Predictive metabolomics of pearl millet phenotypic traits using a germplasm panel of genetic diversity

Mariana Pinheiro Costa Pimentel
Laboratory of Bioactives, Food and Nutrition Graduate Program (PPGAN), Federal University of the State of Rio de Janeiro (UNIRIO)

Alexandre Martins Abdão dos Passos
Embrapa Maize & Sorghum

Sylvain Prigent
Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS

Cédric Cassan
Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS

Flavio Dessaune Tardin
Embrapa Maize & Sorghum

Mariana Simões Larraz Ferreira
Laboratory of Bioactives, Food and Nutrition Graduate Program (PPGAN), Federal University of the State of Rio de Janeiro (UNIRIO)

Pierre Pétriacq
Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS

Millena Barros Santos

millena.barros-santos@inrae.fr

Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS

Research Article

Keywords: generalised linear modelling, machine learning, metabolites, phenolic compounds, Pennisetum glaucum.

Posted Date: May 31st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4450282/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

Introduction

Pearl millet, dubbed a “Nutri-cereal”, has a high content of protein, starch, fiber, mineral and fatty acids. Its resilience in adverse agro-climatic conditions sets it apart from major cereals. Despite this, understanding how its genetic diversity affects physiological traits and metabolic responses remains limited. Predictive metabolomics, merging metabolomics with artificial intelligence, allows for the comprehensive top-down modelling — from phenotype to the mechanism— of various phenotypic traits.

Objectives

To discover predictive biomarkers for phenotypic traits in the Brazilian germplasm core collection of 203 genotypes of pearl millet through the combination of predictive metabolomics with machine learning.

Methods

Untargeted metabolomics was conducted using UHPLC-LTQ-Orbitrap-HRMS to obtain metabolite profiles, from the central and specialised metabolism of the pearl millet core collection. Generalised linear modelling with penalisation (GLMNET) was applied to explore the correlation between metabolism and phenotypic traits.

Results

Our model successfully predicted eight qualitative traits from the pearl millet core collection, with accuracy ranging between 74% and 87%. From, 834 potential unique biomarkers (575 annotated-ion features and 259 unknowns) have been annotated as top metabolic predictors. It is noteworthy that the majority of the top metabolic predictors were from the carbohydrate, amino acid, flavonoid, and terpene subclasses.

Conclusions

This is the first report on leveraging a germplasm bank of pearl millet for metabolome characterisation and subsequent predictive modelling of important agronomic traits. These outcomes hint at the robustness of employing GLMNET for predicting metabolic biomarkers crucial in selecting genotypes for future breeding programmes.

1. INTRODUCTION

Millets emerged as an alternative to major cereal crops, they are grown in more than 93 countries worldwide, reaching a global production of around 32 million tonnes, where Asia produces 50% of the global production (FAOSTAT 2021). India alone is responsible for 39% of global production, followed by Niger (11%) and China (9%) (Bitzer et al., 2023). Pearl millet (Pennisetum glaucum (L.) R. Br.) contributes to around 50% of global millet production (Singh et al. 2017).
It is known as a gluten-free "Nutri-cereal" due to the high protein, carbohydrate, fibre, mineral, and fatty acids contents. The major specialised compounds of pearl millet grains were phenolic compounds such as $p$-coumaric acid, ferulic acid and gallic acids (Nani et al. 2015; N'Dri et al. 2012). Additionally, kaempferol glycoside, catechin, epicatechin, and procyanidin B1 and B2 are the main flavonoids (Hassan et al. 2020). All of it contributes to the antioxidant, anti-inflammatory and hypoglycemic properties (Annor et al. 2017; Dias-Martins et al. 2018). Moreover, this cereal is gluten-free, offering an alternative for individuals with celiac disease and other sensitivities triggered by gluten consumption (Hassan et al., 2021).

Beyond its nutritional value, pearl millet stands out as a resilient crop, displaying inherent resistance to environmental challenges such as low soil fertility, high temperatures, and limited rainfall (Varshney et al. 2017), emphasising its importance in addressing agricultural challenges and contributing to food security (Hassan et al. 2021). Despite this, very little is known about how the genetic diversity of pearl millet can impact physiological functions, specifically metabolic responses. One approach is to rely on a mini-core collection, representing 1% of the germplasm bank, to mimic the genetic diversity and assist improvement programmes (Upadhyaya et al. 2011).

Consequently, delving into the metabolism of pearl millet is expected to enhance yield understanding and advance breeding programs. The use of omic tools provides a holistic approach and valuable perspectives on the complex biochemical pathways that influence plant adaptability, as well as their effects on crucial agronomic traits and health-related benefits. Galactose metabolism and flavonoid biosynthesis were identified as predominant pathways from pearl millet grains genotype collection (197) in a metabolic genome-wide marker association study using high-resolution mass spectrometry (HRMS) (Yadav et al. 2021). These pathways play crucial roles, contributing to cereal L-ascorbic acid biosynthesis and the stress-responsive antioxidant system of plants (Allwood et al. 2021; Gill and Tuteja 2010; G.-Y. Zhang et al. 2015). In genotype breeding selection, HRMS of 54 health-beneting metabolites, pinpointed two top hybrids out of eleven, rich in slowly digestible and resistant starch, noted for their stability and performance across five environments (Yadav et al. 2022).

Predictive metabolomics brings together artificial intelligence, particularly machine learning, enabling comprehensive top-down modelling—from phenotype to the mechanism—of physiological traits (e.g. yield, quality, response to climatic/edaphic factors) based on metabolic data (Hajjar et al. 2023; Luna et al., 2020). Predictive metabolomics revealed the contributions of the plant metabolome to biotic stress resistance against pathogens in young plants after immune priming (Luna et al. 2020). In fruits, this strategy was further validated across eight species, highlighting the link between biomass composition and relative growth rate, revealing biomass components that drive fruit growth (Roch et al. 2020). This approach also effectively anticipated the plant climatic/edaphic environments when applied to 24 extremophiles from the Atacama desert, uncovering essential metabolites for resilience in extreme habitats (Dussarrat et al. 2022). Remarkably, redox and hormonal metabolites, implicated in trade-offs, played a significant role in this predictive framework (Decros et al. 2019; Dussarrat et al. 2021). In the cereal crop field, metabolite levels in the tyrosine pathway were effective predictors of yield when
analysing five hybrid rice varieties (Dan et al. 2021). One step further, predictive multi-omics successfully predicted 21 agronomic traits across 197 recombinant inbred lines of foxtail millet over three distinct years (Wei et al. 2023).

To gain a better understanding of pearl millet metabolism in an agricultural context, the present study seeks to explore the metabolic diversity in genotypes of the germplasm core collection from Embrapa, Brazil. To this end, we used UHPLC-HRMS to profile metabolites, including central and specialized compounds. Exploiting predictive metabolomics via machine learning of important phenotypic traits of pearl millet, the metabolic profiles were screened for predictive biomarkers that could help future breeding programmes.

2. MATERIAL AND METHODS

The Materials and Methods section is summarised in the text; a detailed version can be found in the Supplementary Methods.

2.1. Plant Material

Grains from the core collection of the pearl millet germplasm bank (203 genotypes) were provided by Embrapa Maize and Sorghum (Minas Gerais, Brazil). These genotypes originated from 14 different countries, including accessions from Icrisat, India, some African countries, the United States of America, and Brazil (Supplementary Table S1).

2.2. Metabolite Extraction

Lyophilized grains were grounded using a ball mixer for 4 min at 1500 rpm (Spex SamplePrep 2010, Geno/Grinder®, France) after adding three metal beads (Beads inox AISI 8 mm, CIMAP, France). Ground grains (10 mg d.w. \(^{-1}\)) underwent sequential ethanolic extraction three times using a robotised method (Luna et al. 2020, with some modifications).

2.3. Untargeted Metabolic Profiling of Pearl Millet Core Collection

Untargeted metabolic profiling was conducted by using an Ultimate 3000 ultra-high-pressure liquid chromatography (UHPLC) system within a GEMINI UHPLC C18 column (150 × 2 mm, 3 μm, Phenomenex, France) coupled to an LTQ-Orbitrap-HRMS, equipped with an electrospray ionization (ESI) source (ThermoScientific, Germany).

2.4. Generalised multilinear models

Supervised machine learning classes such as classification and regression were applied to the normalised data from MetaboAnalyst. Generalised linear modelling with ridge, Elastic-Net and LASSO penalisation (GLMNET) as described by Dussarrat et al. (2022) was used, with some modifications, to
explore the quantitative correlation between metabolism and the agronomic traits of the pearl millet core collection. The experimental workflow is summarized in Fig. 1.

3. RESULTS

3.1. Phenotypic traits of pearl millet

In plant breeding, comprehending the relationships among morphological traits is crucial for identifying potential phenotype criteria linked to yield and plant performance. To address this question, we conducted an initial hierarchical cluster analysis (HCA) involving all traits using Pearson's distance and Ward clustering (Fig. 2). The vertical dendrogram displays a distinct division into two primary clusters (Fig. 2a), where cluster I represents flowering range, stem internode length, leaf length, leaf width, origin, panicle height, panicle length, panicle thickness, stem thickness, and genetic type; and cluster II represents bristle length, days to flowering (DTF), grain colour, overall plant aspect, panicle shape, peduncle elongation, green fodder yield potential, number of leaves, total number of tillers, number of productive tillers, and 1,000 grains weight.

Pearson's correlation coefficient testing revealed the most prominent positive correlation between stem thickness and leaf width (0.723) (Fig. 2b). Significant positive correlations (0.50 ≤ r < 0.75) between growth parameters were found: stem internode length with plant height and leaf length; total number of tillers with number of productive tillers and number of leaves; leaf length with width, plant height, and panicle length; panicle length with leaf width and stem thickness; stem thickness with leaf length and panicle thickness. Green fodder yield potential, an indicator of biomass production, showed the strongest correlations with plant height (r = 0.490) and leaf length (r = 0.462). Although these correlations are lower than other metrics, they suggest that taller plants with longer leaves could produce more biomass and therefore be preferred for forage production.

In brief, while only weak correlations were found directly linked to 1,000 grains weight, strong positive correlations were found with other traits that indirectly affect grain yield, notably tiller aspects and panicle characteristics, suggesting other traits are positively interrelated and serve as indicators of plant performance, thereby influencing overall yield potential.

3.2. Overview of the metabolomic profiling

The exploration of the metabolic profile reflects its responses to both biotic and abiotic stimuli, genetic factors, and physiological status, thereby establishing a crucial link between genotypes and phenotypes (Gundaraniya et al. 2020). Consequently, we conducted a metabolomics approach using UHPLC-LTQ-Orbitrap-HRMS on 203 genotypes of pearl millet representing the core collection of germplasm.

After preprocessing, cleaning and normalising the data using MS-DIAL and MetaboAnalyst 5.0 online software, a dataset comprising 2,767 molecular features (rt, m/z) was obtained as described in section 2.4.
Principal Component Analysis (PCA) was then employed to have a global view of the genotypes and their metabolomes (Fig S1a). The first two components of the score plot explained 36% of the total dataset variance, with principal component 1 explaining 31.1% and principal component 2 explaining 4.9%. Taken all together, the analysis revealed that the metabolomes of this collection did not exhibit significant metabolic differentiation which may be related to the high intersect variability.

Furthermore, the results of the HCA heatmap built with the top 100 features ranked by ANOVA, from the pearl millet core collection's metabolome, failed to distinguish genotypes based on the relative ion abundance of metabolites (Fig S1b). Despite this, it is possible to see three different regions (see in Fig S1 b.; [I, II, III]) with dissimilar abundance of metabolites and, in the middle, region II highlighting 30 similar genotypes with a significant abundant profile. Given the overall findings, obtaining a clear distinction among the accessions was not possible. In light of this, directing the data to a supervised analysis, such as supervised learning, could allow us to find similarities between the traits as well as identify biomarkers.

### 3.3. Predictive modeling of agronomical traits from metabolome data

#### 3.3.1. Predictive performance

The utilization of predictive analysis through GLM with Ridge, Elastic-Net and LASSO penalization emerges as a pivotal strategy for predicting phenotypic traits based on the metabolome and highlighting key metabolites. Since the biochemical diversity oscillates with the genotype variety, GLMNET models (Friedman et al. 2010) were built using the 2,767 features. For each phenotypic trait, 50% of the dataset was used for the training of the model. Subsequently, the obtained model was applied to predict physiological values associated with the genotypes for the remaining 50% of the dataset, which served as a testing set.

#### 3.3.2. Classification

Arista/edge length emerged as the most predictable trait, achieving an average accuracy of 100 models of 89%. Interestingly, the qualitative traits resulted in an average accuracy ranging between 74% and 89% above the no-information rate of 44% and 82%. The $p$-value of 2.2E-16 indicates that the GLMNET model's accuracy is significantly superior to the no-information rate, implying that the model learned to predict better than a random one (Fig. 3a). These findings unveiled that the metabolome encompasses morphological variations in plants. Phenotypic traits exhibited distinct metabolic patterns wherein compounds correlated with the specific characteristics of the genotypes, enabling inference of traits such as panicle shape from which the grain was obtained.

#### 3.3.3. Regression

The quantitative traits that exhibited the most accurate predictions based on average $R^2$ over 100 repeats were peduncle elongation (0.66), leaf width (0.51), steam thickness (0.5), and plant height (0.5).
Overall, 1,000 grains weight (0.49), number of productive tillers (0.48), leaf length (0.44), days to flowering (0.42), number of leaves per plants (0.41), panicle thickness (0.39), number of total tillers (0.37), panicle length (0.35), and stem internode length (0.34) had a limited predictive reach (average $R^2 < 0.5$) (Fig. 3b). Unquestionably, the quantitative traits demonstrated a weak performance on both training and testing datasets. In other words, the metabolome has a low ability to predict phenotypic quantitative traits.

Although an average $R^2$ of 0.5 for some quantitative traits such as 1,000 grains weight can represent a good estimation in the practical of grain yield, we set a confidence threshold of $R^2$ or accuracy exceeding 0.7. With this criteria, only the dataset derived from the classification analysis proceeded to the subsequent step of exploiting the metabolic predictors.

### 3.4. Exploitation of metabolic predictors in pearl millet grains

#### 3.4.1. Selection, distribution and classification of the top metabolic predictors

Since we are using Elastic-Net and LASSO classification, the predictive models select specifically the metabolic features the most linked with the phenotype to predict. Hence, the top metabolic predictors were selected based on their occurrence in 100 models. When combining all models, 1,467 features (981 annotated-ion features and 486 unknowns) (Table 1) were selected with a threshold at 60% of occurrences in the models. There is a major difference when looking for the occurrence threshold of these variables in the specific traits. To reduce the risk of erroneously considering a variable as main predictor, we selected the top metabolic predictors with an occurrence above 80% (624 know-ion features and 258 unknowns), ensuring that only the most consistent and reliable biomarkers were included in the analysis.

To illustrate the distribution of the top metabolic predictors for the eight qualitative traits, we employed a Venn diagram for traits with six or fewer classes and an upset graph for traits with more than six classes. Our focus relied on pinpointing unique metabolic features that distinguish each trait category.
Concerning the bristle length, about 23 unique biomarkers predict short length, 19 for medium and 6 for long (Fig. 4a). The distribution of qualitative traits shows the occurrence of six panicle shapes (cylindrical, conical, spindle, club, candle, and lanceolate). When examining unique features, the candle shape exhibited the highest number (58), followed by lanceolate (37), cylindrical (29), club (14), conical (10), and spindle (10) (Fig. 4b). Pearl millet genotypes can be categorised by genetic type as variety, pollinator, line, compound, synthetic or silvestre. The compound type exhibited the highest number of unique features (50), followed by pollinator (45), silvestre (28), variety (24), line (11), and synthetic (10) (Fig. 4c). There were three flowering range classes: short, long, or discontinuous. Short flowering genotypes exhibit 52 exclusive features, while discontinuous flowering showed 42 and genotypes with long flowering periods 23 (Fig. 4d). The overall plant aspect had 52 exclusive features for the genotypes classified as good, 37 for the intermediate class, and 38 for poor class (Fig. 4e). The green fodder yield potential production did not revealed any unique feature (Fig. 4f).

Finally, the core collection of pearl millet grains was established through a series of donations from various germplasm bank collections, in addition to what already existed from Brazilian farmers. In this manner, the 18 locations of origin for the grains were ranked, including locations in Africa (unspecified countries), Brazil, Congo, Ghana, ICRISAT, India, Kenya, Mali, Namibia, Nigeria, Senegal, Serere-Uganda, Sudan, Tanzania, Uganda, the United States of America (USA), West Africa, and Zimbabwe.

Unique features were identified across all locations of origin. Specifically, 74 metabolic markers were exclusive to grains from the ICRISAT collection, 39 from India, 34 from Zimbabwe, 32 from the USA, 31 from Africa, 30 from Brazil, 21 from Mali, 20 from Nigeria, 16 from Sudan, 13 from Ghana, 13 from Kenya, 12 from Namibia, 11 from Serere-Uganda, 10 from Senegal, 10 from Uganda, 8 from West Africa, 8 from Congo, and 8 from Tanzania (Fig. 4g). The grains were categorized into seven colour classes. Among these, genotypes with gray grain colour showed the highest number of unique metabolic markers (52), followed by brown-gray (42), a mixture of yellow and gray grains (29), deep gray (17), yellow (11), brown (10), and cream (7) (Fig. 4h).

3.4.2. Annotation of unique top metabolic predictors/markers

To gain further insights into the metabolic aspects related to the agronomic predictors, we performed metabolic annotation. Our dataset consisted of 834 unique biomarkers (575 annotated-ion features and 259 unknowns), of which 42 compounds were identified as level 1, 116 were annotated as level 2, 417 were annotated as level 3 with defined molecular formula and class indicative and 259 unknowns as level 4. Moreover, all annotated-ion features could be confirmed by the class or subclass level. The MSI annotation level for each marker is presented in Supplementary Table S4.

It is noteworthy that the majority of the unique top metabolic predictors were involved in specialised metabolism (58%), while 42% represented central metabolism. Among the identified level 1 metabolites from specialised metabolism are vanillic acid, caffeic acid, ferulic acid, xanthine, theophylline, diosmetin, quercitrin, naringenin, biochanin A, neochlorogenic acid, (-)-epicatechin, petroselinate and
phenylpropionylglycine. The subclass that gathered the highest number of top metabolic markers alone was amino acids, peptides, and analogues (n = 57), and carbohydrates and carbohydrate conjugates (n = 36). Therefore, we will explore the other subclasses that stood out among the biomarkers considered unique to each agronomic trait. All of this information is represented in Fig. 5.

The most significant unique predictors, comprising 23% of the bristle length subclasses, were associated with lipids fatty acids and conjugates, linoleic acids and derivatives, fatty acyl glycosides, tricarboxylic acids and derivates, ethers and hydroxysteroids. Phenolic compounds had also a clear importance encompassing 20% of the bristle subclasses, including flavonoid glycosides, benzoic acids and derivatives, hydrolysable tannins, methoxybenzenes and stilbene glycosides. Additionally, the terpene pathways, which include diterpenoids, sesquiterpenoids, terpene glycosides, and terpene lactones, were also highlighted as valuable, accounting for 16% of the predictors.

Phenolic compounds were the best unique predictors for panicle shape, being responsible for 25% of subclasses. This group includes benzoic acids and derivatives, hydroxycinnamic acids and derivatives, flavones, flavonoid glycosides, O-methylated isoflavonoids, isoflavonoid O-glycosides, isoflavonoid C-glycosides, hydrolysable tannins, cinnamic acid esters, methoxyphenols, coumarin glycosides. Additionally, the terpenes accomplished 10% of the total unique biomarkers. Within genetic type trait, phenolic compounds were distributed across 12 subclasses (including benzoic acids and derivatives, hydroxycinnamic acids and derivatives, flavones, flavans, flavonoid glycosides, O-methylated flavonoids, isoflavonoid O-glycosides, biflavonoids and polyflavonoids, hydrolysable tannins, tetrahydrofuran lignans, methoxybenzenes, and biphenols), accounting for 26%.

The flowering range trait was distinguished by a high proportion of the unique markers referred to as lipids (23%), encompassing fatty acids and conjugates, linoleic acids and derivatives, eicosanoids, steroid lactones, glycerophosphoglycerols, and hydroxysteroids. Additionally, phenolic compounds were 21% derived from benzoic acids and derivatives, hydroxycinnamic acids and derivatives, flavans, flavonoid glycosides, O-methylated flavonoids, isoflavonoid O-glycosides, O-methylated isoavonoids, cinnamic acids, methoxybenzenes and biphenols.

Within the overall plant aspect, the distribution of unique biomarkers showed 18% belonging to the lipids and fatty acids group (fatty acids and conjugates, fatty acyl glycoside and linoleic acids and derivatives), followed by 14% of phenolic compounds (hydroxycinnamic acids and derivatives, cinnamic acid esters, benzoic acids and derivatives, flavonoid glycosides, isoflav-3-enes, O-methylated isoflavonoids) and 7% of terpenes (terpene lactones, terpene glycosides and diterpenoids). This pattern was also observed in green fodder yield potential trait, with 18% of the unique biomarkers from lipids and fatty acids group (fatty acyl glycosides, fatty acids and conjugates and fatty alcohols), 14% phenolic compounds (flavonoid glycosides, isoflavonoid O-glycosides, O-methylated isoflavonoids, cinnamic acids, methoxybenzenes and biphenols) and 7% terpenes (terpene lactones, terpene glycosides and diterpenoids).

Concerning the origin, phenolic compounds accounted for a significant portion of the unique biomarkers (24%). They were distributed across eight subclasses: cinnamic acid esters, hydroxycinnamic acids and
derivatives, benzoic acids and derivatives, flavans, flavones, flavonoid glycosides, O-methylated flavonoids and isoflavonoid O-glycosides, coumarin glycosides, bioflavonoids and polyflavonoids, stilbene glycosides, isoflavonoid C-glycosides, neoflavones, hydroxycoumarins.

In grain colour, mostly part of them was represented by phenolic compounds 30% (benzoic acids and derivatives, hydroxycinnamic acids and derivatives, flavans, flavonoid glycosides, biflavonoids and polyflavonoids, O-methylated flavonoids, O-methylated isoflavonoids, isoflavonoid O-glycosides, hydrolyzable tannins, methoxyphenols, pyranocoumarins, and biphenols)

4. DISCUSSION

4.1. Exploring yield: Unveiling the strong connection beyond grain attributes

The assessment of phenotypic traits in germplasm is recognized as a crucial stage in identifying genotypes that can thrive in various environments and possess desirable agronomic characteristics and performance (Sood et al. 2015). Notably, certain traits exhibit high heritability (> 80%), including panicle length, number of productive tillers per plant, plant height, 1000 grains weight, and green fodder yield per plant, while days to 50% flowering are considered a phenological trait in pearl millet (Kaushik et al. 2018). In this study, we evaluate the correlation of 21 phenotypic traits of 203 pearl millet accessions (Fig. 1), including 13 quantitative traits, some of them directly related to grain yield, such as 1000 grains weight. The major quantitative trait positively correlated with 1000 grains weight showing a correlation coefficient > 0.3 was panicle thickness (0.37). Corroborating this study, Zhang et al. (2024) examined 21 agronomic traits, including 7 grain yield-related traits across over 1200 foxtail millet varieties and showed agronomic traits weakly correlated with 1000 grains weight (correlation coefficient ≤ 0.2) such as plant height and stem node number. Moreover, this result can be related to the heterogeneity of collected data (location, year, planting time etc) since pearl millet is sensitive to photoperiod, impacting traits such as plant height, days to flowering, biomass/grain yield potential. In this case, phenotypic differences observed between materials could be more strongly influenced by environmental factors rather than genetic ones. However, positive correlations were observed with leaf features and plant height with panicle length. Leaf attributes play a central role in governing light interception and photosynthetic capacity, which are critical for efficiently partitioning photoassimilated carbon and impacting grain development (Horton 2000). Thus, the productivity of agriculture is intimately connected to the growth and development of plants.

4.2. Insights into plant phenotypic plasticity: a comprehensive exploration of metabolomics and predictive modeling

Metabolomics plays a pivotal role in comprehending the phenotypic plasticity of plants, offering unique insights into their characteristics (Weckwerth et al. 2020). Metabolite biomarkers serve as distinct
chemical signatures for trait phenotypes, making them valuable tools for precision breeding. However, the PCA overview did not reveal a clear differentiation between genotypes based on the metabolome. Similar findings were observed in the PMiGAP collection (197 pearl millet accessions), where the metabolome was primarily associated with galactose metabolism, flavonoid biosynthesis, and amino sugar and nucleotide sugar metabolism (Yadav et al. 2021).

Predictive modelling using machine learning technology is increasing towards early predictions before harvest, predictions at the scale of field or region, and predictions for different types of crops. This great diversity of prediction tasks requires a proper choice of specific machine learning techniques to attain high levels of performance (Leukel et al. 2023). The significance of predicting yield and other phenotypic traits has already been emphasised. From 210 rice hybrid lines, it is evident that metabolomics is more effective in predicting yield than genomic data (Xu et al. 2016). This could be attributed to metabolites serving as intermediate phenotypes. Notably, a study utilising the UHPLC-QqQ-MS/MS metabolomics approach, which compared the metabolomes of whole grains from proso millet varieties, revealed significant enrichment in pathways such as flavonoid biosynthesis, isoflavonoid biosynthesis, and flavone and flavonol biosynthesis (Li et al. 2021). Here, the combination of UHPLC-HRMS and GLM approach reveals the relationship among specialised metabolites, mainly flavonoids and terpenes, in predicting qualitative traits.

The top specialised makers associated with grain colour include compounds from the benzoic acids and derivatives, hydroxycinnamic acids and derivatives, flavonoids, hydrolysable tannins, methoxyphenols, and biphenols subclasses. Pearl millet displays a diverse spectrum of pericarp colours in its grains, ranging from ivory, cream, yellow, gray, brown, and a mixture of yellow and gray. These colour variations are associated with the presence of specialised metabolites, which play a crucial role in influencing the colour of both flowers and seeds (Tanaka et al. 2008). Zhou et al. (2020) assessed 217 metabolites that exhibited significant differences concerning the grain colour of white, red, and purple sweet sorghum (*Sorghum bicolor*), with 199 metabolites specifically associated with the red phenotype. The predictive capability of grain colour could also serve as an indicator of the antioxidant and nutritive properties of these grains. Besides this, specialized metabolites could also be evaluated to predict the testa thickness. For tannin-rich sorghum, for instance, varieties with thicker testa might have higher tannin content than those darker-pigmented, thinner testa (Zhang et al. 2023); the levels of phenolics and antioxidant activity are associated with thick and colored pericarp, and a pigmented testa related to S gene (Dykes et al. 2005).

In our model, quantitative traits exhibited an $R^2$ lower than 0.7 and did not proceed to the biomarker annotation selection. Indeed, quantitative traits are usually controlled by numerous genes and have a greater influence on the environment when compared to qualitative traits, which are often controlled by one or a few genes. The environmental effect associated with quantitative data may explain the difficulty of adjusting the prediction model. In this context, the use of the BLUP method (rrBLUP) for predicting agronomic traits like plant height, stem height, and 1000 grain weight in 197 recombinant inbred lines of foxtail millet outperformed the LASSO methods (glmnet). Furthermore, integrating multi-omics data
provided more accurate results than solely using genomic or metabolomic data. This way, incorporating genomic information in our dataset and testing different methods to determine the best fit could prove to be an effective strategy (Wei et al. 2023).

5. CONCLUSION

To the best of our knowledge, the present study represents the first attempt to leverage a germplasm bank of pearl millet for metabolome characterisation and subsequent predictive modelling of important phenotypic traits. Given pearl millet resilience and its potential as a climate change-resistant cereal, this research holds significant implications for various fields including agronomy, nutrition, and health. Focusing on leaf characteristics and general plant aspects offers a valuable perspective on agronomic yield. Although unsupervised analysis failed to differentiate phenotypic traits from the metabolome, supervised machine learning proved to be effective. Employing GLMNET, we successfully predicted eight qualitative phenotypic traits from the pearl millet metabolome. Notably, the top metabolite predictors were from the carbohydrate, amino acid, flavonoid and terpene subclasses. Further investigations in predicting quantitative traits could be deeply explored by working with a multi-omics dataset that includes both metabolic and genomic information. Testing different predictive models would also be crucial in determining the most suitable approach. Moreover, creating a ‘validation’ set to perform real predictions based on the selected biomarkers and genotypes grown under controlled conditions will be instrumental in advancing future breeding programs.

Declarations

ACKNOWLEDGMENTS AND FUNDING

The authors express their gratitude to Dr. Carlos Wanderlei Piler de Carvalho from Embrapa - Brazilian Agricultural Research Corporation for their support in obtaining the pearl millet grains used in the study; to the Agromix-MetaboHUB for providing the customized plant databank used in this work to annotate metabolite; and to Luciana Ribeiro da Silva Lima from the Graduate Program in Food and Nutrition (PPGAN/UNIRIO) for essential teachings in annotation process.

The authors are grateful for financial support from Bordeaux Metabolome Facility, MetaboHUB (ANR-11-INBS-0010) and PHENOME (ANR-11-INBS-0012); Federal University of the State of Rio de Janeiro (UNIRIO); Foundation for Research Support of the State of Rio de Janeiro (FAPERJ) (26/210.008/2024; 26/202.598/2023; 26/201.623/2023; 26/210.152/2023; 26/201.317/2022), National Council for Scientific and Technological Development (CNPq) (407350/2023-3; 314100/2023-7), Coordination for Improvement of Personnel with Higher Education (CAPES) (financial code 001).

AUTHORS CONTRIBUTION

MSLF, PP and MCBS designed and planned the project. AMAP provided the samples and agronomic data. MPCP, CC and MCBS performed experimental analyses. MPCP, SP, PP and MCBS conducted
metabolic or bioinformatic experiments and analyses. MPCP, PP and MCBS, integrated and analysed the experiments results. AMAP and FDT review and agronomic interpretations of results. MPCP writing – original draft. MPCP, MSLF, PP and MCBS writing – review and editing. MSLF, PP and MCBS, supervised the project. MSLF, PP and MCBS funding acquisition. All authors read and approved the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The metabolomics and metadata reported in this paper are available via https://doi.org/10.57745/GU6WDG.

References


Figure 1

Experimental workflow of predictive modelling.
Figure 2

Metadata analysis of quantitative traits. a. Heatmap result of hierarchical cluster analysis (Pearson’s) correlation with Ward clustering applied to 21 phenotypic traits. Characteristics with relatively low and high levels are displayed in light blue and red, respectively. b. Correlogram for phenotypic traits. Red indicates positive correlation and blue negative correlation. Colour intensity is proportional to the correlation coefficient (r).
Figure 3

Prediction of predictive traits from metabolic markers. **a.** Accuracy correlation between predicted and measured qualitative traits based on machine learning. **b.** $R^2$ score obtained by predicting quantitative traits using RT-m/z features based on GLMNET models.
Figure 4

Venn diagram and upset graphs illustrating the predicted RT-\(m/z\) features for each phenotypic trait:  

- **a** Bristle length  
- **b** Flowering range  
- **c** Overall plant aspect  
- **d** Green fodder yield potential  
- **e** Panicle shapes  
- **f** Genetic type  
- **g** Grain colour  
- **h** Origin.
Figure 5

Sunburst plot of the 575 unique metabolic features. The inner ring represents the class, followed by subclass.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFileS1.txt
- SupplementaryTables.xlsx
- SupplementaryMethodsBAG.docx
- Supplementaryfigure.docx