seed)

[6] Little et al (1958) Differential effect of dilute alkali on 25 varieties of milled white rice. Cereal Chem. 35:111-126

[7] Gealy D.R. and R.J. Bryant. (2009). Seed physiochemical characteristics of field grown US weedy red rice (*Oryza sativa*) biotypes: Contrasts with commercial cultivars. Journal of Cereal Science. 49:239-245

[8] Marchetti et al (1987) Inheritance of resistance to *Pyricularia oryzae* in rice cultivars grown in the United States. Phytopath. 77:799-804.

[9] Agrama and Yan (2009) Association mapping of straighthead disorder induced by arsenic in *Oryza sativa*. Plant Breed. 128:551-558

[10] Dilday et al (2000) Straighthead of rice as influenced by arsenic and nitrogen. In: R.J. Norman and C.A. Beyrouty (eds), B. R. Wells Rice Research Studies 1999, Univ. of Arkansas Arkansas Agric. Exp. Stn. Res. Ser. 476. p. 201-214.

Supplementary Method

Statistical models used in genome-wide association

For the naïve model, without correcting for population structure, a simple linear model was used for continuous traits with the following equation:

$$Y = \beta X + \varepsilon \quad \text{(S1)}$$

and logistic regression was used for binary traits with the following equation:

$$\text{logit}[\frac{P(Y)}{1-P(Y)}] = \beta X + \varepsilon \quad \text{(S2)}$$

For the mixed model, we used software implemented in the R package EMMA. In analysis with all samples, we used the equation:

$$Y = \alpha X + \beta P + u + \varepsilon \quad \text{(S3)}$$

In analysis within each subpopulation, we used the equation:

$$Y = \alpha X + u + \varepsilon \quad \text{(S4)}$$

Where Y represents the vector of phentoype, X the vector of SNP genotype, P the matrix of 4 top principle components (PCs), α the SNP effect, β the PC effects; For random effects. $u \sim N(0,\sigma_g^2 K)$, $\varepsilon \sim N(0,\sigma_e^2 I)$, K the kinship matrix estimated as the IBS matrix from the whole genome genotypes.

To estimate the phenotypic variance contribution of significant loci, ANOVA was used with contrasting linear models of $Y = \alpha X + \beta P + \varepsilon$ and $Y = \beta P + \varepsilon$. Thus, the variance due to each SNP is reported after adjusting for the population structure effects. When all m SNP loci are considered, ANOVA was used with contrasting linear models of $Y = \alpha \sum_{i=1}^{m} X_i + \beta P + \varepsilon$ and $Y = \beta P + \varepsilon$.