Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From phenotypic diversity to genotype selection: Integrating trait relationships and machine learning for high-yield Quinoa

Ngoc Dinh Nguyen
Vietnam National University of Agriculture
Van Loc Nguyen
Vietnam National University of Agriculture
Thi Hong Nhung Phan
Vietnam National University of Agriculture
Thai Hoang Dinh
Vietnam National University of Agriculture
Thi Thu Ha Phung
Vietnam National University of Agriculture
Trung Duc Nguyen
Vietnam National University of Agriculture
Viet-Long Nguyen
Vietnam National University of Agriculture
Duc Ha Chu
Vietnam National University, Hanoi
Van Nhan Le
Vietnam Academy of Science and Technology
Ramiro Curti
National University of Salta
Daniel Bertero
Universidad de Buenos Aires
Ton Viet Ta
tavietton@agr.kyushu-u.ac.jp

Kyushu University

Research Article

Keywords: Quinoa, diversity, genotype selection, machine learning, Logistic, LDA, SVM RBF, RF

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11041951/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Quinoa (*Chenopodium quinoa* Willd.) possesses substantial phenotypic diversity, but the identification of high-yielding genotypes from large germplasm collections remains challenging because yield is determined by multiple interacting traits. This study evaluated the agro-morphological diversity of 142 quinoa genotypes and investigated the phenotypic profile associated with high-yield performance using multivariate and machine-learning approaches. Hierarchical cluster analysis classified the genotypes into five distinct groups, confirming substantial phenotypic diversity within the panel. Growth traits showed strong correlations across developmental stages; however, most vegetative traits were weakly associated with individual grain yield (IGY). Higher IGY was primarily associated with greater leaf area (LA) and thousand-seed weight (P1000), whereas excessive late plant height (PH) and number of branches (NoB) showed a negative relationship with IGY. Phenotyping across developmental stages improved discrimination between high-yield and other genotypes, and the complete predictor set including P1000 provided the strongest performance. Among the evaluated classifiers, including Logistic Regression, Linear Discriminant Analysis (LDA), Support Vector Machine with a radial basis function (RBF) kernel, and Random Forest, LDA achieved the highest PR-AUC and ROC-AUC, whereas Logistic Regression attained the highest balanced accuracy and identified a larger proportion of the observed high-yield genotypes. The high-yield phenotype was characterized by greater canopy development and chlorophyll status near 58 DAS, favorable seed mass, and moderate plant stature and branching. Consensus ranking further prioritized both observed high-yield genotypes and phenotypically similar candidates for subsequent evaluation. These findings demonstrate that integrating large-scale phenotypic characterization with multivariate and machine-learning analyses provides a useful framework for germplasm screening and candidate prioritization in quinoa breeding, although independent and multi-environment validation is required before predictive application.

1 Introduction

Quinoa (*Chenopodium quinoa* Willd.) is increasingly recognized as a promising crop for agricultural diversification (Yadav et al., 2023) and food security (Jacobsen, 2003) because of its nutritional value and broad adaptation to diverse agroecological conditions (Angeli et al., 2020; del Castillo et al., 2007). A major advantage of quinoa germplasm is its substantial phenotypic variation, which provides an important resource for identifying desirable trait combinations and developing improved cultivars (Patiranage et al., 2022). However, the existence of extensive variation within a germplasm collection does not necessarily translate into efficient breeding progress (Nazeer and Akram, 2025). The major challenge is to characterize this multidimensional variation and identify genotypes that combine desirable phenotypic characteristics with superior grain-yield potential (Mousa et al., 2025; Nazeer and Akram, 2025). Thus, effective utilization of quinoa germplasm requires approaches that move beyond simple description of phenotypic variation toward the systematic identification and prioritization of promising genotypes.

Grain yield is a complex quantitative trait resulting from the combined expression of multiple morphological and physiological characteristics throughout plant development (Curti et al., 2014; Dumschott et al., 2022). Consequently, direct selection based on final grain yield is generally time- and resource-intensive, particularly when large numbers of genotypes need to be evaluated. Early-stage phenotyping offers the possibility of accelerating selection by identifying promising materials before crop maturity (Rakasi et al., 2026). However, the extent to which early phenotypic characteristics reflect subsequent grain-yield performance remains uncertain. Traits expressed at an early developmental stage may capture only part of the biological variation contributing to final yield (Maldonado-Taipe et al., 2022; Patel et al., 2026), whereas measurements collected across several developmental stages may provide a more comprehensive representation of genotype performance (Saravani et al., 2026). Determining the relative discriminatory value of early versus comprehensive phenotyping is therefore important for improving the efficiency of quinoa germplasm screening.

The difficulty of identifying high-yield genotypes is further increased by the multidimensional nature of quinoa phenotypes. Genotypes may differ simultaneously in plant height, leaf number, branching, stem diameter, leaf area, chlorophyll status, and grain characteristics, with the relationships among these traits potentially changing across developmental stages (Maestro-Gaitán et al., 2022). Analyses based on individual traits may therefore fail to capture combinations of characteristics associated with superior yield performance. Multivariate approaches such as principal component analysis (PCA) provide a useful framework for summarizing correlated phenotypic variables and representing genotypes within a multidimensional phenotype space (Aschard et al., 2014). PCA can reveal major sources of phenotypic variation and facilitate the identification of contrasting genotype profiles (Hazra et al., 2026). Nevertheless, PCA is primarily descriptive, and substantial overlap between high- and low-performing genotypes may remain even when the major dimensions of phenotypic variation have been identified (Küçüközdemir et al., 2026). Thus, phenotypic diversity characterization alone may not be sufficient for objective genotype selection.

Machine-learning approaches (Küçüközdemir et al.) offer a complementary strategy by enabling the simultaneous analysis of multiple phenotypic variables to discriminate genotypes according to predefined performance classes (Aggarwal et al., 2026). Rather than relying on a single trait or a small number of conventional selection criteria, classification algorithms can integrate information distributed across multiple dimensions of the phenotype (Coppens et al., 2017). This is particularly relevant when high-yield genotypes cannot be clearly separated using simple visual inspection or the first few principal components. However, the effectiveness of ML-based classification depends strongly on the information contained in the predictor variables (Elbasi et al., 2023). In particular, an important practical question is whether a relatively small set of early phenotypic measurements can provide sufficient information for high-yield discrimination, or whether reliable classification requires phenotyping across multiple developmental stages and the inclusion of grain-related characteristics.

Importantly, high-yield selection should not be considered independently of the phenotypic diversity present within a germplasm collection. Selection based solely on observed or predicted yield may favor a relatively narrow group of genotypes with similar phenotypic profiles, potentially reducing the diversity

represented among selected materials (Bista et al., 2026). Conversely, genotypes that do not exceed a predefined yield threshold may nevertheless possess combinations of traits resembling those of high-yielding genotypes and could therefore represent potentially valuable materials for further evaluation (Hassani et al., 2024). A useful selection framework should consequently distinguish between observed high-yield genotypes and genotypes exhibiting high-yield-like phenotype profiles, while retaining information on their position within the broader phenotypic diversity of the germplasm. Integrating diversity characterization with high-yield classification may therefore provide a more informative basis for prioritizing breeding materials than ranking genotypes solely according to final yield.

Despite the increasing availability of multidimensional phenotypic data and the growing application of ML approaches in crop improvement (Aggarwal et al., 2026; Coppens et al., 2017; Sheikh et al., 2024; Tsega and Mullualem, 2026), an important gap remains in the integration of phenotypic diversity, developmental phenotyping, and high-yield genotype discrimination in quinoa. Existing approaches do not fully clarify how much information can be obtained from early phenotyping compared with comprehensive phenotyping, which phenotypic characteristics are most strongly associated with high-yield membership, or whether combining multivariate phenotype information with classification models can effectively prioritize promising genotypes within a diverse germplasm collection. In addition, a model that identifies high-yield-like genotypes should not simply be interpreted as a predictor of final yield; rather, its value lies in providing a systematic means of identifying and prioritizing candidate materials for subsequent validation. Addressing these gaps is therefore important for developing more efficient phenotype-based strategies for quinoa germplasm evaluation and breeding. Therefore, this study aimed to develop an integrated framework combining phenotypic diversity characterization, PCA, and machine-learning classification to discriminate and prioritize high-yield quinoa genotypes. Specifically, we hypothesized that early-stage phenotypic measurements alone would provide limited discrimination of high-yield genotypes, whereas the integration of phenotypic information collected across developmental stages, particularly when supplemented with thousand-grain weight, would improve the identification of high-yield-associated phenotype profiles.

2 Materials and methods

Plant material and experimental design

The panel comprised 142 quinoa genotypes, denoted G1-G142 (Table 1), grown at the experimental field of Vietnam National University of Agriculture, Gia Lam, Hanoi (21°00' N, 105°56' E). Genotypes were arranged in a randomized complete block design with two replications; each genotype occupied one plot per block, giving two plots per genotype (Fig. 1). Each plot comprised two 3-m rows spaced 0.70 m apart, with hills spaced 0.15 m within rows and 7–10 seeds sown per hill. At the five-leaf stage, each hill was thinned to a single plant, and five representative plants were retained per plot for phenotypic evaluation, yielding 10 measured plants per genotype (1,420 plants in total). Fertilizer was applied at 120 kg N, 90 kg P_2O_5, and 90 kg K_2O per hectare, with nitrogen split 30% basal, 40% at vegetative growth, and 30% at flowering; phosphorus was applied entirely as a basal dressing and potassium split evenly between basal and flowering applications. Weed and pest management followed standard local practice.

Table 1
Name and passport data of quinoa genotype

No.	Genotype	Selected from	Orgin	No.	Genotype	Selected from	Orgin
1	G1	Cahuil	Chile	72	G35	Atlas	Netherland
2	G10	G2	Argentina	73	G36	2-want	Bolivia
3	G100	G23	Argentina	74	G37	2-want	Bolivia
4	G101	Real1xRuy937	US	75	G38	2-want	Bolivia
5	G102	Q4-bis	Argentina	76	G39	Chadmo	Chile
6	G103	Unknown	Argentina	77	G4	G18	Argentina
7	G104	G9	Argentina	78	G40	G18	Argentina
8	G105	G18	Argentina	79	G41	Unknown	Argentina
9	G106	SuramixRuy937	US	80	G42	Riobamba	Netherland
10	G107	Lucaboldo	Chile	81	G43	EDK-4	Netherland
11	G108	Q3	Argentina	82	G44	Real1xRuy937	US
12	G109	2-want	Bolivia	83	G45	Bear Cajon	Chile
13	G11	Real1xRuy937	US	84	G46	Riobamba	Netherland
14	G110	2-want	Bolivia	85	G47	Real1xRuy937	US
15	G111	Real1xRuy937	US	86	G48	Unknown	Argentina
16	G112	G9	Argentina	87	G49	G18	Argentina
17	G113	G18	Argentina	88	G5	G19	Argentina
18	G114	Ru5	United Kingdom	89	G50	Unknown	Argentina
19	G115	FaroxChadmo	Chile	90	G51	Puno	Denmark
20	G116	SuramixRuy937	US	91	G52	Chadmo	Chile
21	G117	G18	Argentina	92	G53	2-want	Bolivia
22	G118	Q4	Argentina	93	G54	Atlas	Netherland
23	G119	Real1xRuy937	US	94	G55	Q4	Argentina
24	G12	G18	Argentina	95	G56	2-want	Bolivia
25	G120	G3	Argentina	96	G57	Cahuil	Chile
26	G121	Riobamba	Netherland	97	G58	Titicaca	Denmark
27	G122	G49	Argentina	98	G59	G1	Argentina

No.	Genotype	Selected from	Orgin	No.	Genotype	Selected from	Orgin
28	G123	SuramixRuy937	US	99	G6	G1	Argentina
29	G124	SuramixRuy937	US	100	G60	G3	Argentina
30	G125	Chadmo	Chile	101	G61	Lucaboldo	Chile
31	G126	G26	Argentina	102	G62	Real1xRuy937	US
32	G127	G49	Argentina	103	G63	Lucaboldo	Chile
33	G128	G20	Argentina	104	G64	Titicaca	Denmark
34	G129	SuramixRuy937	US	105	G65	Real1xRuy937	US
35	G13	Riobamba	Netherland	106	G66	G21	Argentina
36	G130	Real1xRuy937	US	107	G67	G46	Argentina
37	G131	G18	Argentina	108	G68	G1	Argentina
38	G132	Ru2	United Kingdom	109	G69	42-Test	Chile x Bolivia
39	G133	G1	Argentina	110	G7	42-Test	Chile x Bolivia
40	G134	Real1xRuy937	US	111	G70	Cahuil	Chile
41	G135	G18	Argentina	112	G71	Q2	Argentina
42	G136	G9	Argentina	113	G72	Q4	Argentina
43	G137	Real1xRuy937	US	114	G73	Q1	Argentina
44	G138	G18	Argentina	115	G74	2-want	Bolivia
45	G139	Unknown	Argentina	116	G75	Chadmo	Chile
46	G14	G15	Argentina	117	G76	59ALC	Chile x Bolivia
47	G140	G18	Argentina	118	G77	37-Test	Chile x Bolivia
48	G141	Chadmo	Chile	119	G78	Sayana	Bolivia
49	G142	EDK-4	Netherland	120	G79	2-want	Bolivia
50	G15	G49	Argentina	121	G8	Real1xRuy937	US
51	G16	Atlas	Netherland	122	G80	Real1xRuy937	US
52	G17	Chadmo	Chile	123	G81	Sayana	Bolivia
53	G18	Atlas	Netherland	124	G82	G26	Argentina

No.	Genotype	Selected from	Orgin	No.	Genotype	Selected from	Orgin
54	G19	Isluga	Chile	125	G83	Real1xRuy937	US
55	G2	Atlas	Netherland	126	G84	Lucaboldo	Chile
56	G20	Lucaboldo	Chile	127	G85	Ahachuha	Bolivia
57	G21	Chadmo	Chile	128	G86	Lucaboldo	Chile
58	G22	Riobamba	Netherland	129	G87	FaroxChadmo	Chile
59	G23	Unknown	Argentina	130	G88	Cahuil	Chile
60	G24	G2	Argentina	131	G89	Chadmo	Chile
61	G25	Q4	Argentina	132	G9	Black quinoa	Bolivia
62	G26	Real1xRuy937	US	133	G90	G2	Argentina
63	G27	Q3	Argentina	134	G91	Atlas	Netherland
64	G28	Chadmo	Chile	135	G92	Chadmo	Chile
65	G29	Real1xRuy937	US	136	G93	Isluga	Chile
66	G3	G20	Argentina	137	G94	Q1	Argentina
67	G30	Ahachuha	Bolivia	138	G95	G3	Argentina
68	G31	G26	Argentina	139	G96	G18	Argentina
69	G32	EDK-4	Netherland	140	G97	Q5	Argentina
70	G33	Riobamba	Netherland	141	G98	Real1xRuy937	US
71	G34	Atlas	Netherland	142	G99	Q1	Argentina

Phenotypic measurements

Measurements were taken at seven-day intervals, indexed by days after sowing (DAS). Plant height (PH, cm, soil surface to the tip of the main stem) and stem diameter (STD, mm, 2 cm above the soil surface, digital caliper, 0.01-mm precision) were recorded at 30, 37, 44, 51, 58, 65, and 72 DAS. Leaf number (NoL, leaves longer than 1 cm on the main stem, excluding cotyledons and senescent leaves) and branch number (NoB, axillary branches longer than 1 cm originating from the main stem) were recorded at 30, 37, 44, and 51 DAS. Leaf area (LA, cm^2, CID-202 leaf area meter, CID Bio-Science) and SPAD chlorophyll index (SPAD-502, Konica Minolta; mean of three readings on fully expanded upper leaves) were recorded destructively at 30 and 58 DAS. At harvest, individual grain yield (IGY, g $plant^{-1}$) and thousand-grain weight (P1000, g) were determined after threshing and drying to 14.0% moisture, verified with a PM650 portable grain-moisture meter. Branch number at 30 DAS was constant across all genotypes and was excluded from trait-relationship and diversity analyses, leaving 27 traits.

Trait relationships and diversity structure

Pairwise Pearson correlations among the 27 traits were computed with corr_coef (metan v1.19, R v4.5.1) (Olivoto and Nardino, 2021) and summarized as a correlation network restricted to edges with $|r| \geq 0.2$. Principal component analysis (PCA) was fit on the same 27 traits after mean-centering and unit-variance scaling. Diversity structure was resolved separately, on the eight traits measured at their final observation time (PH72, STD72, NoL51, NoB51, SPAD58, LA58, P1000, IGY) rather than the full time course, to avoid the temporal autocorrelation inherent to repeated within-plant measurements. Genotype means for these eight traits were subjected to PCA (three components retained; FactoMineR v2.11; (Lê et al., 2008) followed by hierarchical clustering on principal components (HCPC), which partitioned the panel into five clusters.

Machine-learning classification of high-yield genotypes

Data aggregation and definition of the high-yield group

This study analyzed 1,420 plant records from 142 quinoa genotypes, with 10 plants measured per genotype. We treated the genotype as the unit of analysis and calculated the arithmetic mean of the 10 plants for each measured variable, including individual grain yield (IGY). This aggregation produced one phenotype vector and one mean IGY value per genotype and avoided treating multiple plants from the same genotype as genetically independent observations. We defined the high-yield class using the observed distribution of genotype-mean IGY. Genotypes with mean IGY greater than or equal to the 80th percentile (33.4954) were classified as high yield, whereas the remaining genotypes were assigned to the other class. This criterion identified 29 high-yield and 113 other genotypes (Fig. 2).

Phenotype predictor sets

Three prespecified predictor sets were compared to determine how classification performance changed with the amount and timing of phenotype information. The early-day-30 set contained six measurements collected at day 30: plant height (PH30), number of leaves (NoL30), number of branches (NoB30), stem diameter (STD30), SPAD value (SPAD30), and leaf area (LA30). The complete-vegetative set contained 26 direct morphological, SPAD, and leaf-area measurements collected across the available observation times; IGY was excluded because it defined the response class and excluded thousand-grain weight (P1000) from this predictor set. The complete-plus-P1000 set contained the same 26 predictors together with P1000. Table 2 summarizes the composition and PCA dimensionality of the three predictor sets.

Table 2
Composition of the three phenotype predictor sets and the number of principal components retained to explain at least 85% of the standardized predictor variance.

Predictor set	Measurements	No. inputs	No. PCs	Variance (%)
Early day 30	PH30, NoL30, NoB30, STD30, SPAD30, and LA30	6	4	91.36
Complete vegetative	All direct morphology, SPAD, and leaf-area measurements; IGY and P1000 excluded	26	10	86.65
Complete plus P1000	Complete vegetative set plus P1000	27	10	85.02

Standardization and principal component analysis

The predictor data were confirmed to contain no missing values; therefore, no imputation was required. For each predictor set, we standardized every variable across the 142 genotype means by subtracting its mean and dividing by its standard deviation. We then fitted principal component analysis (PCA) separately to each standardized predictor matrix. We retained components in descending order of explained variance until the cumulative explained variance reached at least 85%. This criterion retained four components for the early-day-30 set and 10 components for each complete set (Table 2). Figure 3 illustrates the cumulative explained variance for the complete-plus-P1000 set.

We estimated the standardization parameters and PCA transformation once using the complete panel of 142 genotypes and subsequently used the resulting fixed PCA scores in repeated cross-validation. We selected this global-PCA design to provide a common exploratory representation of the observed genotype panel. Because the complete panel contributed to preprocessing, the resulting cross-validation estimates describe within-panel discrimination in this fixed representation and do not constitute a strict estimate of predictive performance in an independent genotype population.

Classification algorithms and repeated cross-validation

Four classification algorithms were compared using the retained PCA scores: class-weighted logistic regression, shrinkage linear discriminant analysis (Elbasi et al.), a radial-basis-function support-vector machine (RBF-SVM), and random forest. Logistic regression used an inverse regularization strength of C = 1, balanced class weights, the liblinear solver, and a maximum of 10,000 iterations. LDA used the least-squares solver with automatic covariance shrinkage. The RBF-SVM used C = 1, the scale-based default for the kernel coefficient, balanced class weights, and probability estimation. Random forest used 100 trees, square-root feature sampling, a minimum leaf size of two observations, and class weights recalculated for each bootstrap sample. We prespecified these settings and did not tune hyperparameters using the reported cross-validation results.

The present study evaluated each predictor-set and classifier combination using stratified five-fold cross-validation repeated 10 times with random seed 20260724. In each repetition, stratification approximately

preserved the proportion of high-yield genotypes across the five folds. Within each repetition, each genotype appeared once in a held-out fold and received one out-of-fold score from a classifier trained on the other four folds using the fixed global-PCA representation. Thus, each genotype received 10 out-of-fold scores for every classifier and predictor set. We used the mean of these 10 scores to summarize model-specific classification for each genotype. To obtain a descriptive multi-model ranking, we averaged the four model-specific mean scores with equal weight and refer to the resulting value as the consensus score.

Performance measures and implementation

The precision-recall curve (PR-AUC) was used as the primary performance measure because only 29 of the 142 genotypes belonged to the high-yield class. For a non-informative ranking, the expected PR-AUC is approximately equal to the high-yield prevalence (29/142 = 0.204). This study also calculated the area under the receiver-operating-characteristic curve (ROC-AUC), balanced accuracy, sensitivity, specificity, precision, and F1 score. PR-AUC and ROC-AUC evaluate the ranking of genotypes across possible decision thresholds, whereas the remaining measures require a specified threshold. We calculated balanced accuracy, sensitivity, specificity, precision, and F1 using a threshold of 0.5. For each metric, we first evaluated the complete set of out-of-fold predictions from each repetition and then reported the mean and standard deviation across the 10 repetitions.

The analysis was implemented in Python using pandas, NumPy, scikit-learn, Matplotlib, and openpyxl. We fixed the random seed at 20260724 for all stochastic operations. The analysis used 10 repetitions of stratified five-fold cross-validation, the predictor sets and model settings specified above, and a decision threshold of 0.5 for threshold-dependent performance measures.

3 Results

Phenotypic diversity and hierarchical classification of quinoa genotypes

Hierarchical cluster analysis classified the 142 evaluated genotypes into five distinct groups (Fig. 4). Group I comprised 17 genotypes, including G60, G109, G46, G10, G100, G12, G49, G55, G133, G34, G50, G51, G13, G47, G28, G11, and G32. Group II consisted of 14 genotypes: G16, G35, G21, G94, G18, G9, G39, G59, G4, G7, G37, G96, G83, and G44. Group III, containing 35 genotypes, included G40, G135, G112, G139, G3, G42, G121, G5, G6, G25, G29, G107, G95, G98, G102, G103, G61, G105, G67, G120, G131, G122, G126, G56, G45, G87, G136, G86, G119, G132, G75, G97, G123, G113, and G99. Group IV comprised 15 genotypes, namely G20, G24, G17, G22, G52, G104, G48, G27, G80, G57, G1, G31, G30, G117, and G8. The largest cluster, Group V, contained 61 genotypes: G101, G127, G111, G137, G124, G129, G106, G19, G23, G26, G71, G72, G73, G54, G2, G36, G78, G83, G82, G84, G85, G125, G110, G15, G81, G108, G89, G74, G93, G63, G90, G76, G88, G65, G58, G91, G69, G66, G68, G115, G116, G134, G130, G62, G53, G141, G79, G128, G140, G77, G92, G14, G41, G114, G33, G118, G138, G43, G142, G64, and G70.

Developmental relationships among agro-morphological traits and individual grain yield

Correlations between growth parameters and yield across the quinoa germplasm collection ranged from − 0.33 to 0.92 (Figs. 5, 6). Remarkably, the growth parameters, specifically PH, STD, and NoL, exhibited strong positive correlations (Figs. 5, 6).

PH showed the strongest relationship among growth stages. The correlation coefficient was highest between PH at 30 DAS and 37 DAS (r = 0.91***). Meanwhile, PH at 58, 65, and 72 DAS also showed strong positive correlations with each other, with correlation coefficients ranging from 0.82*** to 0.86***. Similarly, STD exhibited strong positive correlations, particularly between measurement values at late growth stages (from 58 to 72 DAS). These findings indicate that genotype strongly contributed to growth parameters. Genotypes exhibiting greater PH and STD during the early growth stages can maintain higher PH and STD in later stages. A positive correlation was also observed between PH and STD, especially at the end of the vegetative growth phase, suggesting that genotypes with greater PH tended to have larger stem diameters (Figs. 5, 6).

Similar results were observed in the relationships between NoL and NoB among growth stages, with correlation coefficients ranging from 0.53*** to 0.89*** and from 0.38*** to 0.43***, respectively. A moderate correlation was also found between NoB and NoL (0.18* to 0.40***) and between NoB and PH (r = 0.19*** to 0.42***) during the period from 37 to 51 DAS. This indicates that the vigor of a genotype's vegetative growth can be forecast at the early stage (30 DAS) (Figs. 5, 6).

A positive correlation ranging from 0.21* to 0.35*** was observed between NoL and LA; however, this correlation tended to decrease during the later stages of growth. This suggests that plants producing a higher NoL during the early growth stages can maintain a larger LA in the subsequent stages (Figs. 5, 6).

Notably, significant negative correlations were observed among the parameters, with coefficients ranging from − 0.14* to -0.43***. Negative correlations appeared between SPAD and PH and STD; stronger correlations were found at 30 DAS compared to those at 58 DAS. This can be explained by the fact that during the early stage, genotypes with vigorous vegetative growth allocate nutrients primarily to cell division for stem development in increases in PH and STD rather than to accumulation in the leaves (Figs. 5, 6).

Grain yield showed virtually no correlation with most growth parameters. A weak but significant negative correlation was found only between IGY and PH at 72 DAS. Whereas stronger positive correlations were observed between IGY and P1000 (0.23**) and between IGY and LA at 58 DAS (0.44***). Thus, LA at the stage of full leaf development plays a crucial role in crop yield, making it a viable parameter for yield prediction. Furthermore, potential yield can be predicted based on leaf development across growth stages, given its significant contribution to maintaining LA at the 58 DAS maturity stage (Fig. 5, 6).

Multivariate structure of phenotypic variation and yield-class distribution

PCA representation of the phenotype panel

We retained four principal components for the early-day-30 set, which together explained 91.36% of the standardized predictor variance. The complete-vegetative and complete-plus-P1000 sets each required 10 components to reach the prespecified 85% criterion; these components explained 86.65% and 85.02% of the variance, respectively (Table 2). The larger number of retained components indicates that the complete phenotype profiles contained multiple dimensions of variation that could not be represented adequately by only one or two axes.

In the complete-plus-P1000 analysis, PC1 and PC2 explained 30.1% and 15.0% of the standardized predictor variance, respectively (Fig. 7). The high-yield and other genotypes overlapped substantially in this two-dimensional projection, although some high-yield genotypes occupied relatively distinct positions. This overlap indicates that the first two components alone did not separate the classes. The classification analysis therefore used all 10 retained components, whereas Fig. 3 displays only the two leading axes for visualization.

Classification performance across phenotype predictor sets and algorithms

We found that the day-30 measurements provided limited discrimination between the high-yield and other genotypes. Random forest produced the highest PR-AUC for this predictor set, but its mean PR-AUC of 0.249 was only modestly above the prevalence reference of 0.204. Across the four algorithms, ROC-AUC ranged from 0.471 to 0.583 (Table 3 and Fig. 8). Shrinkage LDA and RBF-SVM did not identify any high-yield genotypes at the 0.5 threshold, resulting in zero sensitivity and F1 scores. These findings indicate that the available day-30 measurements alone were insufficient for reliable classification within this panel.

Table 3
Classification performance for the three predictor sets and four algorithms, summarized across 10 repetitions of stratified five-fold cross-validation.

Predictor set	Model	PR-AUC, mean (SD)	ROC-AUC	Bal. acc.	Sens.	Spec.	Prec.	F1
Early day 30	Logistic	0.232 (0.014)	0.583	0.587	0.624	0.550	0.262	0.369
	LDA	0.227 (0.014)	0.573	0.497	0.000	0.994	0.000	0.000
	SVM RBF	0.204 (0.018)	0.471	0.500	0.000	1.000	0.000	0.000
	Random forest	0.249 (0.024)	0.524	0.505	0.100	0.911	0.228	0.137
Complete vegetative	Logistic	0.532 (0.045)	0.786	0.743	0.731	0.755	0.434	0.544
	LDA	0.557 (0.042)	0.808	0.638	0.324	0.951	0.631	0.428
	SVM RBF	0.442 (0.054)	0.773	0.551	0.148	0.954	0.456	0.222
	Random forest	0.528 (0.057)	0.790	0.590	0.224	0.957	0.574	0.320
Complete + P1000	Logistic	0.626 (0.038)	0.829	0.762	0.734	0.790	0.474	0.576
	LDA	0.643 (0.033)	0.845	0.674	0.400	0.949	0.667	0.499
	SVM RBF	0.496 (0.047)	0.807	0.592	0.245	0.939	0.514	0.330
	Random forest	0.616 (0.062)	0.836	0.653	0.348	0.958	0.675	0.459

Note: Values are means across 10 repetitions, except that PR-AUC is reported as mean (SD). Balanced accuracy, sensitivity, specificity, precision, and F1 were calculated using a decision threshold of 0.5. PR-AUC, area under the precision-recall curve; ROC-AUC, area under the receiver-operating-characteristic curve; LDA, linear discriminant analysis; RBF-SVM, radial-basis-function support-vector machine.

Classification improved markedly when we used the complete vegetative profile. In this predictor set, LDA achieved the highest PR-AUC (0.557), whereas logistic regression provided the highest balanced accuracy (0.743) by combining a sensitivity of 0.731 with a specificity of 0.755. The complete-plus-P1000 set produced the strongest overall discrimination among the three predictor sets evaluated. With the complete-plus-P1000 set, LDA achieved the highest PR-AUC at 0.643 (SD 0.033) and the highest ROC-AUC at 0.845. Logistic regression achieved a similar PR-AUC of 0.626 and the highest balanced accuracy of

0.762, with sensitivity of 0.734 and specificity of 0.790. Random forest achieved a PR-AUC of 0.616 and the highest specificity of 0.958, but its sensitivity was only 0.348. Thus, the complete-plus-P1000 predictors produced the strongest ranking performance, while the preferred classifier depended on whether the objective emphasized sensitivity or specificity.

Figure 9 summarizes the trade-offs among performance measures. For the complete-plus-P1000 set, LDA and random forest favored specificity and precision over sensitivity, producing relatively selective high-yield classifications. Logistic regression identified a larger proportion of the observed high-yield genotypes and provided the highest balanced accuracy. Under the fixed hyperparameter settings, RBF-SVM produced lower PR-AUC than logistic regression, LDA, and random forest for the complete-plus-P1000 set. These differences show that the apparent performance of a classifier depends on whether the application prioritizes recovering more high-yield genotypes or limiting false-positive selections.

We additionally applied a threshold of 0.5 to each genotype's mean out-of-fold LDA probability from the complete-plus-P1000 analysis. The model correctly classified 108 of the 113 other genotypes and incorrectly selected five as high yield. Among the 29 observed high-yield genotypes, it detected 12 and missed 17 (Fig. 10). This aggregate confusion matrix confirms that LDA produced a conservative selection rule: it excluded most other genotypes but recovered fewer than half of the observed high-yield genotypes.

Phenotype profile associated with high-yield membership

We compared the standardized phenotype means of the high-yield and other groups to identify the largest descriptive differences between the two classes (Fig. 11). P1000 showed the largest positive difference: its mean in the high-yield group was 1.100 standard deviations higher than that in the other group. The high-yield group also had higher LA58 (0.840 SD), NoL51 (0.819 SD), SPAD58 (0.642 SD), NoL37 (0.533 SD), NoL44 (0.483 SD), and NoL30 (0.436 SD). Within this panel, the observed high-yield profile was therefore associated with greater thousand-grain weight, larger late leaf area, more leaves, and higher late SPAD values.

By contrast, the largest negative differences were observed for NoB51 (-0.531 SD), PH72 (-0.318 SD), and PH51 (-0.254 SD). The observed high-yield group was therefore not characterized simply by greater plant height or branch number. These comparisons are descriptive associations between phenotype measurements and the yield-defined classes; they do not establish that modifying any individual trait would cause a corresponding change in yield.

Consensus-based prioritization of high-yield and phenotypically similar genotypes

We calculated a consensus score for each genotype by averaging its four model-specific mean out-of-fold scores with equal weight. For the complete-plus-P1000 predictor set, G12 ranked first with a consensus score of 0.859, followed by G49 (0.776), G10 (0.743), G128 (0.736), G44 (0.715), G30 (0.711), and G55 (0.698) (Fig. 12 and Table 4). G60 had the highest mean IGY among the 15 leading genotypes (54.048) but

ranked tenth by consensus score (0.678). The ranking therefore reflects similarity to the multivariate phenotype patterns learned by the four classifiers rather than yield magnitude alone.

Table 4
Top 15 genotypes ranked by the consensus of four model-specific mean out-of-fold scores for the complete-plus-P1000 predictor set

Rank	Genotype	Mean IGY	Observed high yield	Consensus score
1	G12	52.751	Yes	0.859
2	G49	48.619	Yes	0.776
3	G10	41.109	Yes	0.743
4	G128	36.112	Yes	0.736
5	G44	50.327	Yes	0.715
6	G30	26.320	No	0.711
7	G55	34.713	Yes	0.698
8	G81	30.232	No	0.696
9	G15	48.336	Yes	0.683
10	G60	54.048	Yes	0.678
11	G100	23.757	No	0.656
12	G18	51.901	Yes	0.631
13	G117	23.875	No	0.628
14	G14	26.546	No	0.627
15	G33	27.962	No	0.620

Note: The consensus score is the equally weighted mean of the model-specific mean out-of-fold scores from logistic regression, LDA, RBF-SVM, and random forest. It is used for descriptive ranking and should not be interpreted as an externally calibrated probability of high-yield performance.

Nine of the top 15 consensus-ranked genotypes belonged to the observed high-yield class, whereas six-G30, G81, G100, G117, G14, and G33-fell below the yield threshold. The below-threshold genotypes nevertheless received relatively high scores because their measured phenotype profiles resembled patterns associated with the high-yield class. We therefore interpret them as genotypes with high-yield-like phenotype profiles rather than as confirmed high-yield genotypes. Overall, the results show that complete phenotyping, particularly the predictor set that included P1000, provided useful within-panel discrimination and prioritization. In contrast, the day-30 measurements alone provided alimited discrimination between the high-yield and other genotypes.

4 Discussion

The hierarchical cluster analysis (HCA) successfully partitioned the 142 evaluated quinoa genotypes into five distinct groups (Groups I to V) based on their agro-morphological traits (Fig. 4), providing a more comprehensive representation of phenotypic diversity than previous germplasm characterization studies, which generally involved smaller genotype collections evaluated under specific environmental conditions. For example, earlier studies examined 32 genotypes for morphological diversity (El-Harty et al., 2021), 52 genotypes for salinity responses (Nguyen et al., 2020), and (Nguyen et al., 2022), whereas the present analysis integrates a substantially larger diversity panel, thereby providing a higher-resolution characterization of the phenotypic landscape of quinoa. The identification of phenotypically divergent clusters provides valuable opportunities for selecting complementary parental materials for crossing programs aimed at broadening the genetic base and exploiting heterosis, while the systematic characterization of large and diverse germplasm collections is essential for minimizing genetic erosion associated with intensive selection for a limited number of elite traits. In this context, the present findings contribute to the broader research priority of large-scale diversity screening for identifying genetically diverse and stronger foundation for quinoa germplasm utilization, parent selection, and future genetic improvement programs (Hinojosa et al., 2018).

The wide phenotypic variation observed among the 142 quinoa genotypes indicates that the panel comprises genetically diverse materials with distinct growth, physiological, and grain characteristics; however, this high level of diversity also complicates the direct identification of genotypes with superior yield potential. Correlations that ranged from − 0.33 to 0.92 (Fig. 5) and the requirement for 10 principal components to explain at least 85% of the variance in the complete predictor sets show that phenotypic differences occupied several partly independent dimensions (Table 2, Fig. 3), while the overlap between yield classes on PC1 and PC2 shows that the dominant axes of variation were not equivalent to yield ranking (Fig. 7). Comparable diversity has been reported in a world core collection of 360 accessions, in which agronomic and seed traits separated the accessions into four groups and exposed useful high-yield materials outside established breeding lines (Craine et al., 2023), whereas multiyear analysis of 310 accessions identified stable associations for 17 traits across a highly variable quinoa panel (Patiranage et al., 2022). These results support selection for a high-yield phenotype that is defined by a combination of canopy, physiological, and grain traits instead of selection from plant size or a single principal axis. A breeding decision based on this multivariate phenotype can prioritize high-yield genotypes while retaining candidates from different parts of the phenotypic distribution, which limits premature narrowing of the available germplasm.

The strong correlations among repeated PH, STD, and NoL measurements indicate that genotypic differences in vegetative architecture persisted through much of development, but persistence of plant size did not confer a comparable advantage in IGY. Most vegetative traits had weak relations with IGY, PH72 had a negative relation, and the stronger positive coefficients for LA58 and P1000 identify canopy area near reproductive demand and grain mass as more useful selection traits (Fig. 5, 6). The positive relation between yield and thousand-seed weight reported in the quinoa world core collection supports this result (Craine et al., 2023), while the high-density mapping study identified pleiotropic loci for flowering time, plant height, and thousand-kernel weight (Maldonado-Taipe et al., 2022). The relation

between NoL and LA weakened at later dates, which indicates that leaf number loses value as a surrogate for canopy area after differences in leaf expansion, senescence, and architecture become pronounced, and the early negative associations of SPAD with PH and STD may indicate allocation differences between structural growth and leaf chlorophyll status (Figs. 5, 6). The previous study identified grain number as the principal yield component in coastal lowland quinoa and reported negative associations between grain number and grain weight (Dumschott et al., 2022), so P1000 should be treated as an informative component of the high-yield profile and not as an independent cause of productivity (Dumschott et al., 2022). The desirable phenotype in this panel therefore combines sufficient canopy area and chlorophyll status near 58 DAS, moderate late stature, and favorable seed mass, resulting in a balanced plant architecture characterized by moderate height, limited branching, and a compact, well-developed panicle with abundant seeds (Fig. 13). This phenotype contrasts with excessively tall and highly branched plants with loose panicles and small seeds, as well as short plants with sparse panicles and consequently very low yield potential.

The comparison of predictor sets indicates that phenotyping effectiveness depends on developmental coverage and the inclusion of reproductive traits. Measurements at 30 DAS primarily reflected early vegetative development and showed limited discriminatory value for high-yield membership. Expanding the vegetative profile and incorporating P1000 improved discrimination, as reflected by PR-AUC and ROC-AUC, highlighting the value of integrating complementary vegetative and reproductive traits for identifying high-yielding genotypes (Table 3, Fig. 8). This improvement agrees with evidence that phenology, height, and thousand-kernel weight can share pleiotropic loci (Maldonado-Taipe et al., 2022) and with multiyear evidence that agronomic performance in quinoa has a complex genetic basis across numerous trait-associated regions (Patiranage et al., 2022). Within the present panel, the high-yield group had higher standardized means for P1000, LA58, NoL51, and SPAD58, while NoB51 and PH72 were lower (Fig. 11). The overlap of the two yield classes on the first two PCA axes does not conflict with the classifier results, since the models used information distributed across 10 retained components that included later canopy traits and grain mass. Phenotyping near 58 DAS and measurement of P1000 therefore provided the most informative combination for identification of the yield-associated phenotype in this experiment.

Differences among classifiers have direct consequences for germplasm screening, since logistic regression recovered more observed high-yield genotypes, whereas LDA and random forest produced higher specificity and precision and selected a smaller candidate set. The consensus ranking placed G12, G49, G10, G128, and G44 among the leading observed high-yield genotypes, while G30, G81, G100, G117, G14, and G33 ranked highly despite yields below the empirical threshold (Fig. 12). The latter group should be considered phenotypic analogues for field reassessment, and their consensus scores should not be treated as calibrated probabilities or proof of superior yield. Two analytical choices also restrict the predictive claims, since the high-yield threshold was derived from the same 142-genotype panel and standardization plus PCA were fitted before cross-validation, which allowed held-out genotypes to influence the predictor representation. Cross-validation studies have shown that preprocessing with the full dataset can produce optimistic performance estimates, especially when sample size is limited or multiple correlated traits enter the model (Runcie and Cheng, 2019; Vabalas et al., 2019). The reported

metrics should therefore be described as within-panel discrimination, and future evaluation should refit standardization and PCA within each training fold, use nested cross-validation when model settings are tuned, retain a fully unseen test population, and examine candidate performance across seasons, locations, and genetically distinct material. Under these conditions, the proposed workflow can serve as a screening tool that reduces the field-testing burden while keeping final selection anchored to replicated grain-yield evaluation.

5 Conclusion

The 142 quinoa genotypes exhibited substantial phenotypic diversity and were classified into five distinct groups. High-yield membership was associated with a balanced phenotype characterized by greater canopy development and chlorophyll status near 58 DAS, favorable seed mass, and moderate plant stature and branching. Complete phenotyping, particularly the inclusion of P1000, substantially improved within-panel discrimination compared with early vegetative measurements alone. Among the machine-learning classifiers, LDA achieved the highest PR-AUC and ROC-AUC, whereas logistic regression provided the highest balanced accuracy and recovered a larger proportion of observed high-yield genotypes. The consensus ranking further prioritized both confirmed high-yield genotypes and phenotypically similar candidates for further field evaluation. The integration of large-scale phenotypic characterization with multivariate and machine-learning analyses provides a useful framework for germplasm screening and candidate prioritization in quinoa breeding, although independent and multi-environment validation is required before predictive application.

Declarations

Declarations

Conflict of interest:

The authors declare no conflict of interest.

Ethical Approval and Consent to Participate:

Not applicable

Funding:

This study was supported from the National Foundation For Science and Technology Development (NAFOSTED) program of Vietnam on “**Study on the waterlogging tolerance of Quinoa (Chenopodium quinoa Willd.) for development of Climate-Resilient Crops in Vietnam**” from 2025–2028 (grant No. 106.03-2025.03) awarded to Vietnam National University of Agriculture (VNUA).

Funding

acquisition: VLN

Methodology: VLN, DB

Writing - original draft: V-LN, TNDN, TTHP, DHC, VNL

Writing - review & editing: All authors

Author Contribution

Conceptualization: VLN, DBFormal analysis: TDN, THNP, NCFunding acquisition: VLNMethodology: VLN, DBWriting - original draft: V-LN, TNDN, TTHP, DHC , VNLWriting - review & editing: All authors

Acknowledgement

We sincerely thank Professor Robert van Loo and Professor Daniel Danial, Wageningen University & Research (WUR), the Netherlands, for providing several quinoa genotypes used in this study, as well as for their valuable guidance and support in developing quinoa cultivation methods suitable for the conditions of Vietnam.

Data Availability

Data sets generated during the current study are available from the corresponding author on reasonable request.

References

1. Aggarwal, N., Rathore, M., Jan, F., Sharma, D., Kumar, S., Thudi, M., Jighly, A., Varshney, R. K., and Mir, R. R. (2026). Harnessing artificial intelligence in plant breeding: innovations in digital phenotyping and breeding methodologies. *Theoretical and Applied Genetics* 139, 188.
2. Angeli, V., Miguel Silva, P., Crispim Massuela, D., Khan, M. W., Hamar, A., Khajehei, F., Graeff-Hönninger, S., and Piatti, C. (2020). Quinoa (Chenopodium quinoa Willd.): An Overview of the Potentials of the “Golden Grain” and Socio-Economic and Environmental Aspects of Its Cultivation and Marketization. *Foods* 9, 216.
3. Aschard, H., Vilhjálmsson, B. J., Greliche, N., Morange, P. E., Trégouët, D. A., and Kraft, P. (2014). Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am J Hum Genet* 94, 662–76.

4. Bista, M. K., Rutland, W., Pieralisi, B., and Bheemanahalli, R. (2026). Genotype × environment interaction and multi-trait stability of cotton lint yield and fiber quality across irrigated and rainfed environments. *Field Crops Research* 343, 110484.
5. Coppens, F., Wuyts, N., Inzé, D., and Dhondt, S. (2017). Unlocking the potential of plant phenotyping data through integration and data-driven approaches. *Current Opinion in Systems Biology* 4, 58–63.
6. Craine, E. B., Davies, A., Packer, D., Miller, N. D., Schmöckel, S. M., Spalding, E. P., Tester, M., and Murphy, K. M. (2023). A comprehensive characterization of agronomic and end-use quality phenotypes across a quinoa world core collection. *Front Plant Sci* 14, 1101547.
7. Curti, R. N., de la Vega, A. J., Andrade, A. J., Bramardi, S. J., and Bertero, H. D. (2014). Multi-environmental evaluation for grain yield and its physiological determinants of quinoa genotypes across Northwest Argentina. *Field Crops Research* 166, 46–57.
8. del Castillo, C., Winkel, T., Mahy, G., and Bizoux, J.-P. (2007). Genetic structure of quinoa (Chenopodium quinoa Willd.) from the Bolivian altiplano as revealed by RAPD markers. *Genetic Resources and Crop Evolution* 54, 897–905.
9. Dumschott, K., Wuyts, N., Alfaro, C., Castillo, D., Fiorani, F., and Zurita-Silva, A. (2022). Morphological and physiological traits associated with yield under reduced irrigation in Chilean coastal lowland quinoa. *Plants* 11, 323.
10. El-Harty, E. H., Ghazy, A., Alateeq, T. K., Al-Faifi, S. A., Khan, M. A., Afzal, M., Alghamdi, S. S., and Migdadi, H. M. (2021). Morphological and molecular characterization of quinoa genotypes. *Agriculture* 11, 286.
11. Elbasi, E., Zaki, C., Topcu, A. E., Abdelbaki, W., Zreikat, A. I., Cina, E., Shdefat, A., and Saker, L. (2023). Crop prediction model using machine learning algorithms. *Applied Sciences* 13, 9288.
12. Hassani, M., Mahmoudi, S. B., Saremirad, A., and Taleghani, D. (2024). Genotype by environment and genotype by yield*trait interactions in sugar beet: analyzing yield stability and determining key traits association. *Sci Rep* 13, 23111.
13. Hazra, S., Ray, D. D., Hazra, A., Chatterjee, A., Mukhopadhyay, A., Ghorai, S., Ghosh, P. K., Bose, S., Roy, S., and Nath, T. (2026). Multivariate and nonlinear analyses to characterize phenotypic variability in landraces and cultivars of rice (Oryza sativa L.). *Plant Physiology Reports* 31, 133–146.
14. Hinojosa, L., González, J. A., Barrios-Masias, F. H., Fuentes, F., and Murphy, K. M. (2018). Quinoa abiotic stress responses: A review. *Plants* 7, 106.
15. Jacobsen, S.-E. (2003). The worldwide potential for quinoa (Chenopodium quinoa Willd.). *Food reviews international* 19, 167–177.
16. Küçüközdemir, Ü., Tosun, M., Dumlu, B., Alipour, H., and Türkoğlu, A. (2026). Do modern winter wheat cultivars outperform traditional landraces under cold stress? A large-scale agro-morphological evaluation. *BMC Plant Biology*.
17. Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of statistical software* 25, 1–18.

18. Maestro-Gaitán, I., Granado-Rodríguez, S., Orús, M. I., Matías, J., Cruz, V., Bolaños, L., and Reguera, M. (2022). Genotype-dependent responses to long-term water stress reveal different water-saving strategies in Chenopodium quinoa Willd. *Environmental and Experimental Botany* 201, 104976.
19. Maldonado-Taipe, N., Barbier, F., Schmid, K., Jung, C., and Emrani, N. (2022). High-Density Mapping of Quantitative Trait Loci Controlling Agronomically Important Traits in Quinoa (Chenopodium quinoa Willd.). *Front Plant Sci* 13, 916067.
20. Mousa, M. A. A., Asiry, K. A., Al-Qurashi, A. D., Almasoudi, N. M., and Elsayed, M. I. (2025). Genotypic variation in yield, physiological traits, and drought tolerance of quinoa (Chenopodium quinoa Willd.) under arid conditions. *Front Plant Sci* 16, 1679444.
21. Nazeer, S., and Akram, M. Z. (2025). Elucidating Genotypic Variation in Quinoa via Multidimensional Agronomic, Physiological, and Biochemical Assessments. *Plants* 14, 2332.
22. Nguyen, L. V., Bertero, D., Hoang, D. T., and Long, N. V. (2022). Variation in quinoa roots growth responses to drought stresses. *Journal of Agronomy and Crop Science* 208, 830–840.
23. Nguyen, L. V., Bertero, D., and Nguyen, L. V. (2020). Genetic variation in root development responses to salt stresses of quinoa. *Journal of Agronomy and Crop Science* 206, 538–547.
24. Olivoto, T., and Nardino, M. (2021). MGIDI: Toward an effective multivariate selection in biological experiments. *Bioinformatics* 37, 1383–1389.
25. Patel, H. P., Nair, R. J., Prajapati, N. N., Zala, H. N., Soni, N. V., and Soni, K. (2026). Unraveling Yield Performance and Nutritional Trait Architecture in Quinoa (Chenopodium quinoa Willd.) Using Integrated Variability, Correlation and Path Analysis. *Archives of Current Research International* 26, 155–169.
26. Patiranage, D. S. R., Rey, E., Emrani, N., Wellman, G., Schmid, K., Schmöckel, S. M., Tester, M., and Jung, C. (2022). Genome-wide association study in quinoa reveals selection pattern typical for crops with a short breeding history. *eLife* 11, e66873.
27. Rakasi, N., Kienbaum, L., Böndel, K. B., Wiederstein, J.-D., Raju, N. K. G., Schmöckel, S. M., and Schmid, K. J. (2026). Phenotypic Differentiation Between Highland and Coastal Quinoa Under Cold Stress Conditions. *Plant, Cell & Environment* n/a.
28. Runcie, D., and Cheng, H. (2019). Pitfalls and Remedies for Cross Validation with Multi-trait Genomic Prediction Methods. *G3 Genes|Genomes|Genetics* 9, 3727–3741.
29. Saravani, A. F., Rabiei, B., and Laki, E. S. (2026). Evaluation of quinoa (Chenopodium quinoa Willd.) genotypes using genotype by trait (GT) and genotype by yield × trait (GYT) methods. *Scientific Reports* 16, 16970.
30. Sheikh, M., Iqra, F., Ambreen, H., Pravin, K. A., Ikra, M., and Chung, Y. S. (2024). Integrating artificial intelligence and high-throughput phenotyping for crop improvement. *Journal of Integrative Agriculture* 23, 1787–1802.
31. Tsega, A., and Mullualem, D. (2026). Machine learning for multi-omics data integration in crop improvement: a systematic review. *BMC Bioinformatics*.

32. Vabalas, A., Gowen, E., Poliakoff, E., and Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE* 14, e0224365.
33. Yadav, R., Gore, P. G., Gupta, V., and Siddique, K. H. (2023). Quinoa (Chenopodium quinoa Willd.)—A smart crop for food and nutritional security. *In* "Neglected and Underutilized Crops", pp. 23–43. Elsevier.

Figures



Figure 1

Experimental field and quinoa plants

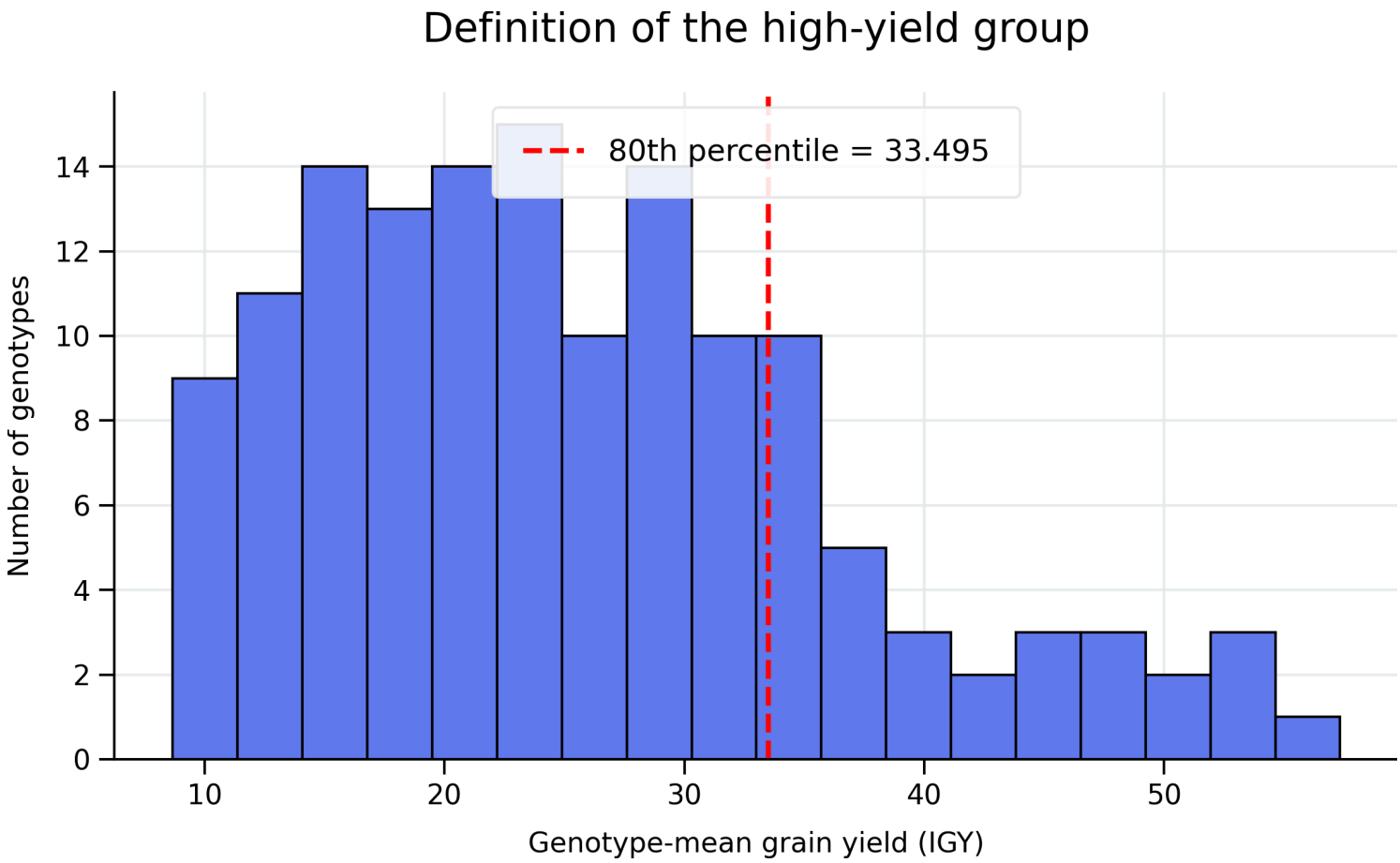


Figure 2

Distribution of mean individual grain yield (IGY) across the 142 quinoa genotypes. The dashed red line marks the 80th-percentile threshold (IGY = 33.4954) used to classify 29 genotypes as high yield and the remaining 113 genotypes as other.

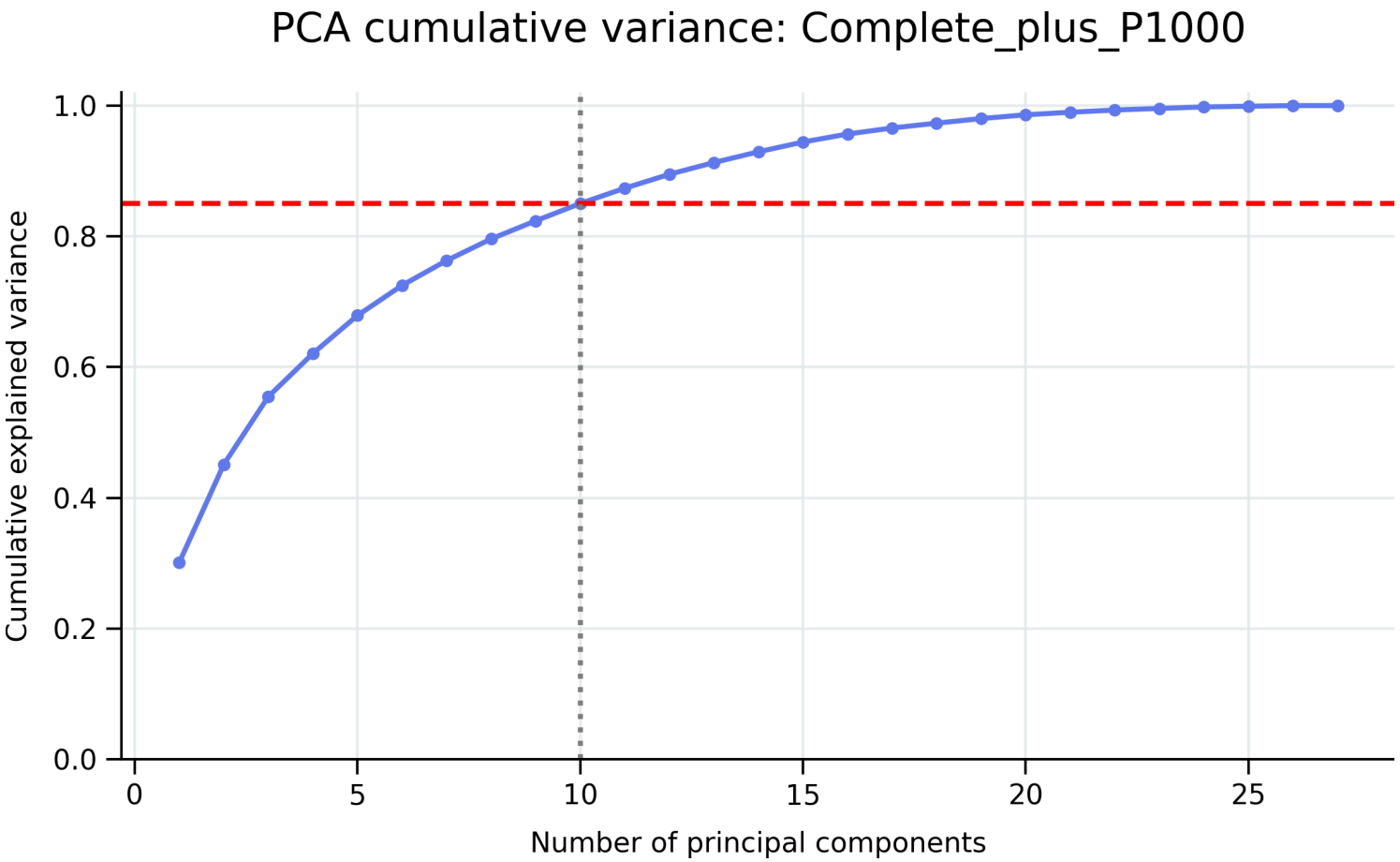


Figure 3

Cumulative proportion of standardized predictor variance explained by PCA for the complete-plus-P1000 set. The horizontal dashed line marks the prespecified 85% retention criterion, and the vertical dotted line identifies the 10 retained principal components, which together explained 85.02% of the variance.

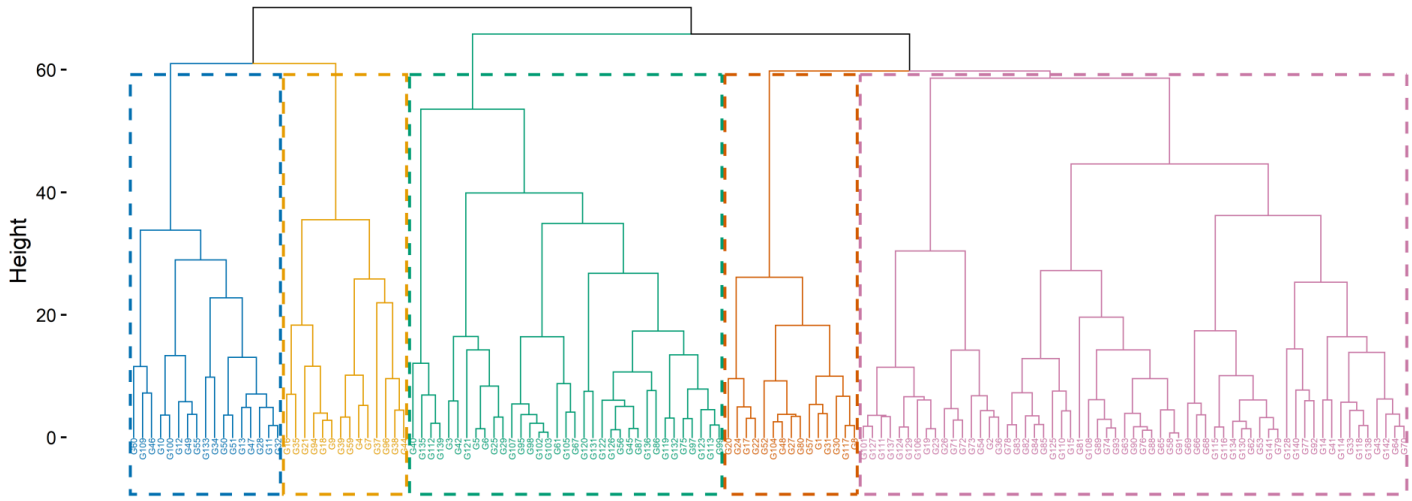


Figure 4

Hierarchical cluster analysis of 142 quinoa genotypes based on agro-morphological traits

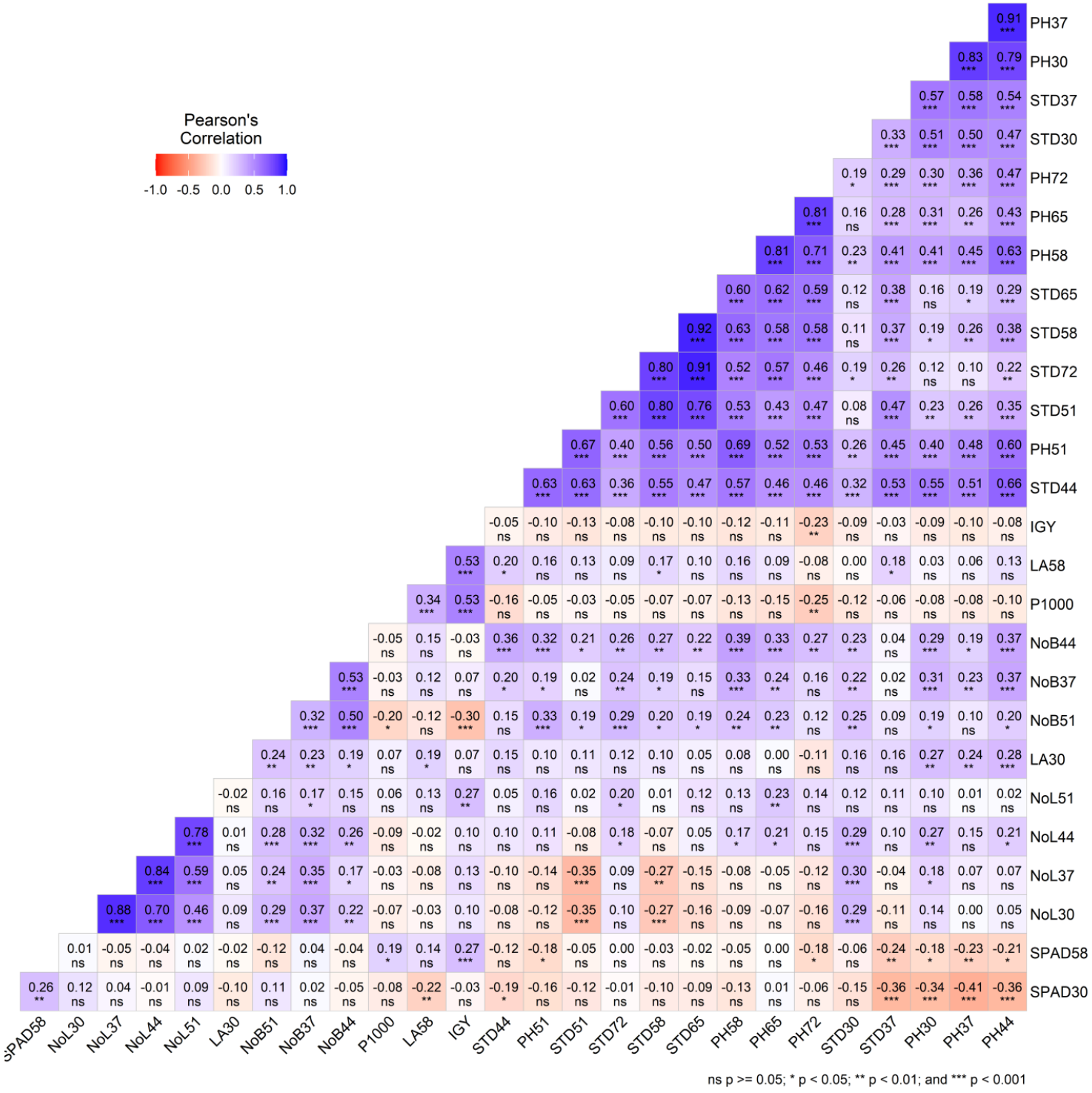


Figure 5

Pearson's correlation matrix of the evaluated traits in quinoa genotypes

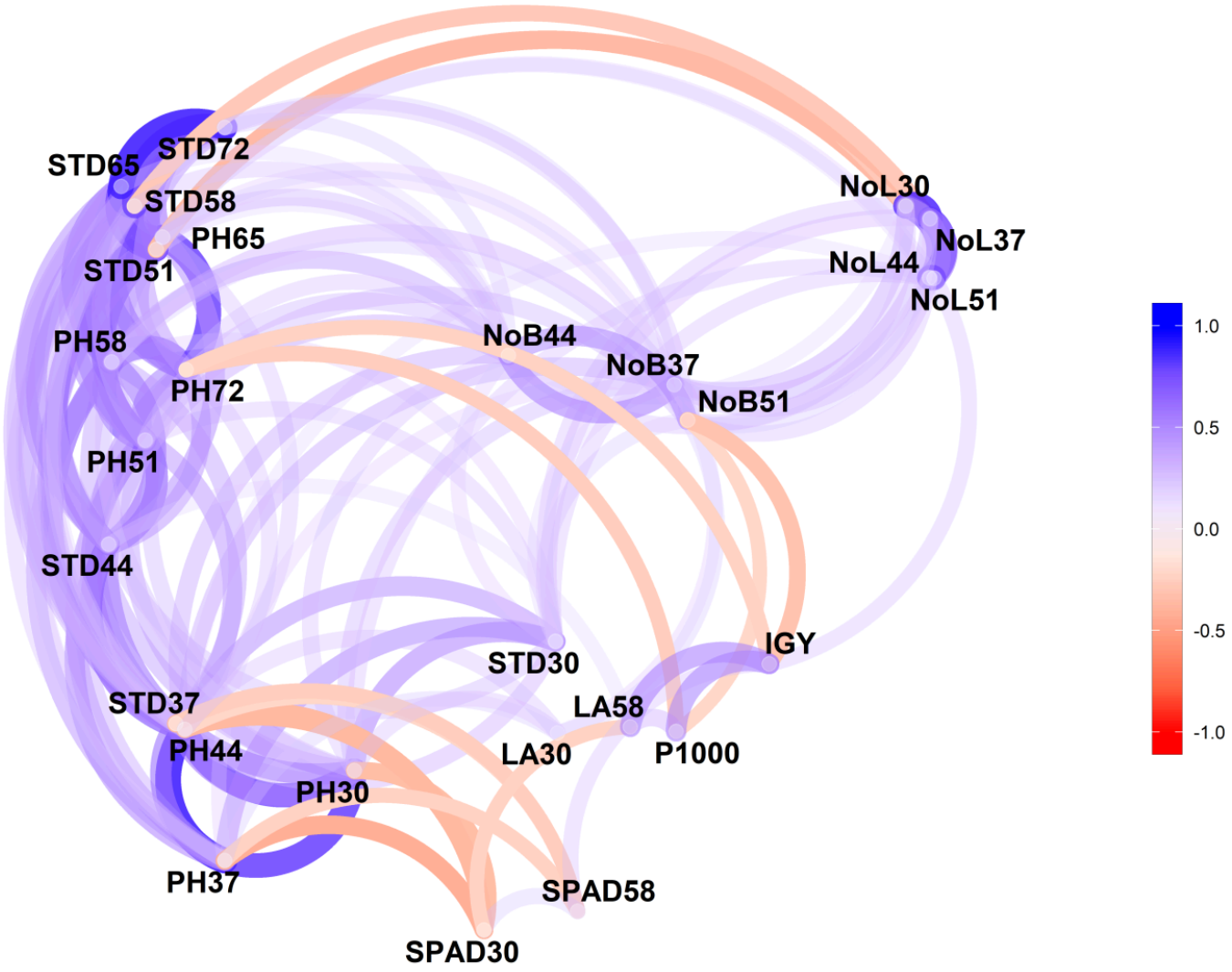


Figure 6

Network visualization of significant Pearson's correlations among the evaluated traits in quinoa genotypes.

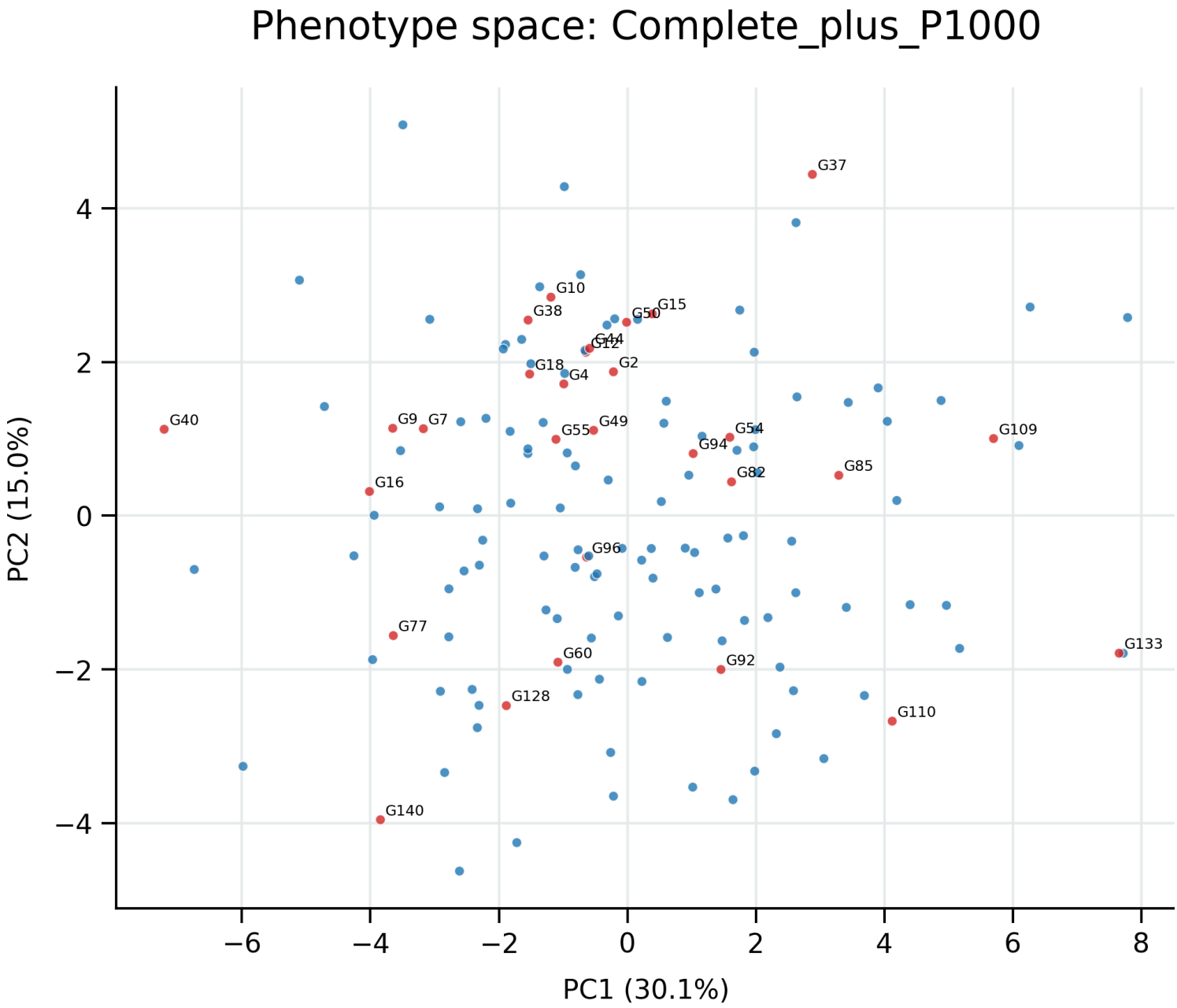


Figure 7

Projection of the 142 quinoa genotypes onto the first two principal components derived from the complete-plus-P1000 predictor set. PC1 and PC2 explained 30.1% and 15.0% of the standardized predictor variance, respectively. Red points represent the 29 observed high-yield genotypes, blue points represent the other 113 genotypes, and labels identify the high-yield genotypes. Classification used all 10 retained components rather than only the two displayed axes.

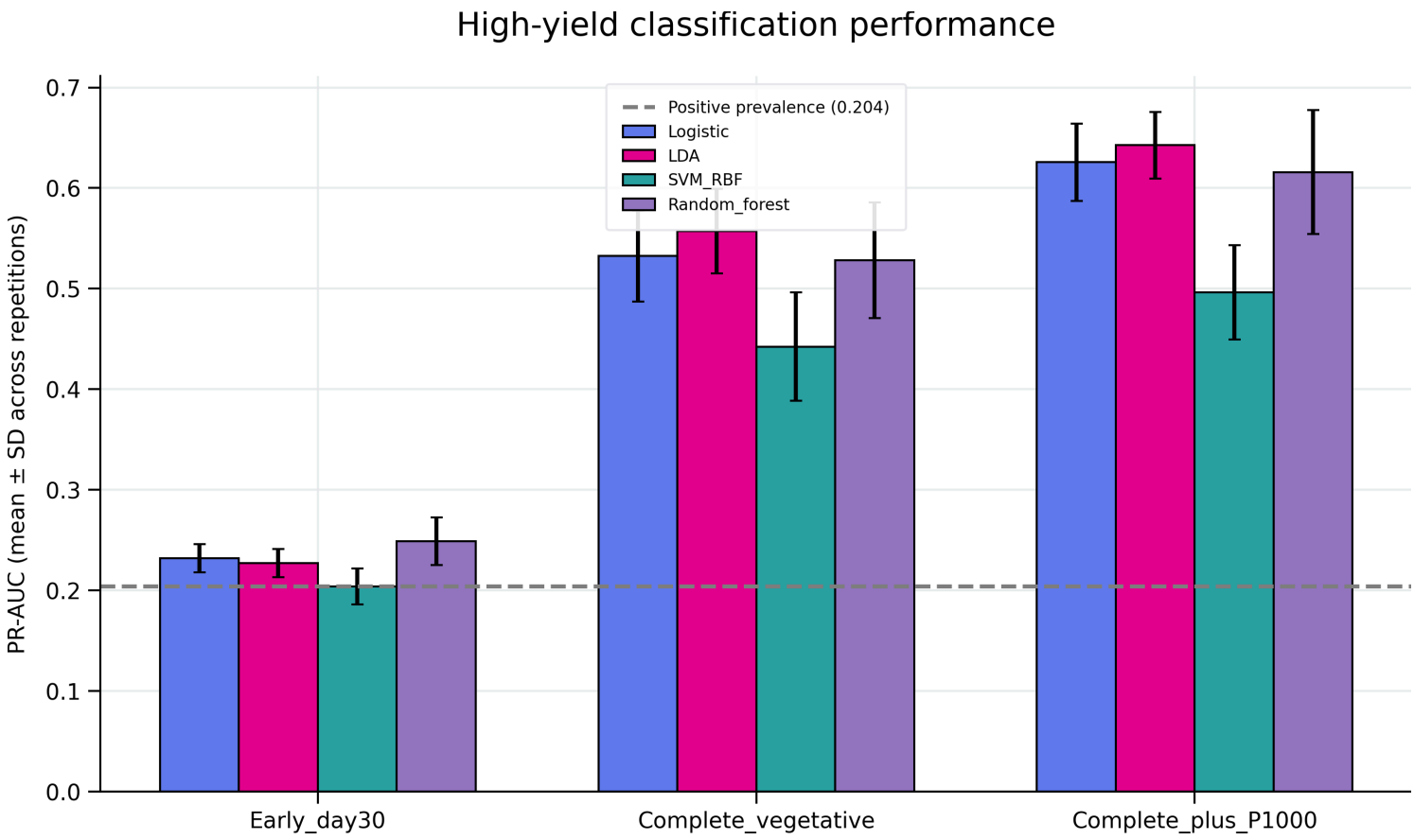


Figure 8

PR-AUC for each predictor-set and classifier combination, averaged across 10 repetitions of stratified five-fold cross-validation. Error bars represent the standard deviation across repetitions. The dashed horizontal line marks the high-yield prevalence (29/142 = 0.204), which provides the approximate PR-AUC expected from a non-informative ranking.

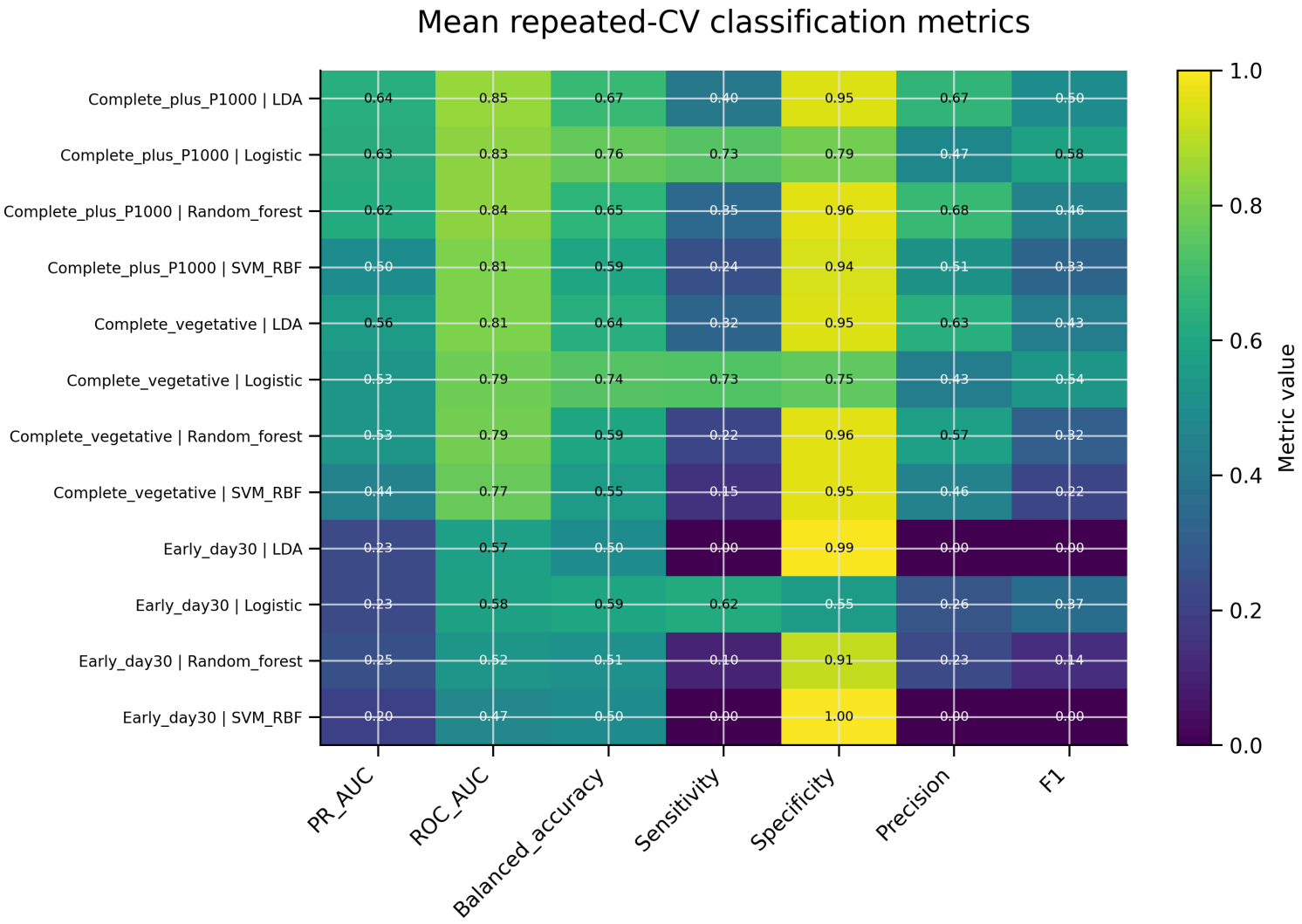


Figure 9

Mean classification metrics for the 12 predictor-set and classifier combinations across 10 repeated cross-validations. Each row represents one combination of predictor set and classifier, and each column represents one performance measure. Values closer to one indicate better performance, although the practical importance of sensitivity, specificity, and precision depends on the intended screening objective.

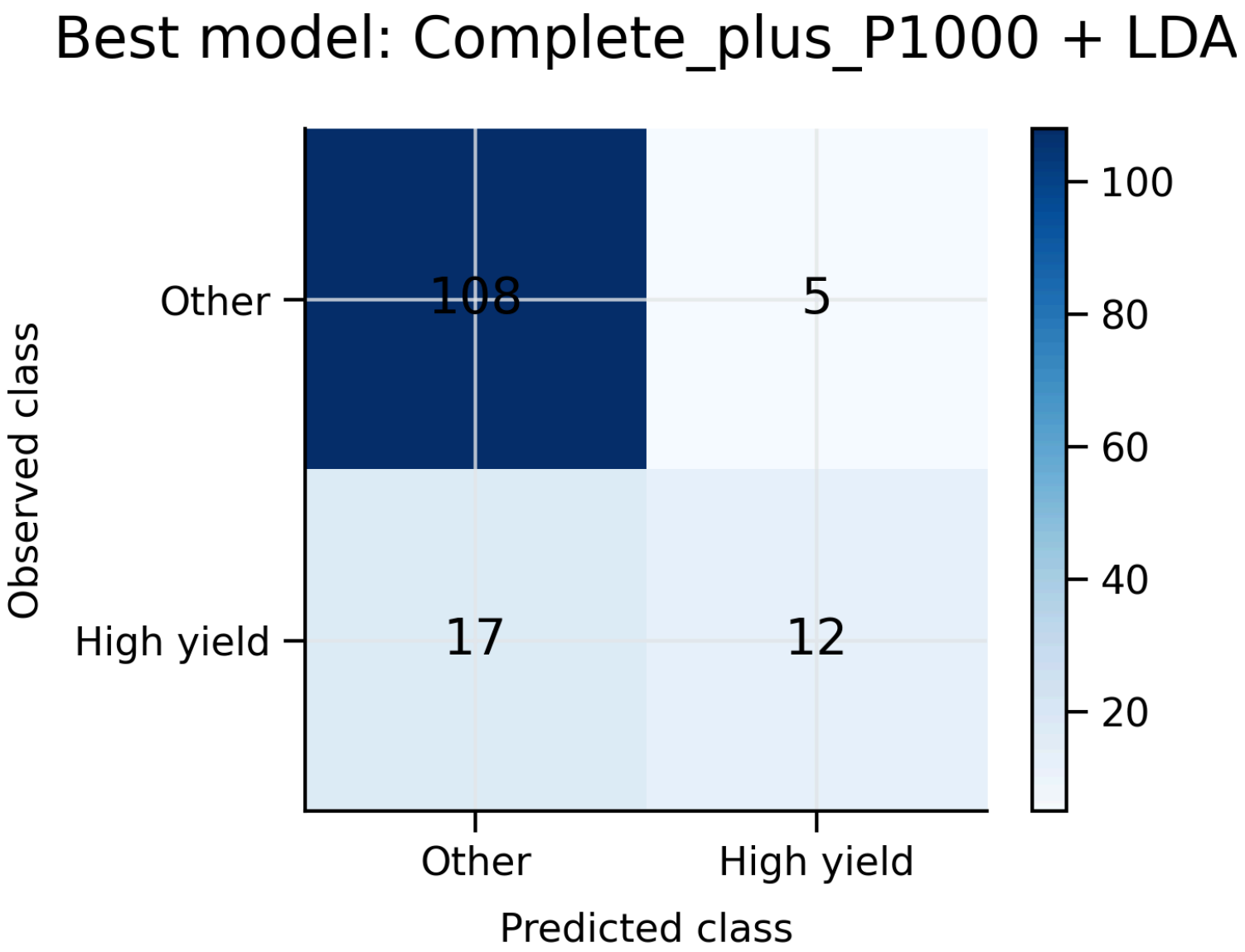


Figure 10

Aggregate confusion matrix for shrinkage LDA using the complete-plus-P1000 predictor set. We assigned each genotype to the high-yield class when its mean out-of-fold probability across the 10 cross-validation repetitions was at least 0.5. Rows indicate observed classes, and columns indicate predicted classes.

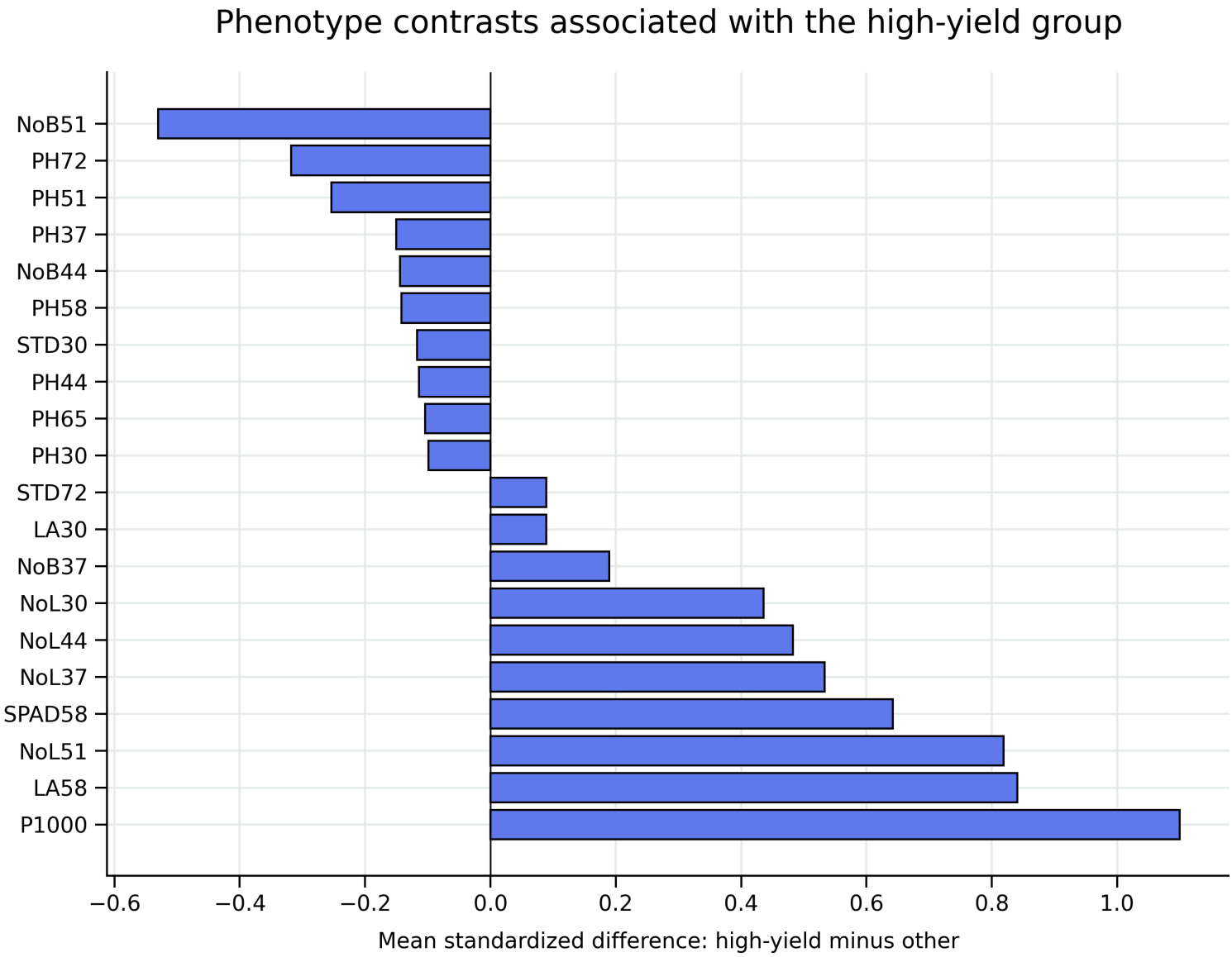


Figure 11

Differences in standardized phenotype means between the high-yield and other genotype groups. Each bar represents the high-yield-group mean minus the other-group mean in standard-deviation units. Positive values indicate higher means in the high-yield group, whereas negative values indicate lower means. The figure shows the 10 largest positive and 10 largest negative differences.

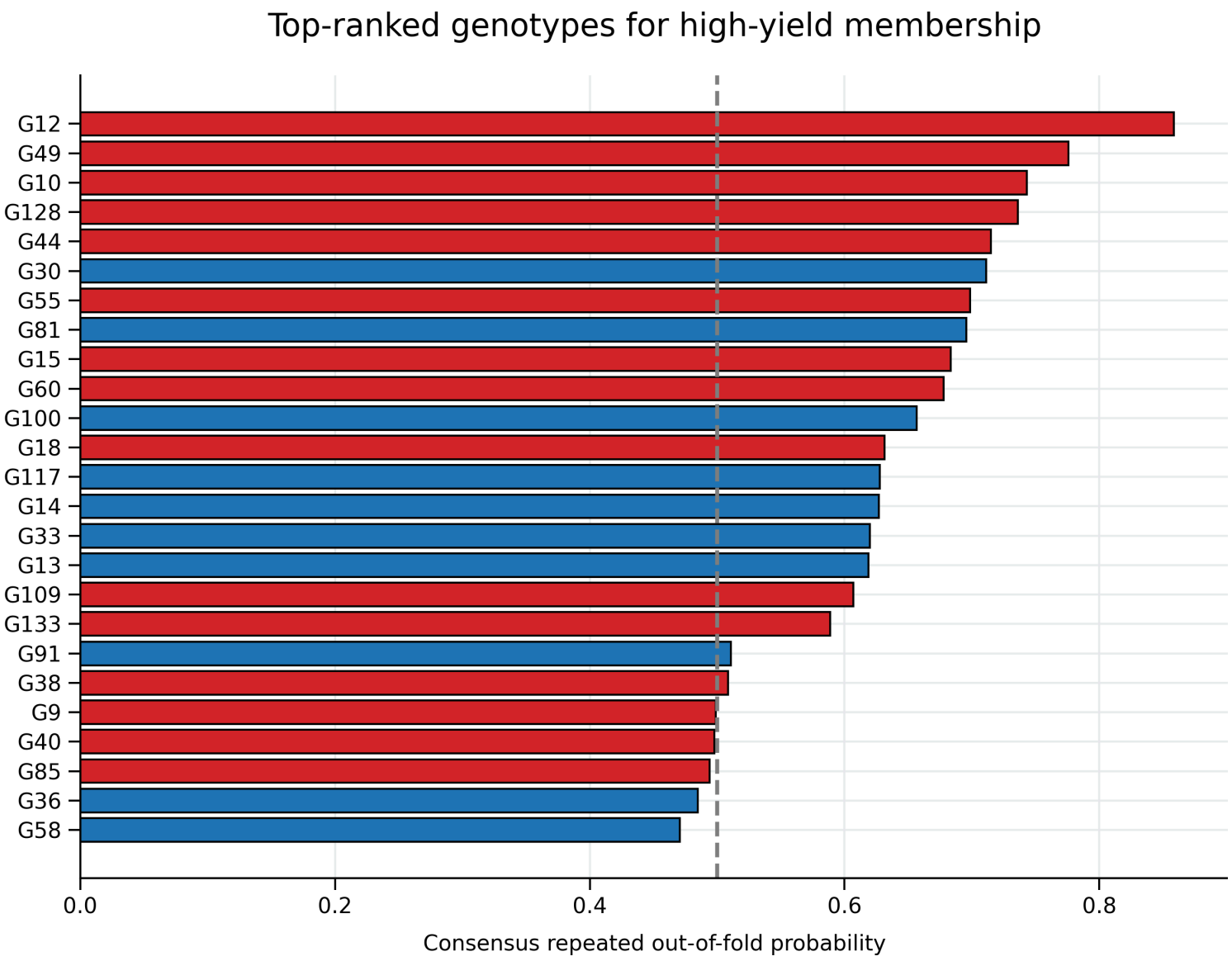


Figure 12

Top 25 genotypes ranked by the consensus score derived from the complete-plus-P1000 predictor set. The consensus score is the equally weighted mean of the four model-specific mean out-of-fold scores. Red bars indicate genotypes in the observed high-yield class, and blue bars indicate below-threshold genotypes. The dashed vertical line marks a score of 0.5 and is shown as a visual reference rather than as a validated decision threshold for the consensus ranking.

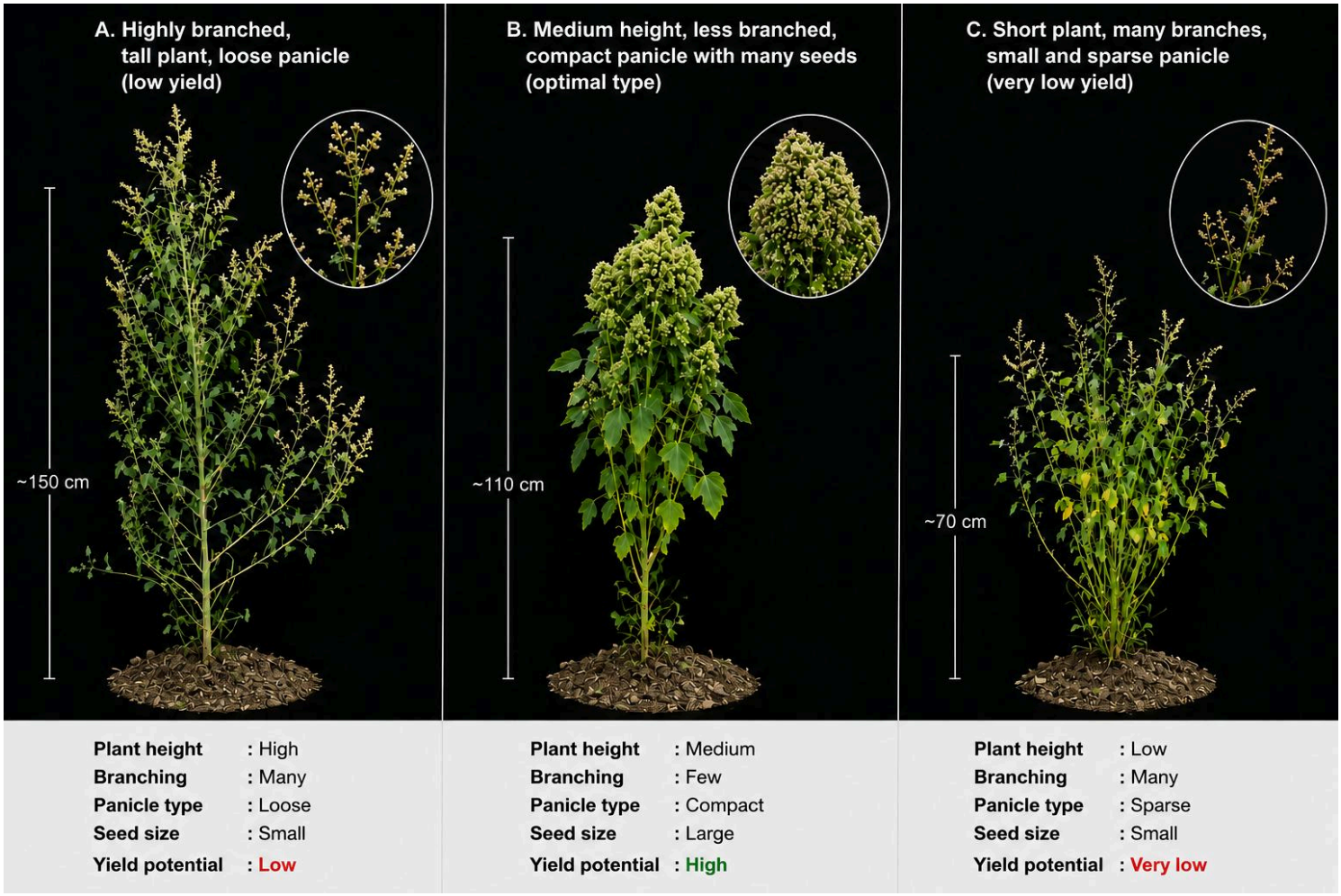


Figure 13

Representative quinoa plant architectures associated with contrasting yield potential