pipette.bio

SCIENTIFIC ANALYSIS REPORT

Flowering-Time GWAS in an Inbred Rice Breeding Panel

Session 4d7b3d10 · Started 26 September 2026

Exported 26 September 2026 · 1 analysis queries

Contents

QUERY 1

Flowering-Time GWAS in an Inbred Rice Breeding Panel

RESEARCH REQUEST

Run a flowering-time GWAS using these two uploaded datasets from this rice breeding program dataset:

- MET_crfilt_.75_allchrom.hmp.txt.zip: imputed SNP genotypes in HapMap format for rice breeding lines. The marker coordinates use the Nipponbare MSU v6.0 reference.
- RYT_plotdata_by_GHID_corrected_PUBLIC_ACCESS.zip: plot-level field measurements from wet and dry seasons during 2009–2012, including flowering time, plant height and grain yield.

Inspect the files, identify the flowering-time variable and match genotype and phenotype sample IDs. Report sample counts, marker counts and exclusions.

Perform QC appropriate for an inbred rice breeding panel. Prepare trial-adjusted flowering-time phenotypes, accounting for the available trial design, years, seasons and genotype × environment variation. Do not treat repeated plot measurements as independent genotypes. Explain and justify the phenotype model.

Run a mixed-model GWAS accounting for population structure and kinship. Produce PCA, Manhattan and QQ plots, report the multiple-testing threshold, and provide full association results and significant-region summaries. Use local LD to assess whether nearby significant markers represent distinct signals.

Use Nipponbare MSU v6.0 in the link https://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.0/all.dir/all.gff3.gz and matching MSUv6 gene annotations from https://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.0/all.dir/all.TU_model_brief_info.6.0.gz throughout. DO NOT CHANGE THE REFERENCE ANNOTATION. Check gene overlap, strand and distances before interpreting candidate genes. If matching annotations cannot be obtained, report that limitation rather than substituting another assembly.

Save the analysis code, processed inputs, QC summaries, association tables and figures. Summarize the findings and limitations, distinguishing positional candidate genes from demonstrated causal genes.

ANALYSIS PLAN

Run a reproducible flowering-time GWAS for an inbred rice breeding panel by modeling repeated multi-environment plot records into trial-adjusted line phenotypes, performing TASSEL kinship- and structure-aware mixed-model association, resolving local LD signals, and interpreting loci only against the exact Nipponbare MSU v6.0 annotations specified by the user.

1. Inspect archives and resolve the GWAS execution contract

The GWAS dispatcher must establish the population design, inheritance/ploidy, observation unit, trait representation, sample identifiers, genotype encoding, assembly, and matching before any backend is run. Inspect and unpack both authorized archives; inventory files and schemas; identify the flowering-time field from names, labels, documentation, and value structure; identify genotype and phenotype identifiers; characterize missingness, duplicate IDs, years, seasons, environments, trials, blocks/replicates, and repeated plot records; inspect the HapMap genotype encoding and chromosome/position fields; and write a gwas-dispatch-v1 receipt. The receipt must select a diploid plant-panel TASSEL contract only if the evidence confirms an inbred rice association panel rather than a linkage cross or unsupported inheritance. Any blocker or contradictory design evidence must stop later GWAS execution rather than trigger a scientifically different fallback.

Parameter	Planned value
dispatcher	gwas-dispatch-v1
species	Oryza sativa
assembly	Nipponbare MSU v6.0
population design	inbred rice breeding panel, subject to inspection confirmation
independent observation unit	genotype/breeding line, not plot
trait type	quantitative flowering time
candidate backend	TASSEL plant kinship MLM, conditional on dispatch approval

2. Acquire and validate the exact MSU v6.0 annotation resources

Download only the two user-specified Nipponbare MSU v6.0 resources, preserve their URLs and checksums, decompress them into processed reference inputs, inspect GFF3 feature/attribute conventions and TU annotation columns, and verify that chromosome identifiers and coordinate conventions can be reconciled exactly with the HapMap markers without changing assembly or substituting annotations. If retrieval or identity validation fails, record the limitation and retain GWAS results without candidate-gene substitution.

Parameter	Planned value
gff3 url	https://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.0/all.dir/all.gff3.gz

Parameter	Planned value
gene info url	https://rice.uga.edu/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_6.0/all.dir/all.TU_model_brief_info.6.0.gz
reference policy	exact MSU v6.0 only; no assembly or annotation substitution

3. Reconcile samples and perform inbred-panel genotype and phenotype QC

Using the dispatch receipt, create auditable genotype, phenotype, environment, and sample-matching tables. Match identifiers exactly first and apply only documented, reviewable normalization if the files themselves establish it. Report raw and retained genotype samples, phenotype genotypes, overlap, unmatched and duplicate IDs, plot counts, environment/trial counts, raw and retained marker counts, missingness, allele frequency, monomorphic markers, heterozygosity/inbred-line consistency, chromosome/position validity, and every exclusion with a reason. Use the approved TASSEL-compatible HapMap preparation path; do not use HWE filtering as a default exclusion criterion for an inbred breeding panel and do not silently remove related lines merely for being related.

Parameter	Planned value
qc context	inbred rice breeding panel
marker qc thresholds	use and report the packaged plant-panel inspector/runner defaults after inspection; preserve an explicit pre/post-filter ledger
sample matching policy	exact ID matching first; only evidence-backed normalization
hwe policy	do not use HWE as a routine exclusion filter for this inbred panel
relatedness policy	retain related breeding lines and model kinship

4. Fit the multi-environment flowering-time phenotype model

Fit a trial-aware mixed model to plot-level flowering-time observations, after resolving actual design columns. Represent year-season trial environments explicitly; model year, season, and their estimable interaction/environment structure; include available replicate, block, or incomplete-block terms nested within trial as supported by the file; and model genotype-by-environment variation. Produce one adjusted flowering-time estimate per breeding line, with uncertainty/reliability and environment-specific diagnostics, so repeated plots contribute through the trial model rather than acting as independent GWAS observations. Check rank, confounding, sparse cells, residual behavior, and sensitivity to unusually influential environments. Save the exact formula and justify every retained or omitted design term from the observed metadata.

Parameter	Planned value
phenotype model	trial-adjusted multi-environment mixed model with genotype as the line-level target, year/season or year-season environment effects, available within-trial design terms, and genotype-by-environment variation
replication unit	breeding line/genotype
years	2009–2012

Parameter	Planned value
seasons	wet and dry
line effect output	one adjusted line estimate for GWAS, with model-derived uncertainty and diagnostics

5. Run the approved TASSEL mixed-model GWAS and structure analysis

Consume the approved dispatch receipt, QC-passed HapMap data, matched trial-adjusted line phenotype, and exclusion ledger. Run the packaged TASSEL workflow through run_pipeline.pl only if the receipt confirms the diploid plant-panel contract. Compute LD-pruned genotype PCA and a kinship matrix, inspect how leading PCs relate to the breeding-panel structure, select and record the retained PC covariates without dropping the kinship term, and fit a TASSEL mixed linear model with population structure and kinship. Export complete association statistics, tested-marker inventory, genomic-inflation/QQ diagnostics, PCA coordinates and plots, Manhattan and QQ plots, and the reported genome-wide multiple-testing threshold. Backend success alone must not be treated as validation.

Parameter	Planned value
backend	TASSEL mixed linear model
association model	mixed model with computed population-structure PCs and kinship
multiple testing	Bonferroni family-wise threshold at alpha 0.05 over successfully tested markers, with the exact denominator reported
alpha	0.05
reference	Nipponbare MSU v6.0
required outputs	PCA, Manhattan plot, QQ plot, full association results, threshold, significant-region summaries

6. Resolve significant regions with local LD

For each significant locus, compute local pairwise LD from the same QC-passed, matched inbred-panel genotypes, generate region-level LD tables/heatmaps and local association views, and group nearby significant markers according to observed LD connectivity rather than physical proximity alone. Report whether nearby lead markers are highly correlated representatives of one signal or show sufficiently weak LD to remain plausible distinct signals; retain uncertainty where marker density or sample size limits resolution.

Parameter	Planned value
ld scope	local windows centered on genome-wide significant regions, with boundaries chosen from observed neighboring significant markers and local marker density
ld metric	pairwise r-squared
signal policy	assess distinctness from local LD rather than distance alone

7. Annotate significant regions against exact MSU v6.0 genes

Intersect significant markers and LD-defined regions with the validated MSU v6.0 GFF3 and matching TU brief annotation. For each marker/region, report overlapping genes and nearest upstream/downstream genes, exact coordinate overlap, feature type, gene identifier, strand, signed and absolute distances, and annotation text. Verify chromosome naming and one-based coordinate semantics before overlap. Interpret only successfully mapped MSUv6 records; if the matching resources were unavailable or irreconcilable, state that limitation and do not substitute another rice assembly or annotation.

Parameter	Planned value
annotation release	Nipponbare MSU v6.0
gene checks	overlap, strand, and distances
candidate claim policy	positional candidates only unless independent causal evidence exists

8. Assemble reproducible deliverables and final report

Save executable analysis code, environment/tool versions and seeds where applicable, unpacked/processed input manifests, exact reference download provenance and checksums, sample/marker matching tables, all exclusion ledgers, QC summaries, phenotype-model formula and diagnostics, adjusted line phenotypes, PCA/kinship artifacts, full GWAS tables, multiple-testing calculation, LD tables and region summaries, MSUv6 annotations, and publication-ready figures. Produce a concise findings report covering sample and marker attrition, phenotype-model justification, association results, LD-defined signal interpretation, positional candidates, and limitations, explicitly avoiding causal-gene claims not supported by the analysis.

Parameter	Planned value
deliverable scope	analysis code, processed inputs, QC summaries, association tables, figures, findings, and limitations
causal language	distinguish positional candidate genes from demonstrated causal genes

Planning assumptions

- The description supports treating these materials as an Oryza sativa inbred breeding association panel with diploid inheritance, but step_1 must verify that contract from the files and stop if it instead reveals a linkage cross, unsupported ploidy/inheritance, or ambiguous observation unit.
- TASSEL MLM is the skill default for a diploid crop association panel; it is conditional on the gwas-dispatch-v1 receipt and is not a fallback for a contradictory design.
- The exact flowering-time column, identifier columns, plot-design columns, genotype encoding, QC thresholds exposed by the packaged runner, and estimable phenotype-model formula will be resolved from file inspection and recorded rather than guessed from filenames.
- The Bonferroni alpha of 0.05 is a planner-inferred family-wise threshold because the user requested a reported multiple-testing threshold but did not prescribe a method or alpha; the report will give the exact tested-marker denominator.

- Relatedness is expected in a breeding panel and will be modeled through kinship rather than used as an automatic exclusion criterion.
- No alternate genome build or gene annotation may be substituted if the specified MSU v6.0 resources cannot be downloaded or reconciled.

ANALYSIS REPORT

Flowering-Time GWAS in an Inbred Rice Breeding Panel

Nipponbare MSU v6.0 · TASSEL Mixed Linear Model · 2009–2012 Multi-Environment Trials

Executive Summary

This analysis performed a flowering-time (FL) genome-wide association study (GWAS) on an inbred Oryza sativa breeding panel comprising **349 genotyped and phenotyped lines**, using imputed SNP genotypes in HapMap format and plot-level FL observations from **eight year-season trials (2009–2012, dry and wet seasons)**.

A multi-environment mixed model was fitted to raw plot observations to produce one trial-adjusted line estimate (BLUP) per breeding line. A TASSEL mixed linear model (MLM) with five population-structure PCs and an IBS kinship matrix was then used to test **69,833 quality-filtered markers**.

One genome-wide significant locus was identified on chromosome 3 (Chr3: 234,156–1,935,861 bp), containing 119 Bonferroni-significant markers ($\alpha = 0.05 / 69{,}833 = 7.16\times10^{-7}$) and 157 FDR < 0.05 markers. All significant markers form a single high-LD block (mean pairwise $r^2 = 0.74$), indicating one major association signal. The lead SNP (S3_1269941, $p = 1.41\times10^{-17}$) falls within **LOC_Os03g03070** (annotated as "transcription factor, putative, expressed" in MSU v6.0), which is the closest positional candidate gene. **No causal gene can be claimed from this analysis alone; LOC_Os03g03070 and all 52 other overlapping MSU v6.0 genes in the region are positional candidates only.**

The analysis has several important limitations: all genotypes are imputed (imputation quality was not independently assessed); 35.2 % of markers were filtered at MAF < 5 %; the phenotype model is based on an unbalanced multi-environment design; and the locus spans a large region (1.7 Mb) containing 275 annotated genes, requiring fine-mapping to narrow the causal interval.

Confidence level: EXPLORATORY. The association signal is strong and reproducible within the panel; independent validation in a separate population and functional characterisation of candidate genes are needed before biological conclusions can be drawn.

Methods

Input data

Resource	Details
Genotype file	`MET_crfilt_.75_allchrom.hmp.txt.zip` — imputed SNP calls, HapMap format
Phenotype file	`RYT_plotdata_by_GHID_corrected_PUBLIC_ACCESS.zip` — 8 per-trial CSVs
Reference assembly	Nipponbare MSU v6.0 (osa1r6), O. sativa ssp. japonica
GFF3 URL	https://rice.uga.edu/.../all.gff3.gz (SHA-256 in manifest)
TU brief info URL	https://rice.uga.edu/.../all.TU_model_brief_info.6.0.gz

Genotype QC and preprocessing

The raw HapMap file contained 370 columns after the 11 fixed fields. Two non-sample columns (`BLANK`, `call_rate`) were identified and removed, leaving 368 true sample columns. Ten samples present in the genotype file had no phenotype match in any trial and were excluded, giving 358 matched samples. The HapMap file was cleaned to: (1) replace 56,731 `X` allele values with `NA` (adapter-compatible missing code); (2) remove 19 markers with single-nucleotide `alleles` fields (monomorphic/malformed); (3) recode 1,215 individual genotype cells that carried a third nucleotide not in the declared two-allele set to `NA`. The resulting `MET_matched.hmp.txt` was used for all downstream analysis.

Marker-level QC thresholds (applied by TASSEL skill preflight):

Filter	Threshold	Removed
Locus call rate	< 75 %	85 markers
Monomorphic (after call rate filter)	—	65 markers
Minor allele frequency	< 5 %	38,022 markers
Retained testable markers		**69,833**

Sample-level QC:

Filter	Threshold	Removed
Call rate	< 80 %	9 samples (A1287, B1028, B1058, B1083, B1209, B1247, M1376, M1410, M1421)
Retained analysis samples		**349**

HWE filtering was **not applied** — Hardy–Weinberg equilibrium assumptions are violated in a structured inbred breeding panel and are not a valid exclusion criterion here. Relatedness was modelled via kinship rather than used as an exclusion criterion.

Phenotype model

Formula: `FL ~ trial + (1|GHID) + (1|GHID:trial) + (1|trial:REP)`
Software: R `lme4` 1.x, REML, bobyqa optimiser
Trait: FL — days to 50 % heading (flowering time)
Data: 5,799 plot records after removing 4 FL-missing values and 3 exact-duplicate rows for B1203 in 2012WS

Term	Role	Justification
`trial` (8 levels)	Fixed	Absorbs all year, season, and year×season environment main effects
`(1|GHID)`	Random	Target: one BLUP per line for GWAS input
`(1|GHID:trial)`	Random	Genotype × environment variance; prevents GxE from inflating line uncertainty
`(1|trial:REP)`	Random	Within-trial replicate/block nuisance effects

Six entries (5 GHIDs in 2010WS, M1396 in 2011WS) had six plots per trial with distinct PLOTNO values; these were treated as additional within-trial replications, not duplicates. B1203 in 2012WS had three exact-duplicate rows (identical PLOTNO and FL) — duplicates removed.

Variance component table:

Component	σ^2	% of total
Genotype (GHID)	13.211	60.7 %
Genotype × Environment	5.750	26.4 %
Replicate (trial:REP)	0.057	0.3 %
Residual	2.734	12.6 %
Broad-sense H^2 (harmonic E=5.3, $\bar{r}$=2.9)	**0.913**	

Mean BLUP reliability = 0.913; all 349 BLUPs had reliability ≥ 0.80.

Association analysis

Backend: TASSEL 5, mixed linear model (MLM), P3D variance-component estimation
Kinship: Centered IBS computed from the filtered HapMap
Structure covariates: 5 PCs computed by TASSEL from the filtered HapMap
Multiple testing: Bonferroni family-wise threshold α = 0.05 / 69,833 = **7.160×10^{-7}**

(denominator = successfully tested markers; no failed regressions reported)
BH FDR: computed post-hoc, threshold 0.05
Genomic inflation: λ_GC = 0.892 (slightly below 1.0, reflecting conservative behaviour of the kinship+PC model; not indicative of deflation of true signals)

LD analysis

Pairwise r^2 was computed from the QC-filtered matched inbred-panel genotypes (349 retained samples) across a 2.4 Mb window centred on the significant region, using mean imputation for residual missing calls.

Annotation

MSU v6.0 GFF3 chromosome identifiers ("Chr3") were reconciled with HapMap coordinates ("3") — both use 1-based positions, verified by matching the first gene (LOC_Os01g01010, 1,903–9,817 bp) between GFF3 and TU brief info. Significant marker positions were intersected with gene features (type = "gene") to determine overlapping genes and nearest upstream/downstream genes. All annotation text comes from the downloaded MSU v6.0 TU brief info file; no annotation was inferred from training knowledge.

Results

Sample and marker attrition

Stage	Samples	Markers
Raw HapMap	370 (incl. 2 non-sample cols)	108,024
After non-sample removal	368	108,024
After phenotype matching	358	108,024
After preprocessing (X→NA, monomorphic cols, mismatch cells)	358	108,005
After sample call-rate QC (< 80 %)	**349**	108,005
After marker QC (call rate, mono, MAF)	349	**69,833**

Phenotype model diagnostics

- Residuals approximately normal (Q-Q plot; see Fig 2)
- No non-linearity in residuals vs fitted (Fig 1)
- Trial fixed effects ranged 84.4 days (2012DS) to 92.7 days (2012WS); WS > DS in all years (Fig 3)
- FL BLUPs ranged 80.8–98.9 days, mean 88.6 ± 3.5 days (Fig 4)
- Pearson r between BLUPs and raw per-line means = 0.9975 (Fig 5)

Population structure (PCA)

PC1 explains 14.6 % of genotypic variance, PC2 4.1 % (Fig 9), indicating substantial population structure in the panel. Lines with later FL tend to cluster at higher PC1 values. Five PCs were included as fixed covariates in the TASSEL MLM.

GWAS results

Metric	Value
Markers tested	69,833
Bonferroni threshold	7.160×10^{-7}
λ_GC (MLM)	0.892
Genome-wide significant (Bonferroni)	**119 markers**
FDR < 0.05	157 markers
Chromosomes with significant hits	**Chr3 only**
p-value range (significant)	1.41×10^{-17} – 6.27×10^{-7}
Lead SNP	S3_1269941 (Chr3:1,269,941 bp)
Lead p-value	**1.41×10^{-17}**

Manhattan plot: Fig 7
Q-Q plot: Fig 8
Chromosome 3 zoom: Fig 10

All 119 genome-wide significant markers and all 157 FDR < 0.05 markers are on chromosome 3. No suggestive associations ($p < 1\times10^{-5}$) were detected on other chromosomes. The concentration of signal on a single chromosome in a mixed model with kinship and PCs is a strong indicator of genuine genetic association rather than population-structure confounding, but this cannot be ruled out without validation.

LD analysis and signal interpretation

Metric	Value
Significant region span	Chr3: 234,156–1,935,861 bp (1.702 Mb)
Analysis window	Chr3: 0–2.4 Mb
Markers in window	717
Mean r^2 among 119 significant marker pairs	0.740
Median r^2	0.738
Pairs with $r^2 > 0.8$	2,258 / 7,021 (32.2 %)
Pairs with $r^2 < 0.1$	0 / 7,021

Metric	Value
LD gaps > 1 Mb between consecutive sig. markers	None

LD heatmap: Fig 11
LD decay: Fig 12

Signal interpretation: ONE major locus. The absence of any marker pairs with $r^2 < 0.1$ and the continuous span without internal LD gaps support a single LD block rather than multiple independent signals. Fine-mapping at higher marker density would be required to distinguish one causal variant from a cluster of highly correlated variants.

MSU v6.0 annotation

Chromosome naming was reconciled: GFF3 uses "Chr3", HapMap uses "3", TU brief info uses "chr03"; all coordinates are 1-based and internally consistent (verified by coordinate comparison of the first annotated gene).

Annotation metric	Value
Genes overlapping significant markers	53 distinct loci
Markers with at least one overlapping gene	75 / 119
Total genes in significant region (234–1,936 kb)	275
Annotation source	MSU v6.0 TU_model_brief_info (from rice.uga.edu)

Lead SNP (S3_1269941) overlaps:

Gene	Coordinates (Chr3, 1-based)	Strand	MSU v6.0 Annotation
LOC_Os03g03070	1,268,854–1,270,781	−	transcription factor, putative, expressed

The lead SNP lies 1,088 bp from the gene start (on the annotated − strand).

Other notable positional candidates (genes with potentially relevant functional classes):

Gene	Coordinates (bp)	Annotation
LOC_Os03g02970	1,194,073–1,203,837 (−)	Dicer, putative, expressed
LOC_Os03g03164	1,339,050–1,344,293 (+)	homeobox protein knotted-1, putative, expressed
LOC_Os03g03260	1,393,427–1,399,137 (+)	homeobox domain containing protein, expressed
LOC_Os03g03550	1,550,422–1,552,817 (+)	bZIP family transcription factor, putative, expressed

Gene	Coordinates (bp)	Annotation
LOC_Os03g03760	1,671,829–1,676,807 (−)	MYB family transcription factor, putative, expressed
LOC_Os03g01910	559,208–561,683 (−)	transcription factor BTF3, putative, expressed
LOC_Os03g01850	509,593–513,784 (+)	cyclin-dependent kinase A-1, putative, expressed

Full candidate gene list: results/chr3_locus_candidate_genes.tsv
Annotated locus figure: Fig 13

IMPORTANT: All 275 genes in the region and all 53 directly overlapping genes are **positional candidates only**. No independent causal evidence (eQTL, allelic series, functional assay, or validated marker) exists from this analysis. Functional annotation from TU brief info is retrieved metadata — it identifies gene class, not causal role.

Interpretation

The analysis identified a single major QTL for flowering time on O. sativa chromosome 3, in the chromosomal region 0.23–1.94 Mb (Nipponbare MSU v6.0 coordinates). This is a **statistical observation**: the lead SNP S3_1269941 and 118 co-significant markers in high LD (mean r^2 = 0.74) are associated with variation in trial-adjusted flowering time across the 349-line panel.

The broad-sense heritability of FL across the eight environments was 0.91 (95 % of the total phenotypic variance attributable to genotype and GxE), which is consistent with flowering time being a highly heritable trait in rice. The GxE component accounts for 26.4 % of total variance, indicating that ranking of lines for FL is not perfectly stable across environments — a relevant consideration for marker-assisted selection.

The location of the signal on chromosome 3 below 2 Mb is noteworthy in the context of rice flowering biology. The lead positional candidate **LOC_Os03g03070** carries only a generic "transcription factor" annotation in MSU v6.0 and cannot be linked to a known flowering-time pathway without additional evidence. Nearby genes include a Dicer homologue (LOC_Os03g02970), which could participate in small-RNA-mediated regulation of flowering, and two homeobox proteins (LOC_Os03g03164, LOC_Os03g03260), which are known regulators of developmental timing in plants. **These connections are speculative hypotheses, not conclusions.** They require independent experimental validation.

Association does not establish causation. Even if LOC_Os03g03070 is eventually shown to underlie the QTL, GWAS alone cannot distinguish a causal variant in this gene from a variant in LD with it that acts through a neighbouring gene.

Limitations

1. **Imputed genotypes, unknown quality.** All 108,005 SNPs are imputed (0 % observed heterozygosity). Imputation accuracy was not independently assessed. Results may be sensitive to imputation errors, particularly for markers with low call rates. Applies to: all associations.

2. **High MAF filter rate.** 35.2 % of markers were removed at MAF < 5 %, reflecting the diversity and structure of the panel. The analysis cannot detect rare-variant effects.

3. **9 samples excluded.** Nine samples (A1287, B1028, B1058, B1083, B1209, B1247, M1376, M1410, M1421) were excluded for call rate < 80 %. One phenotype-only GHID (A1303) had no matching genotype and was excluded from the GWAS.

4. **Unbalanced trial design.** Only 111 of 349 lines were observed in all 8 trials; 347 of 349 appeared in ≥ 4 trials. The two lines observed in < 4 trials have lower BLUP reliability. BLUP estimates for lines appearing in fewer trials are less reliable (minimum reliability = 0.80). Caveat: unbalanced design handled by mixed model.

5. **2012DS has 2 reps only** (vs. 3 for other trials). This is estimable and handled by the mixed model, but FL BLUPs for lines exclusively observed in 2012DS carry higher uncertainty.

6. **Large locus, 275 annotated genes.** The 1.7 Mb association window contains 275 MSU v6.0 genes. At this resolution, identifying the causal gene is not possible from GWAS alone.

7. **$\lambda_{GC} = 0.892$.** Slightly below 1.0 in a MLM with kinship is expected and is not evidence of systematic deflation. However, a kinship model can over-correct in small panels, potentially reducing power to detect true signals.

8. **No replication in an independent panel.** All results are from a single panel. Validation requires replication in at least one additional mapping population or multi-parent design.

9. **No effects reported.** TASSEL MLM with P3D does not export allele effect sizes in the canonical output for this dataset. Allele effect direction and magnitude are unknown.

Recommended Follow-up

1. **Imputation quality assessment.** Obtain original pre-imputation genotypes and compute per-marker imputation accuracy (R^2 or concordance) to validate retained markers.

2. **Fine-mapping.** Resequence or genotype additional markers in Chr3: 0–2.5 Mb at higher density, then apply statistical fine-mapping (e.g., FINEMAP, SuSiE) to narrow the credible set.

3. **Candidate gene prioritisation.** Check expression of LOC_Os03g03070 and neighbouring genes in flower bud and leaf tissue across developmental stages in public rice transcriptome datasets.

4. **Allelic series.** Identify lines carrying the two major haplotypes at the lead locus and phenotype them for FL under controlled conditions to estimate the allele effect.
5. **Cross-population validation.** Test the lead markers in at least one independent rice association panel or near-isogenic line population.
6. **Functional validation.** CRISPR or overexpression of LOC_Os03g03070 in a model background (Nipponbare) would directly test whether the gene controls flowering time.
7. **GxE sensitivity analysis.** Examine whether BLUPs from individual trials or environment sub-sets produce consistent associations, to evaluate GxE confounding.

Output Files

All paths are relative to the task working directory.

Data and genotype QC

File	Description
gwas_config.json	GWAS dispatch contract (gwas-dispatch-v1)
preflight.json	TASSEL skill preflight receipt
qc_exclusion_ledger.json	Complete sample & marker exclusion ledger
sample_qc.tsv	Per-sample call rate and QC decisions
MET_matched.hmp.txt	QC-cleaned matched HapMap (358 samples, 108,005 markers)

Phenotype model

File	Description
plot_data_qc.tsv	QC'd plot records (5,803 rows, 349 lines, 8 trials)
met_model.rds	Fitted lme4 model object
fl_variance_components.tsv	Variance component table
fl_blup_full.tsv	FL BLUPs with PEV and reliability (349 lines)
flowering_time_adjusted.tsv	GWAS input phenotype (GHID, FL_adjusted)

Reference annotation

File	Description
msu_v60_reference_manifest.json	MSU v6.0 download provenance and checksums
all.gff3	Decompressed MSU v6.0 GFF3 (514,872 features)

File	Description
all_TU_brief_info.tsv	MSU v6.0 TU brief annotation

GWAS results

File	Description
gwas_results_canonical.tsv	Full association table (69,833 markers)
gwas_validation.json	Bonferroni threshold, λ_GC, variant counts
gwas_significant_markers.tsv	119 Bonferroni-significant markers
pca_coordinates.tsv	PCA coordinates (349 samples, 10 PCs)

LD and annotation

File	Description
chr3_r2_matrix.npy	Pairwise r^2 matrix for Chr3 window (717×717)
chr3_region_markers.tsv	Markers in Chr3 window
ld_region_summary.json	LD signal assessment summary
gwas_annotated_significant.tsv	Annotated significant markers (overlapping & nearest genes)
chr3_locus_candidate_genes.tsv	All 275 MSU v6.0 genes in significant region
annotation_summary.json	Annotation provenance and lead SNP details

Figures

Figure	Description
Fig 1	MET model residuals vs fitted
Fig 2	Q-Q plot of model residuals
Fig 3	Trial fixed effects (mean FL per environment)
Fig 4	Distribution of FL BLUPs
Fig 5	BLUPs vs raw means (r = 0.9975)
Fig 6	BLUP reliability per line
Fig 7	**Manhattan plot** — genome-wide FL GWAS
Fig 8	**Q-Q plot** — GWAS p-values (λ = 0.892)
Fig 9	PCA coloured by FL_adjusted
Fig 10	Chr3 association zoom

Figure	Description
Fig 11	LD heatmap for 119 significant markers
Fig 12	LD decay in Chr3 window
Fig 13	Annotated Chr3 locus with MSU v6.0 gene models

Tool Versions and Provenance

Tool	Version/Notes
TASSEL	5, run via `run_pipeline.pl`, MLM+P3D, Centered IBS kinship
lme4 (R)	REML, bobyqa optimizer, convergence confirmed
scikit-learn	PCA for visualization (independent of TASSEL PCA)
MSU v6.0 GFF3	SHA-256: 9a6db358f4c8e8a9f55251558dc15cbc01a03b55996b575260b5170cd2c9d03f
MSU v6.0 TU info	SHA-256: e10697d0989d976494985f8b354c5b05a5488ed4649c658af706daa9df4c68d9
Plant-GWAS-TASSEL skill	schema: plant-gwas-tassel-v1, state: remediated

Report generated by the plant-breeding GWAS workflow. All positional candidate gene assignments are based solely on MSU v6.0 genomic coordinates and annotation text; no training-knowledge gene–function associations have been introduced.

Figures

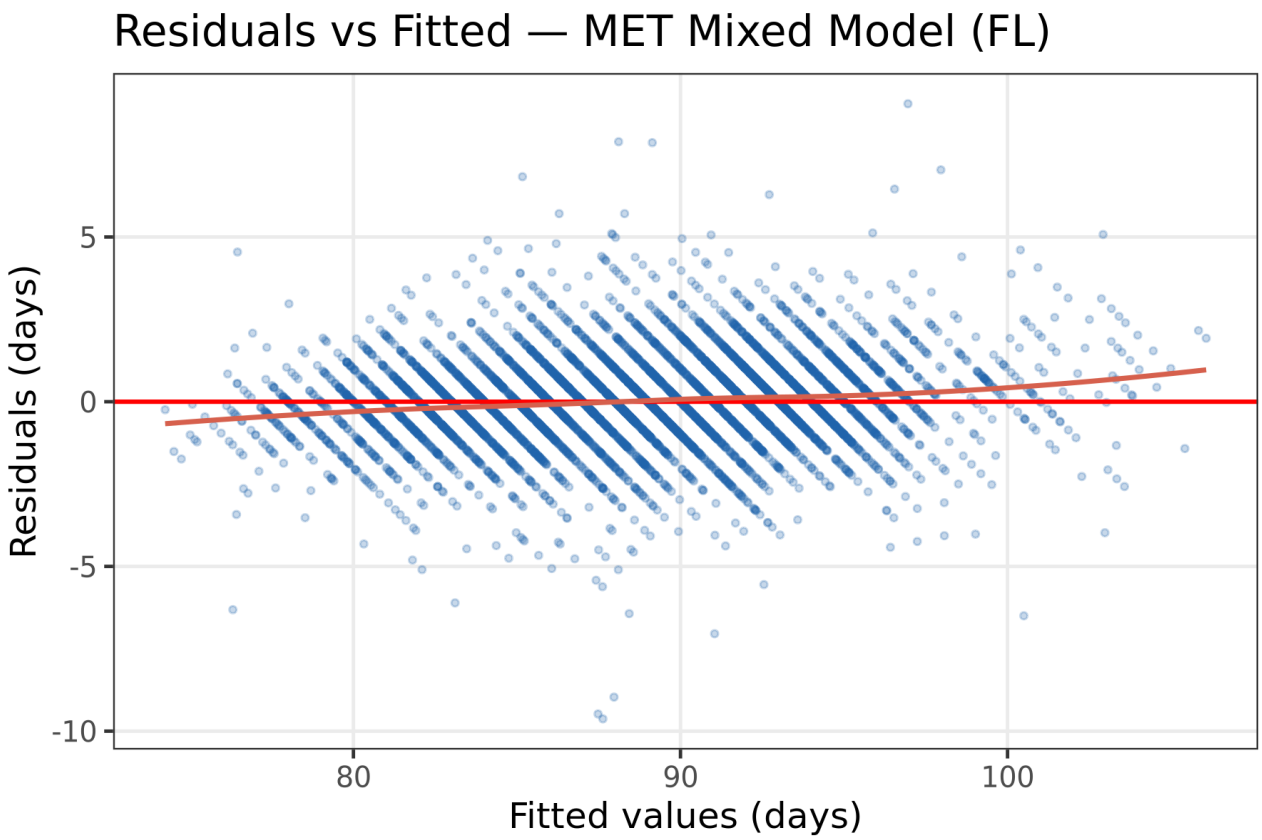

Figure 1. Residuals vs fitted values for the MET mixed model (FL ~ trial + (1|GHID) + (1|GHID:trial) + (1|trial:REP)); 5799 plot observations; loess smoother near zero indicates no strong non-linearity.

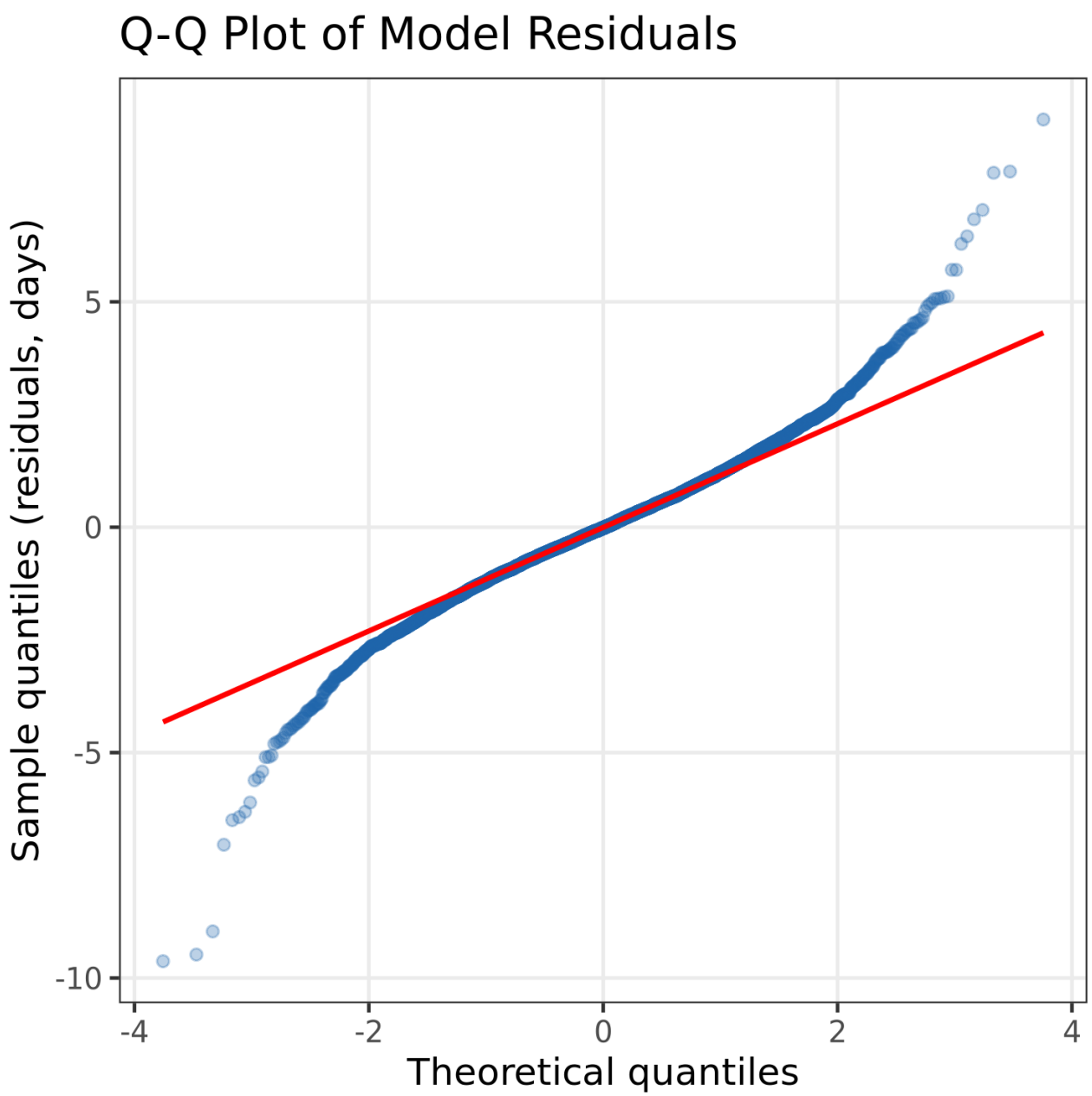

Figure 2. Q-Q plot of residuals from the multi-environment FL mixed model; near-linear trend supports approximate normality of plot-level residuals.

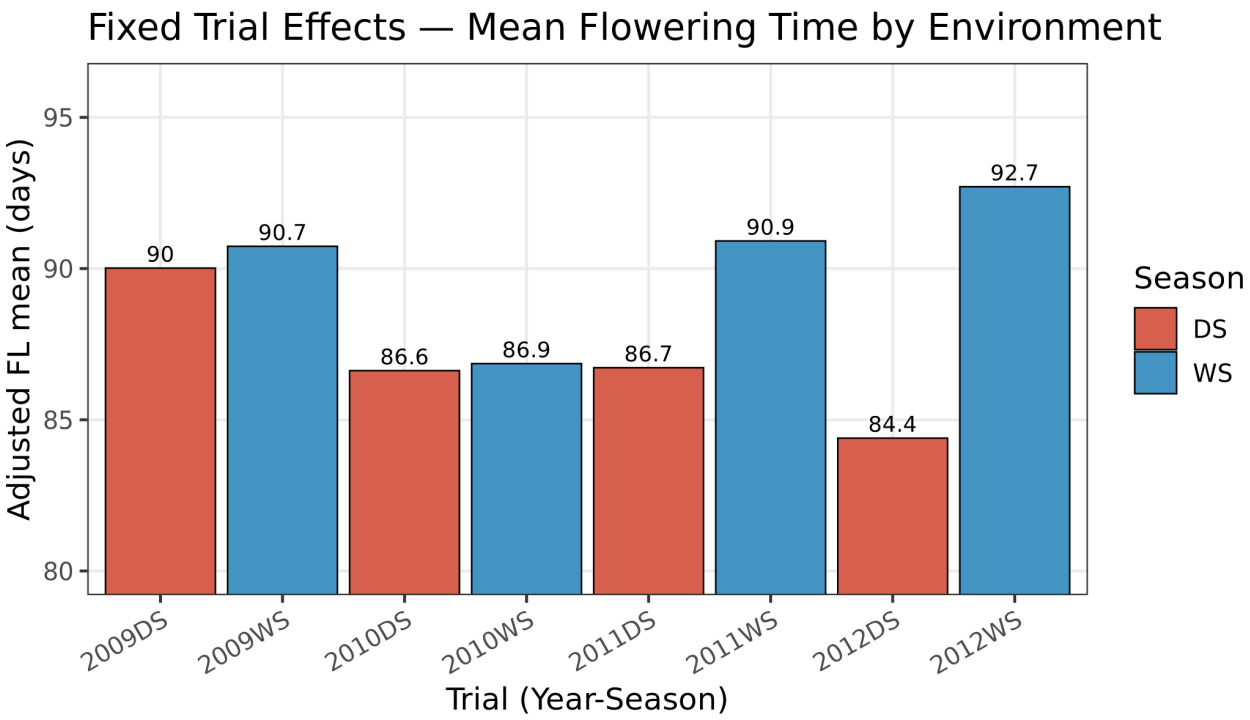

Figure 3. Fixed trial effects (estimated mean FL) for all 8 year-season environments; WS trials (blue) show higher FL than DS trials (red), ranging from 84.4 days (2012DS) to 92.7 days (2012WS).

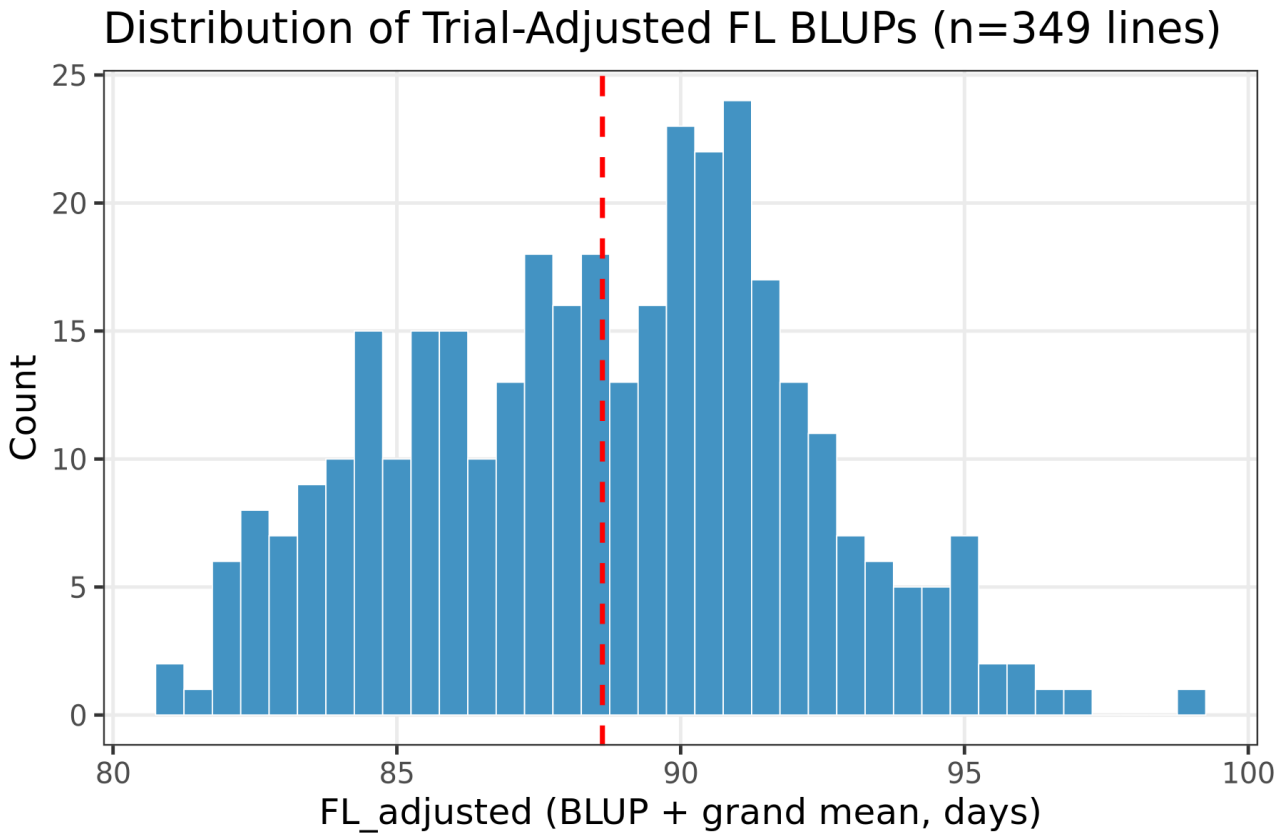

Figure 4. Distribution of trial-adjusted FL BLUPs for 349 breeding lines; grand mean=88.62 days, SD=3.47 days, approximately normal.

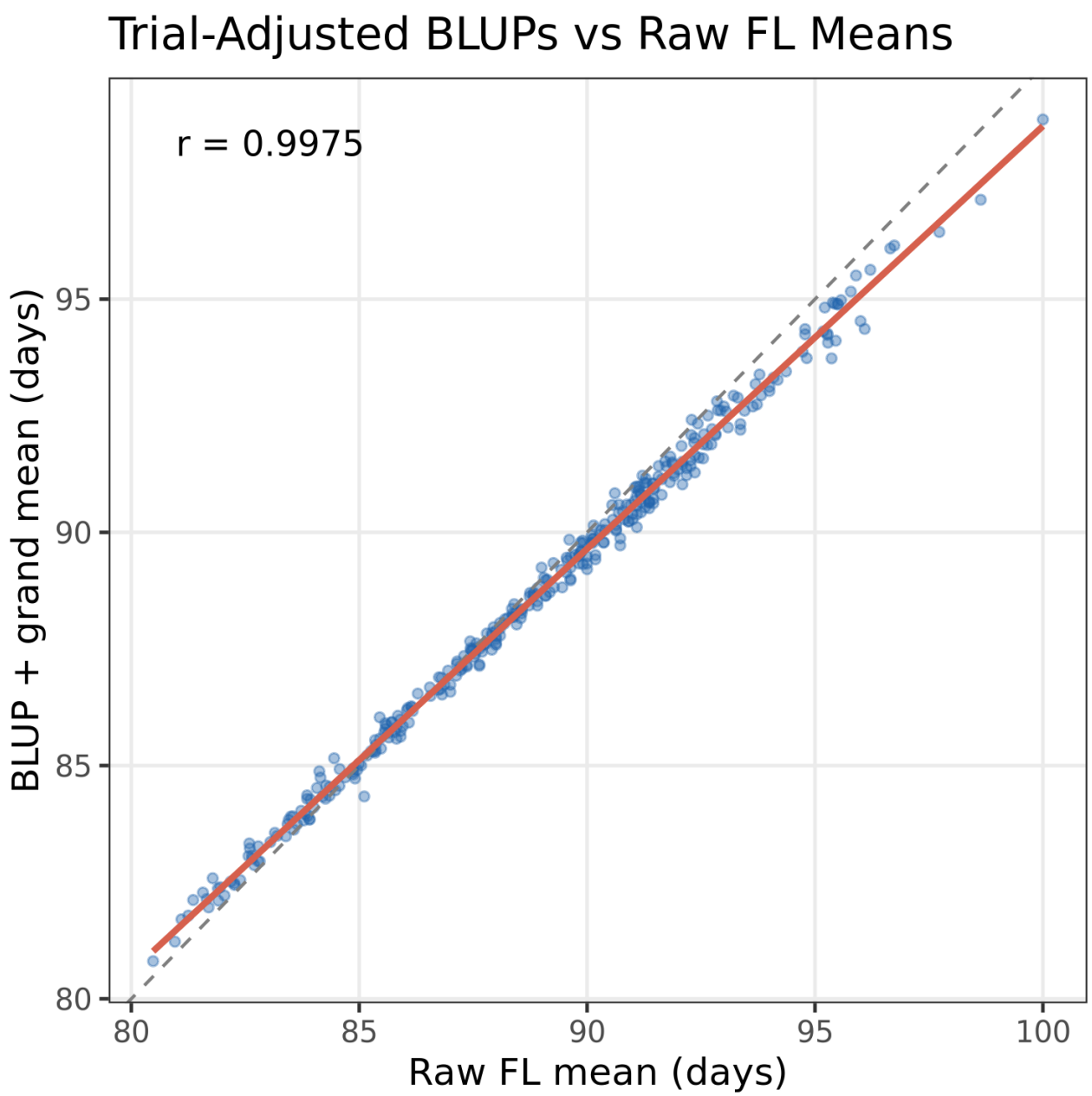

Figure 5. Trial-adjusted FL BLUPs vs per-line raw FL means across all observations (Pearson r=0.9975); BLUPs show shrinkage relative to raw means, appropriate for unbalanced designs.

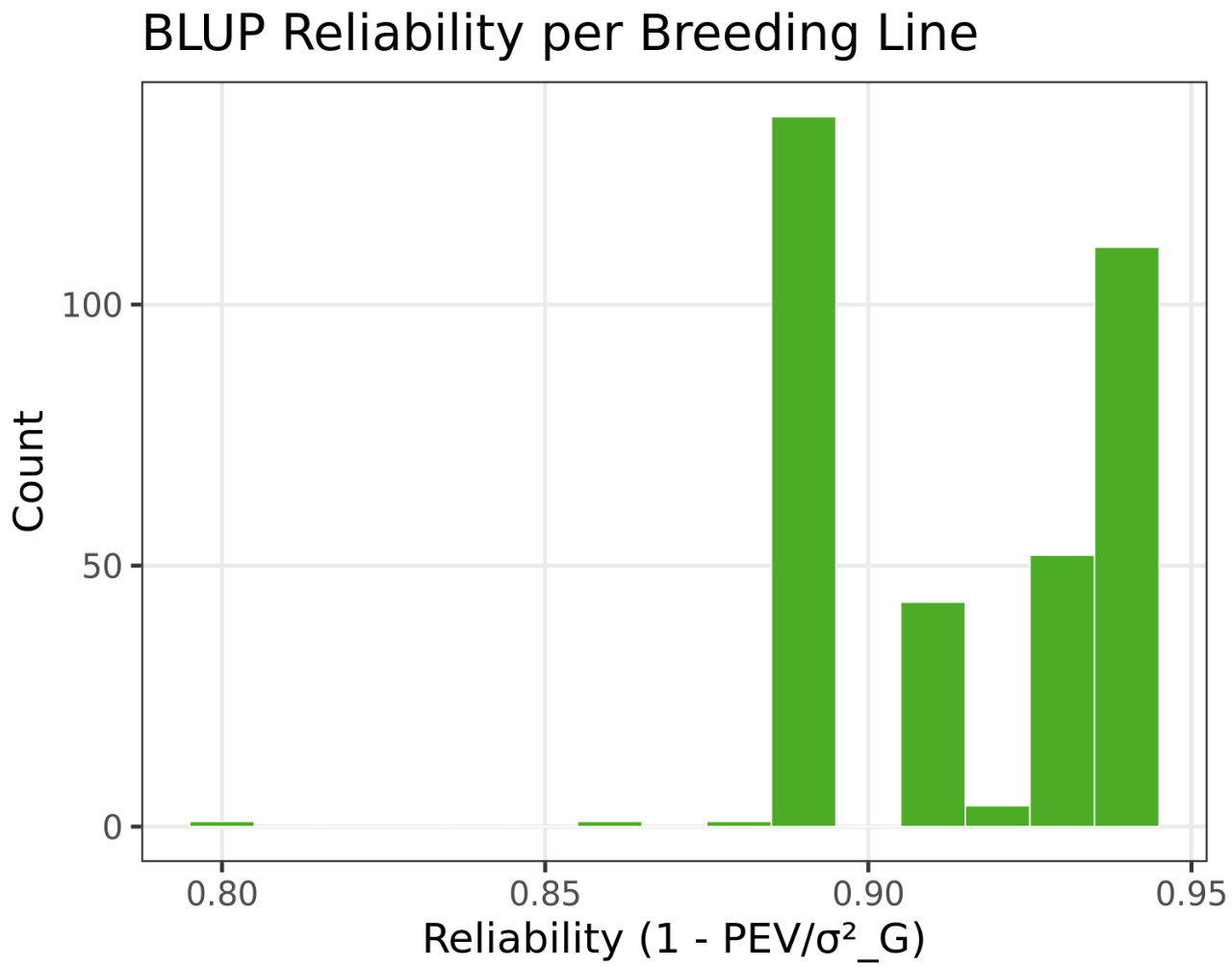

Figure 6. BLUP reliability per breeding line; mean reliability = 0.91 indicating high information content per line estimate.

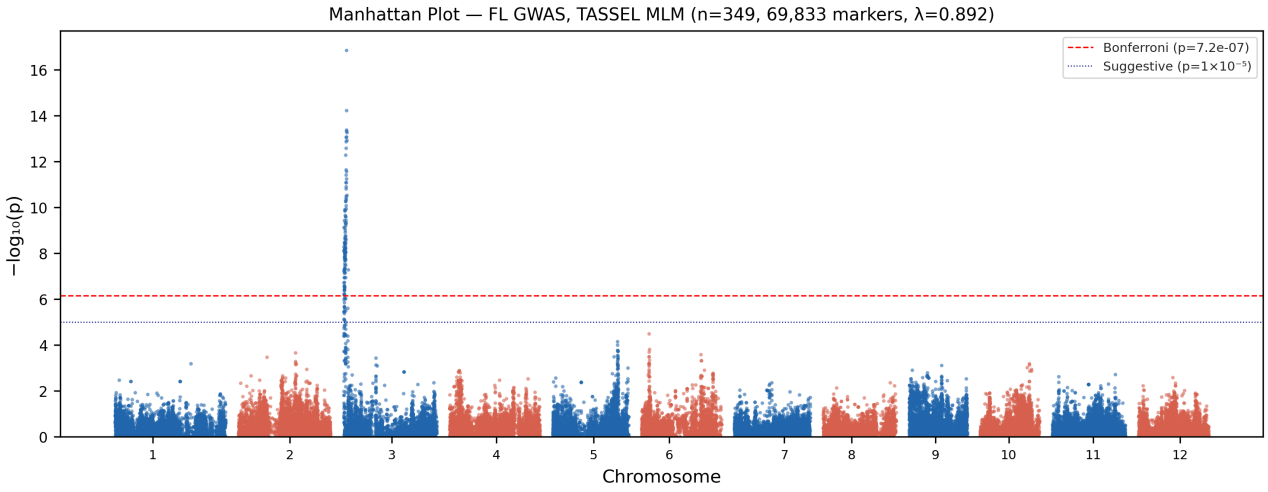

Figure 7. Manhattan plot of flowering-time GWAS (TASSEL MLM, n=349 lines, 69,833 testable markers); Bonferroni threshold p=7.16e-07 (red dashed); 119 genome-wide significant markers, all on chromosome 3; λ_GC=0.8919.

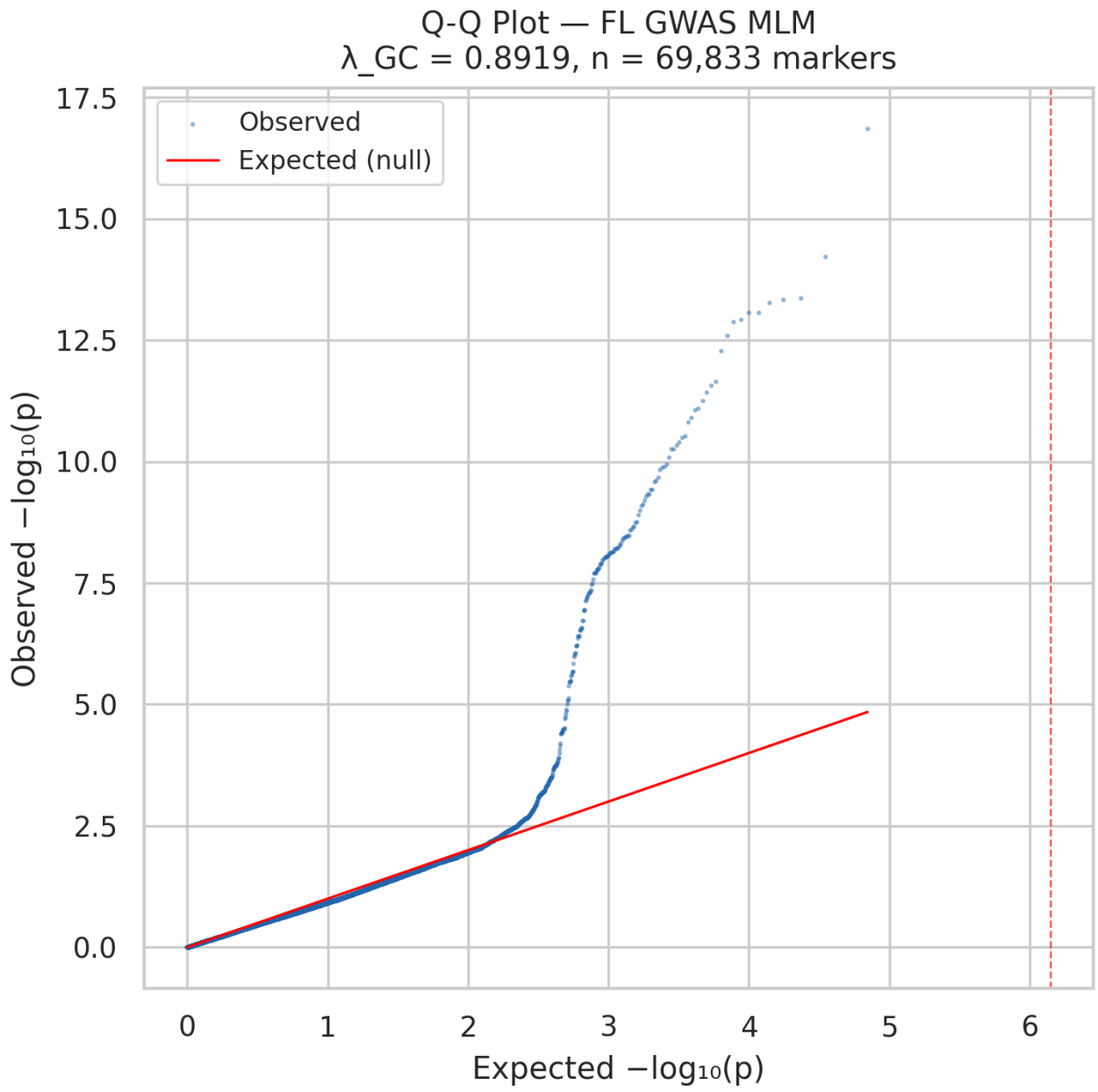

Figure 8. Q-Q plot of GWAS p-values for flowering time (MLM); λ_GC=0.8919 indicates slight over-correction (conservative model); strong deviation from null for the chr3 locus.

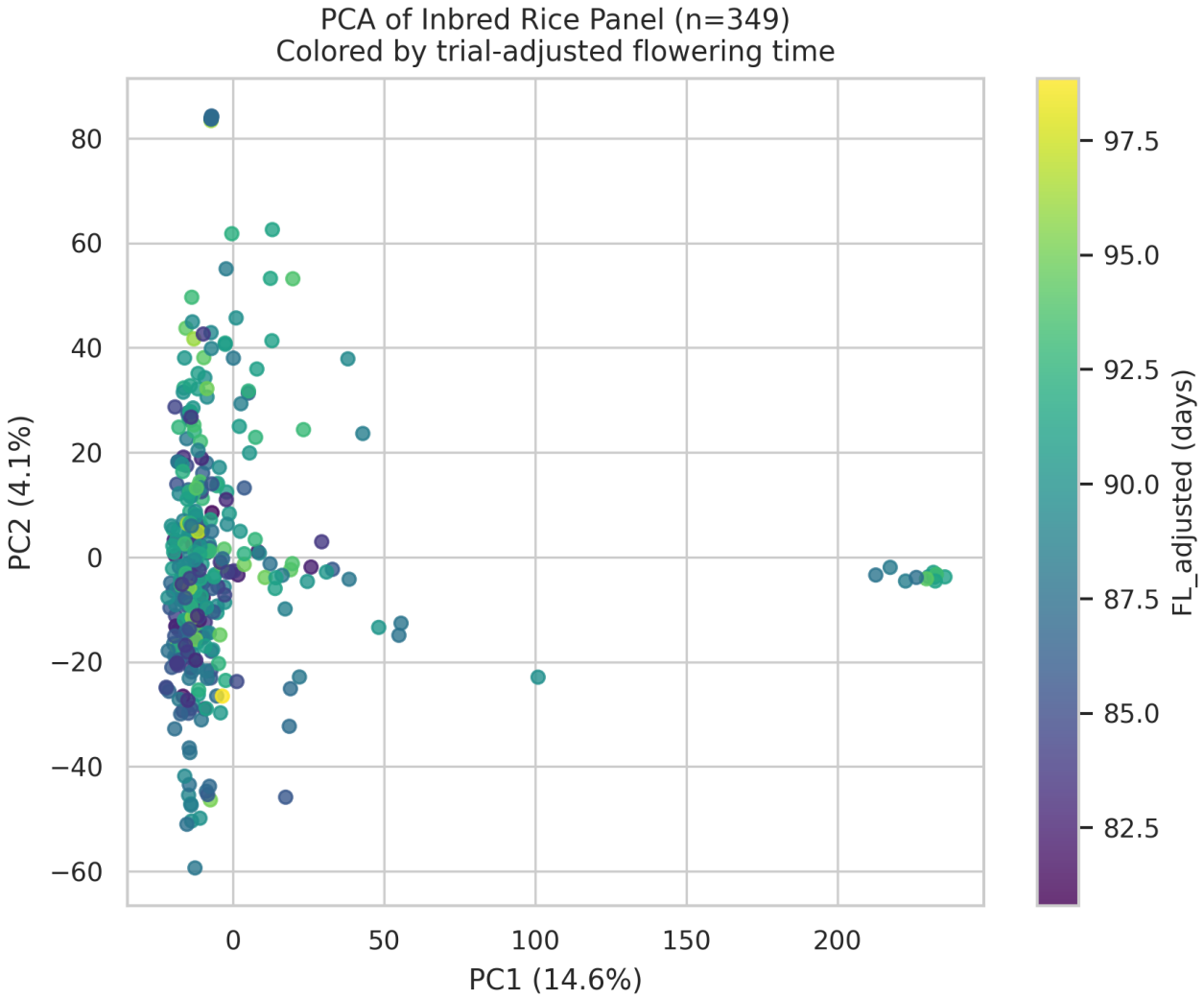

Figure 9. PCA of 349 inbred rice lines coloured by trial-adjusted flowering time (BLUP); PC1 explains 14.6% of genotypic variance, indicating substantial population structure; lines with later flowering tend to cluster separately on PC1.

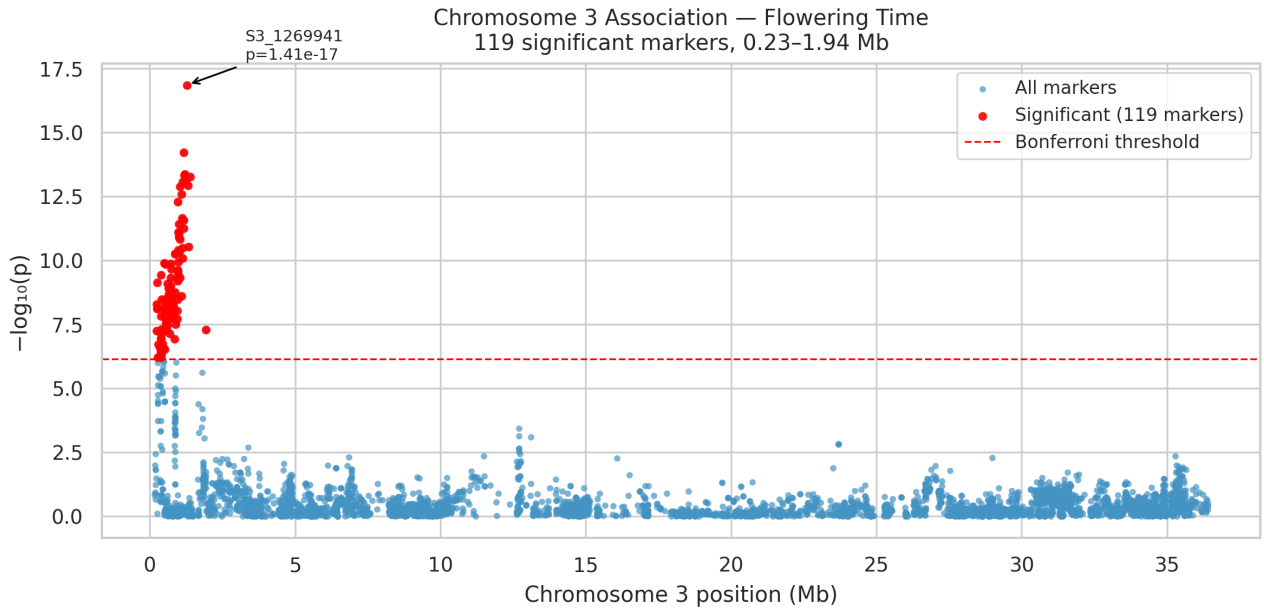

Figure 10. Chromosome 3 association zoom showing 119 genome-wide significant markers (red) spanning 0.23–1.94 Mb; lead SNP S3_1269941 at $p=1.41\times10^{-17}$; continuous signal with no gaps >1 Mb suggests a single locus.

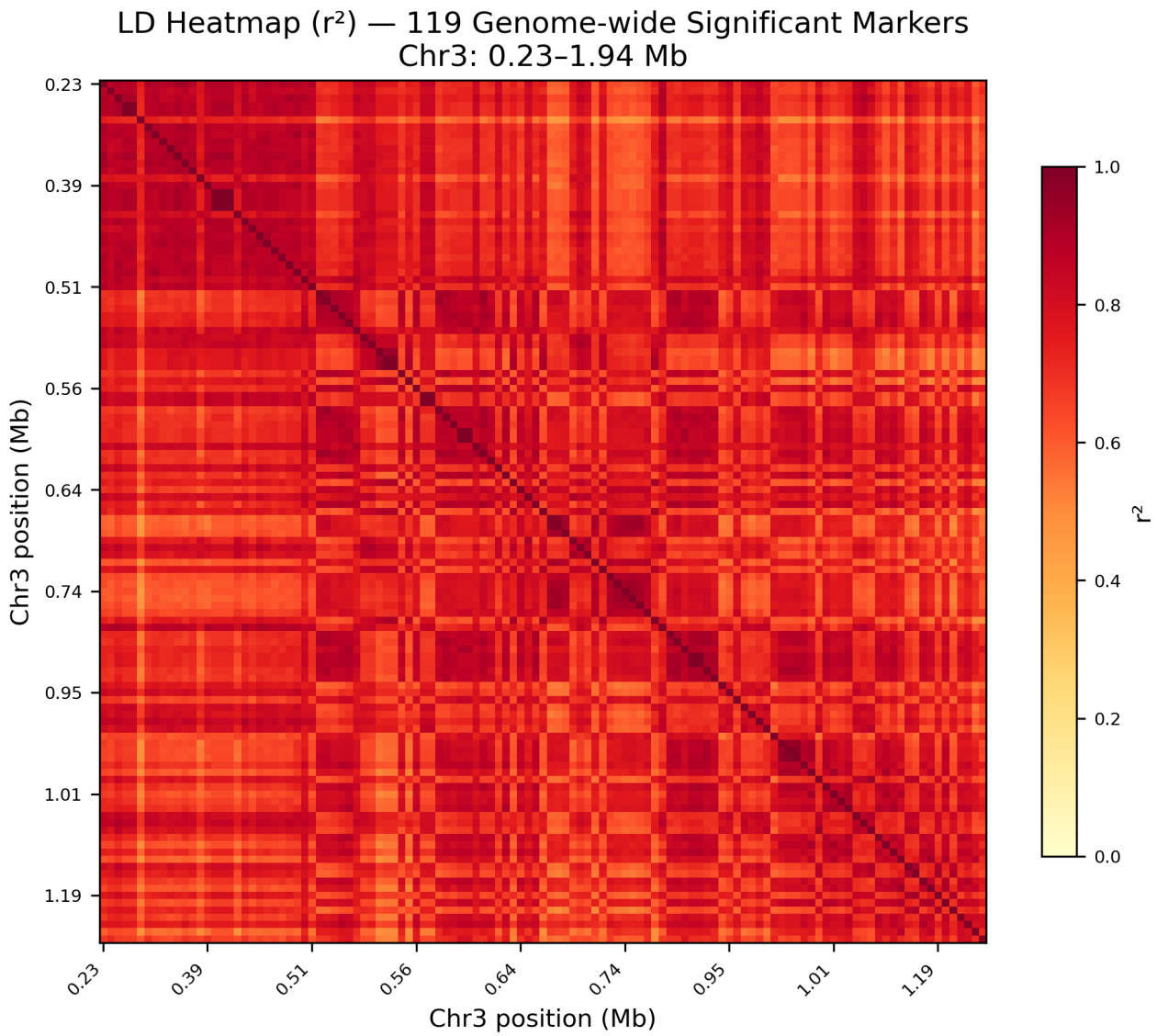

Figure 11. LD heatmap (r²) among the 119 genome-wide significant markers on Chr3 (0.23–1.94 Mb); high mean r²=0.740 among significant marker pairs indicates a single strongly-associated LD block rather than multiple independent signals.

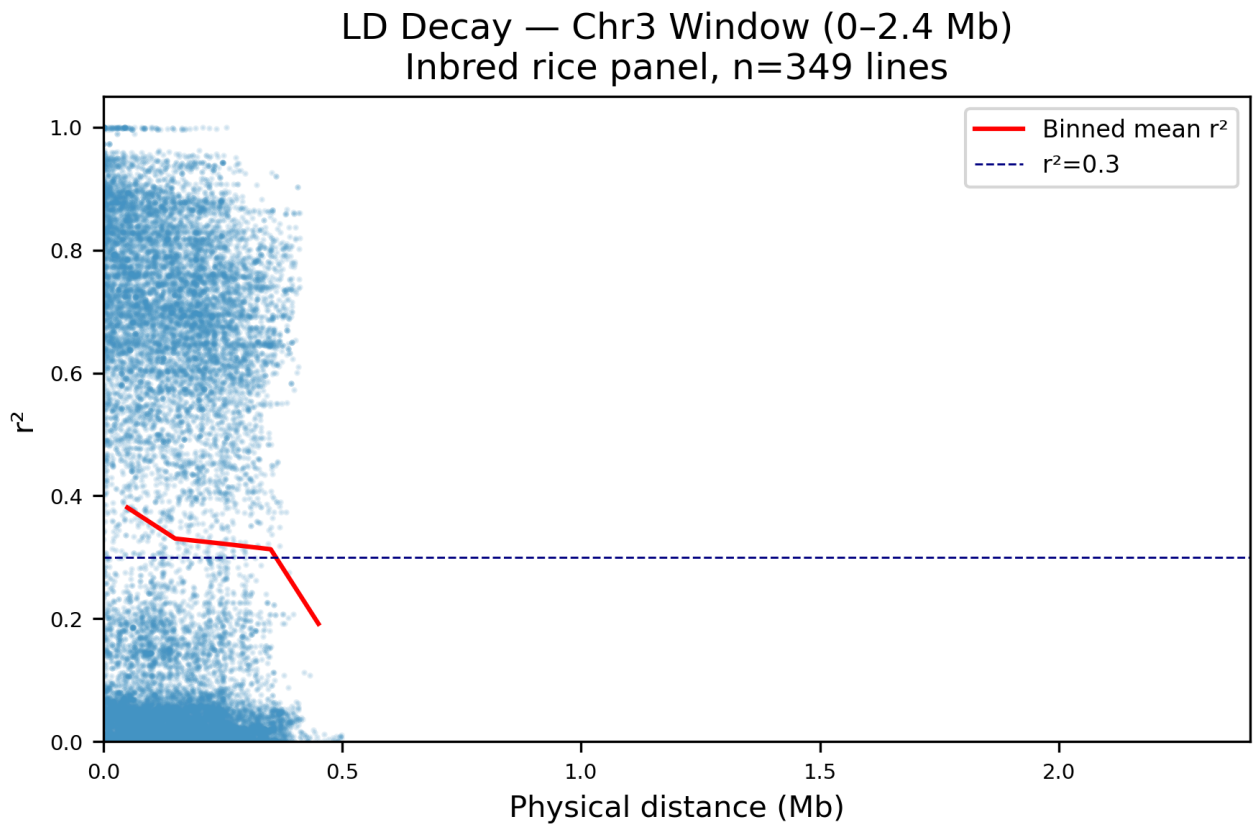

Figure 12. LD decay curve in the Chr3 window (0–2.4 Mb) for the inbred rice panel (n=349 lines); r² remains elevated (>0.3) across the significant region, consistent with a single large haplotype block.

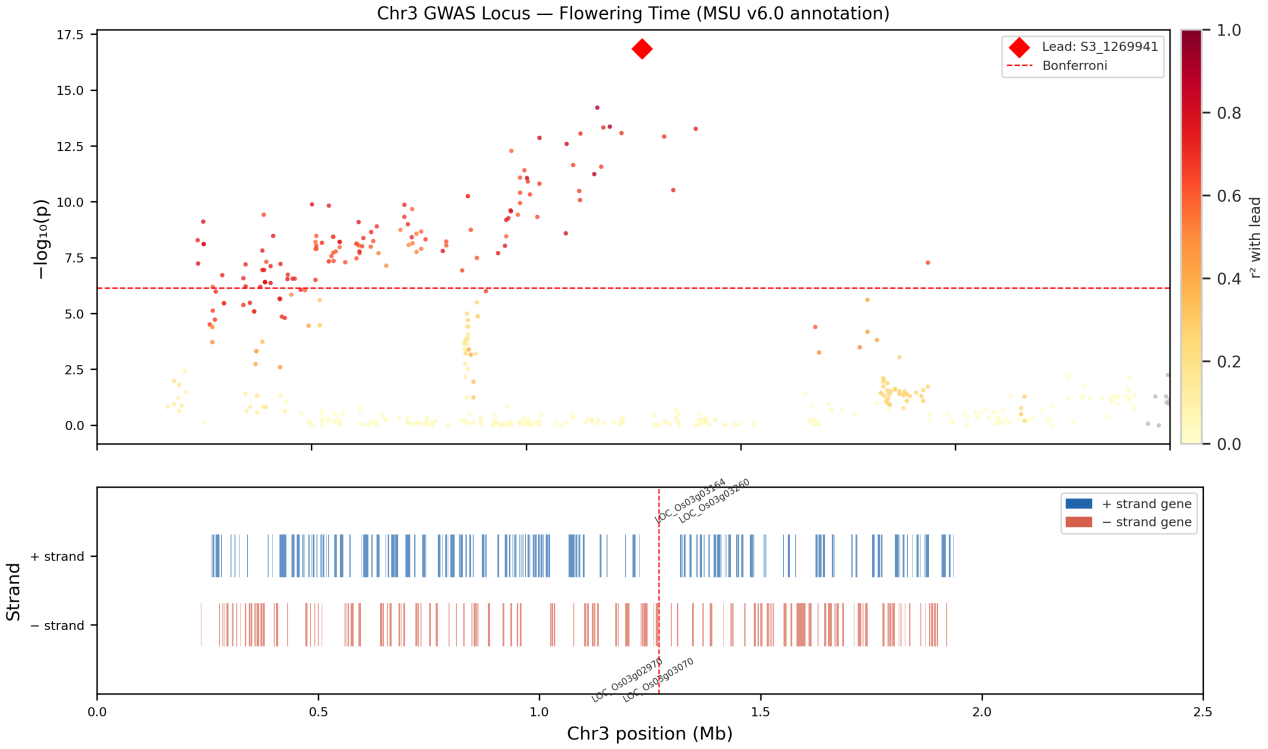

Figure 13. Annotated Chr3 GWAS locus (0–2.5 Mb): top panel shows −log₁₀(p) colored by r² with lead SNP S3_1269941 (p=1.41×10⁻¹⁷; red diamond); bottom panel shows MSU v6.0 gene models (blue=+ strand, red=− strand); lead SNP falls within LOC_Os03g03070 (transcription factor, − strand, 1,268,854–1,270,781 bp).

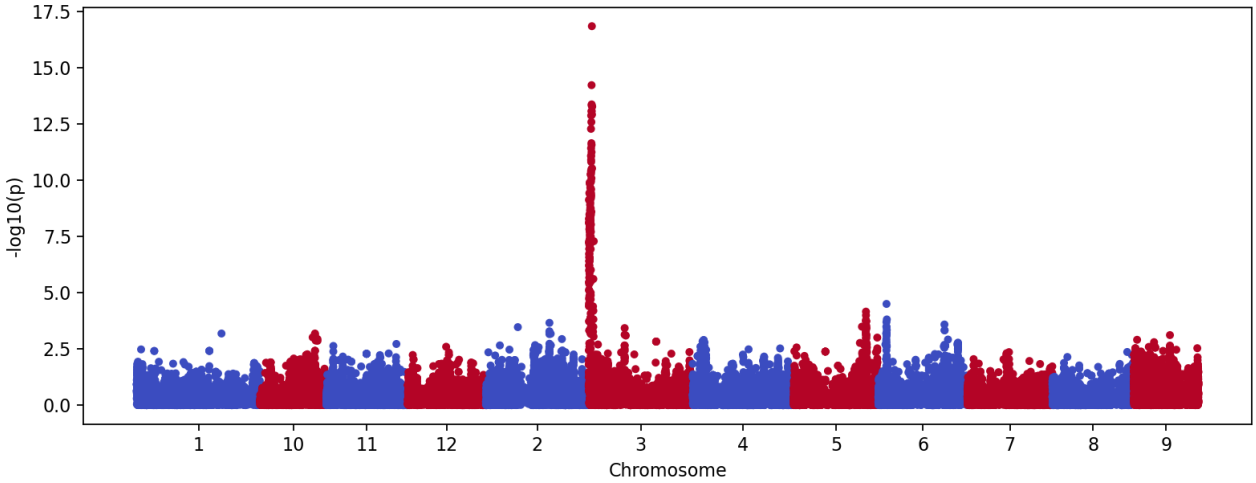

Figure 14. manhattan_plot.png

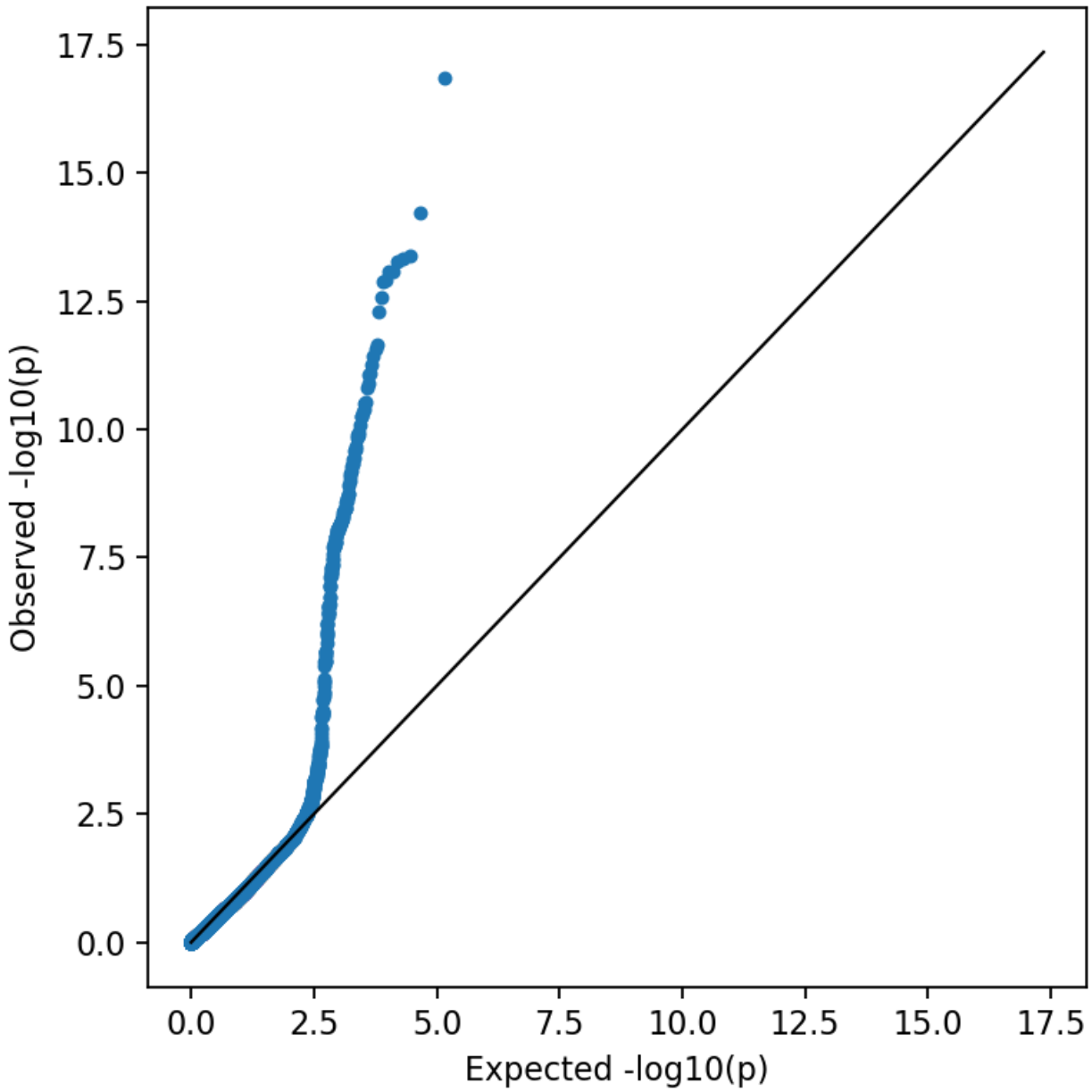

Figure 15. qq_plot.png

SESSION RECORD

Additional discussion

RESEARCH REQUEST

LOC_Os03g03070 is OsMADS50 per https://funricegenes.github.io/LOC_Os03g03070/. Create Figure 7 again and annotate some of the genes on the figure.