BMC Genomics

RESEARCH

Open Access

Genetic architecture of inter-specific and -generic grass hybrids by network analysis on multi-omics data

Elesandro Bornhofen[1*], Dario Fè[2], Istvan Nagy[3], Ingo Lenk[2], Morten Greve[2], Thomas Didion[2], Christian S. Jensen[2], Torben Asp[3] and Luc Janss[1*]

Abstract

Background Understanding the mechanisms underlining forage production and its biomass nutritive quality at the omics level is crucial for boosting the output of high-quality dry matter per unit of land. Despite the advent of multiple omics integration for the study of biological systems in major crops, investigations on forage species are still scarce.

Results Our results identified substantial changes in gene co-expression and metabolite-metabolite network topologies as a result of genetic perturbation by hybridizing *L. perenne* with another species within the genus (*L. multiflorum*) relative to across genera (*F. pratensis*). However, conserved hub genes and hub metabolomic features were detected between pedigree classes, some of which were highly heritable and displayed one or more significant edges with agronomic traits in a weighted omics-phenotype network. In spite of tagging relevant biological molecules as, for example, the light-induced rice 1 (*LIR1*), hub features were not necessarily better explanatory variables for omics-assisted prediction than features stochastically sampled and all available regressors.

Conclusions The utilization of computational techniques for the reconstruction of co-expression networks facilitates the identification of key omic features that serve as central nodes and demonstrate correlation with the manifestation of observed traits. Our results also indicate a robust association between early multi-omic traits measured in a greenhouse setting and phenotypic traits evaluated under field conditions.

Keywords Graphical lasso, Metabolome, Multi-trait mixed model, Network science, Polyploid, Transcriptome

Background

Forage grasses cover large portions of agricultural land worldwide, efficiently converting enormous amounts of natural resources into macronutrients used primarily for feed. Their relevance can be recognized by the extent of the network of researchers and breeding organizations devoted to maximizing production efficiency. This has been largely achieved by conventional breeding techniques aiming to explore genetic variation not only within but also across species and genera over the last decades. As biotechnology surged, breeders advanced in experimenting with hybridizations across species and genera, leading to the release of successful varieties of polyploid

*Correspondence:
Elesandro Bornhofen
bornhofen@qgg.au.dk
Luc Janss
luc.janss@qgg.au.dk
[1] Center for Quantitative Genetics and Genomics, Aarhus University, Aarhus, Denmark
[2] Research Division, DLF Seeds A/S, Store Heddinge, Denmark
[3] Center for Quantitative Genetics and Genomics, Aarhus University, Slagelse, Denmark

hybrid ryegrass (*L. perenne* × *L. multiflorum*) and *Festulolium loliaceum* (*L. perenne* × *F. pratensis*), for example. As high-throughput sequencing platforms reduced genotyping costs, genomics began to play a significant role across grass breeding programs, reshaping breeding pipelines aiming at the optimization of resource allocation mainly via genome-wide selection [1–3]. Recently, the complex problem of predicting phenotypes and finding candidate biological molecules associated with it can also be supported not only by marker information at the DNA level but also via transcriptomics [4] and metabolomics [5], leading to a holistic view of the phenomena controlling the expression of economically important traits.

Improving existing weaknesses of elite genetic materials or simply unlocking genetic variability for breeding exploitation are processes that benefited by leveraging hybridization across genera and species within a genus. In spite of being predominantly diploid ($2n = 2x = 14$), *L. perenne*, *L. multiflorum*, and *F. pratensis* can also be found as or induced to tetraploid states, which is essential for amphidiploid production. However, genomic instability is often reported and a shift to the ryegrass genome over generations can happen in crosses with fescues [6, 7]. Additionally, homeolog expression bias and expression level dominance can be observed in such allopolyploids [8]. Collectively, these phenomena may lead to distinct interactomes when hybridizations are performed across species, which can be analyzed through network reconstruction by leveraging high throughput omics data and appropriate statistical methods. For example, network topologies revealed allopolyplid cotton resembling more to one of the diploid species representing a progenitor besides a substantial domestication impact on the coexpression [9]. Additional studies on expression modifications in allopolyploids remain scarce.

Adding extra layers of biological information also means increasing data dimensionality ($n \ll p$ problem). Reliable inferences in high dimensions require specific statistical procedures and an in-depth understanding of the underlining phenomena. Among the methods proposed for the analysis of high dimensional omics data [10], the reconstruction and analysis of regulatory networks offer the possibility to prioritize omic features [11], significantly reducing the searching space for downstream analyses. Organizing omic features in interacting networks can be seen as an approximation of the true existent interconnected biological system that reads the information encoded on the genome and ends with a functional organism. Reconstructed networks hold biological meaningful topological properties, for example, the presence of modules that might cluster nodes (omic features) performing specific biological functions [12] and the existence of highly connected nodes. These hub nodes arise as biological networks are assumed to be scale-free, meaning that node degrees are power-law distributed [13] and, therefore, few highly connected nodes are expected. The presence of these disproportionally connected hub nodes is an important topological property of networks as it may represent key genes/metabolites associated with biological pathways. Thus, it would be of special interest to investigate the extent to which hub omic features can be significantly linked to biomass yield and other economically important phenotypes of fodder grasses. Researchers have found hub genes affecting biomass accumulation in other families of plants, for example, in *Ulmus pumila* L. [14] and *Arabidopsis thaliana* [15]. That being stated, one needs to first estimate the network to be able to explore its topological properties and this can be accomplished by leveraging graph theory and probability for modeling and representation of complex biological problems as probabilistic graphical models [12].

Omics data as a graphical model is based on the estimation of conditionally independent relationships across random variables in a multivariate setting. Learning a graphic in high-dimension requires dealing with a situation where the number of unknown parameters exceeds the sample size. In this case, ℓ_1-penalization has been one of the main techniques used to make sparse inference in a Gaussian Markov random field [16], yielding a sparse structured precision which, in turn, can be converted into an undirected network and further analyzed for its topological properties. This approach has been applied to the study of gene expression [17, 18] and metabolomic [19] data in humans, with few examples in plants [20–23]. With a selected set of candidate features recovered from gene co-expression and metabolic networks, one can perform omic-phenotype integration. The simple correlation-based integration method of omic variables and phenotypes is widely used, with examples in maize [24] and for the forage species *E. sibiricus* [25]. However, more robust approaches based on multivariate multi-level models have also been applied [26, 27], showing better properties [28]. Finding significant associations of genes and metabolites with dry matter yield and nutritive quality traits in fooder grasses could reveal potential targets for quantitative trait dissection studies, improve the omics-assisted selection of elite families, and shed light on regulatory processes of key traits. Additionally, given the fact that a large part of the above-ground biomass is harvested in forage grasses, it can be hypothesized that randomly sampled hub features are more likely to be linked to a phenotype of interest compared to, for example, grain crops.

The inherent properties of an organism's interactome, especially the power-law distribution of interactions, give plasticity in face of random disturbances. However, interferences on hub nodes may lead to severe product alterations [29], making them targets for genetic studies. Additionally, hub genes appear to be associated with a variety of biological processes [9, 25, 30, 31] and had been mentioned as potential targets for the molecular breeding of forage species [32]. In the present study, we consider the problem of reconstructing the interplay among biomolecules represented by omic data from families of grass hybrids of two pedigree classes and narrowing down the high-dimensional space to fewer conserved hub features between them. By using a sparse estimation technique via undirected graphical models, we filter the relevant omic features and test the conserved hubs for their genetic association with quantitative traits and potential for prediction. The identification of significantly associated hubs will confirm the relevance of these interacting biomolecules, providing insights into potential molecular biology studies and marker-assisted breeding.

Results

Genetic similarity among family pools and omics heritability

We constructed a genomic relationship matrix (GRM, see Eq. 1) for the *L. perenne* × *L. multiflorum* (hybrid ryegrass; HR pedigree class) using 85,283 SNPs and a GRM for the intergeneric crosses of *L. perenne* × *F. pratensis* (*Festulolium loliaceum*; FL pedigree class) using 75,299 SNPs (Fig. 1A and B, respectively). The average genomic relationship was close to zero as expected due to the centering of allele frequencies in both data sets (-0.0178 and -0.024 for hybrid ryegrass and *F. loliaceum*, respectively) but with substantially more variation found in the FL data set (off-diagonal standard deviation equal to 0.21 compared to 0.15 in the HR class). In addition, GRM heatmaps are substantially populated with negative relationships, meaning that many pairs of individuals were less related than the average genomic relationship. Also, the GRMs revealed biparental combinations that substantially deviated from the expected offspring composition of bi-parental crosses of single-plant parents, suggested by the presence of blocks of high genomic relationships (>1.0) among families, especially for the FL data set (Fig. 1B). For instance, the 4×4 block on the top-left side of Fig. 1B holds highly related families that share the same pollen receptor parent crossed with different *F. pratensis* genotypes. As the diallel design was not accounted for, downstream analyses were performed controlling for population stratification due to replicated parents in the crossing scheme using principal component (PC) scores as covariates (See Supplemental Fig. S1 for a visual representation of the population dispersion). The first 10 PCs of the GRM matrices explained a cumulative percentage of variation equal to 75% and 82% for HR and FL data sets, respectively. Additionally, adjusted means on the right-hand side of Fig. 1 reveal blocks of families with similar trait-specific performance as they were hierarchically clustered by IBS-based measurement of relatedness.

The GRMs displayed in Fig. 1 were also used in a linear mixed model to estimate the genomic heritability of NMR variables and gene expression entities. The density plots of the heritabilities for both pedigree classes are displayed in Fig. 2. For the HR class, median heritabilities of 0.047 and 0.122 with an interquartile range (IQR) of 0.177 and 0.311 were observed for NMR variables and gene expression, respectively. For the FL class, we observed median heritability of 0.162 and 0.165 with IQR of 0.273 and 0.295 for NMR variables and gene expression, respectively. Distributions are positively skewed and a higher quantity of high heritable variables can be detected for gene expression data in comparison to NMR variables. Additionally, the figure suggests a slightly higher proportion of more heritable features measured on samples from the FL class, especially for metabolomic data. Finally, Fig. 2B and C reveal the similarity in heritability between pedigree classes according to the spectrum and genomic position, respectively. Overall, there is a high correspondence between classes for regions displaying high and low heritability.

Hyperparameter tuning of joint graphical lasso

The search for the appropriate values of λ_1 and λ_2 (Eq. 4) that returned the smallest Bayesian information criterion (BIC, see Eq. 5) was computationally intensive as the model was fitted for all combinations of the penalties defined in the grid search, requiring several days of CPU time for joint graphical lasso (JGL) model of transcriptomic data but only using few wall time hours by taking advantage of multi-core processing. A total of 939 connected nodes were estimated for gene expression. Within data sets, four sparse subnetworks and 4,038 edges were obtained for HR whereas five sparse subnetworks and 2,182 edges were identified for the FL class given the tuning parameters selected via BIC (Fig. 3). Additionally, 462 edges were found to be shared by the two pedigree crossing classes. For the next omic layer, all 556 nodes (NMR variables) were connected, one sparse network on each pedigree class was estimated, 7,757 and 4,789 edges were available for HR and FL data sets, respectively, and 2,371 common interactions shared by all classes. Less than half of the edges were shared between the two classes, indicating that hybridization had a significant impact on regulatory processes. The HR class exhibited more

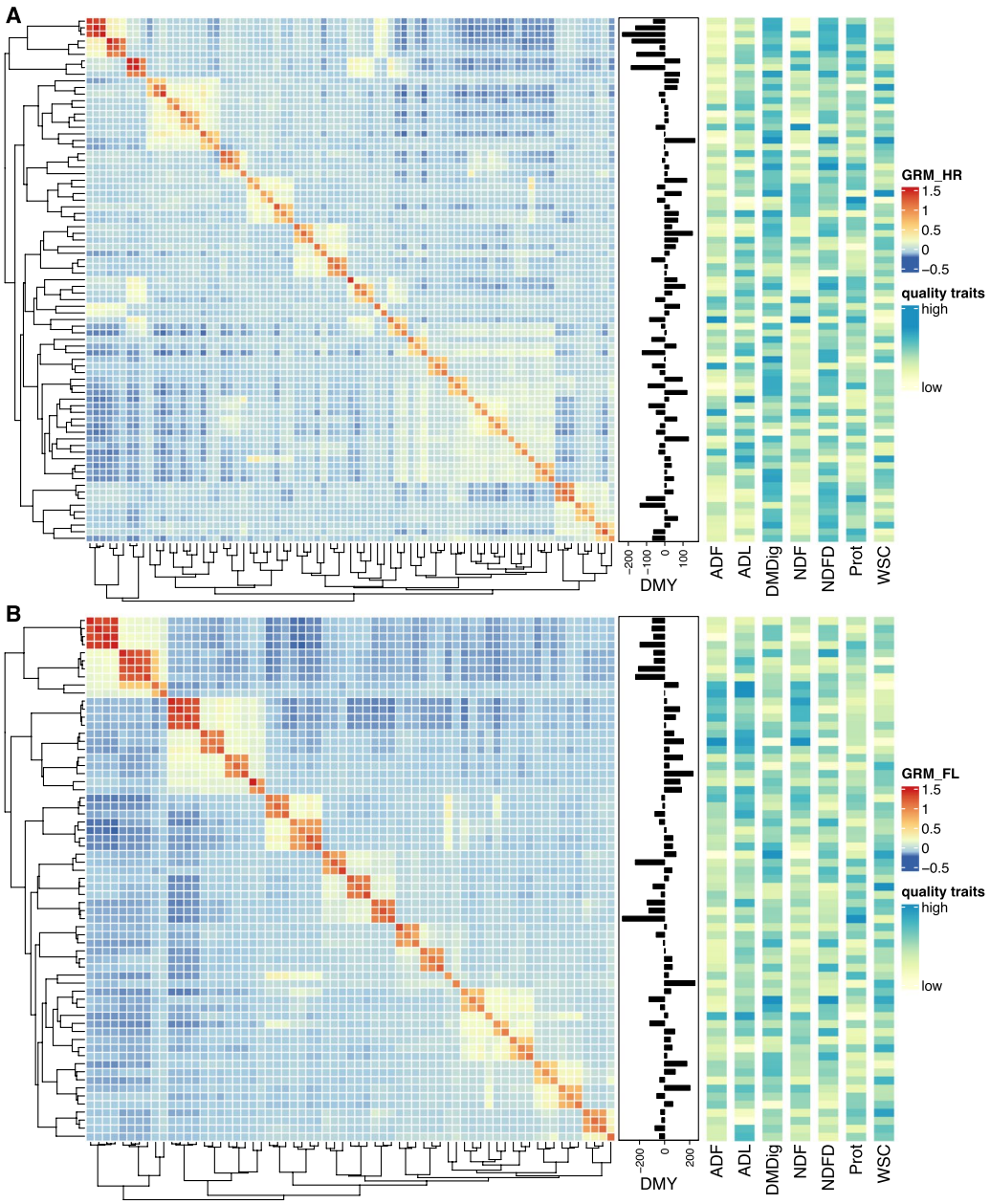

Fig. 1 Heatmaps of genomic relationship matrices annotated with trait means. **A** and **B** show genomic (co)variances between all pairs of 79 families of hybrid ryegrass [HR] while **B** and 65 families of *Festulolium loliaceum* [FL], respectively. Annotations on the right-hand side of each matrix depict the best linear unbiased estimators (BLUEs) for dry matter yield (DMY) and each of the seven nutritive quality traits: ADF - acid detergent fiber; ADL - acid detergent lignin; DMDig - digestible dry matter; NDF - neutral detergent fiber; NDFD - digestible NDF; Prot - protein; and WSC - water-soluble carbohydrates. Partially surrounding dendrograms were produced using Euclidean as the distance measure and the agglomerative complete-linkage method to build the hierarchy of clusters

complex interacting networks in both omics, possibly due to higher hybridization success rates, mixing specific genomes and resulting in novel genetic combinations and regulatory mechanisms.

For the gene expression data, λ_2 was optimized at $\lambda_2 = 0$. This implies different networks for each pedigree class with a different arrangement of non-zero positions for the gene expression data. On the other hand, for NMR data, the best combination of λ_1 and λ_2 that minimized the BIC found a small non-zero value for λ_2, implying a small level of similarity on the sparsity pattern across precision matrices for NMR data. Overall and across omic layers, the

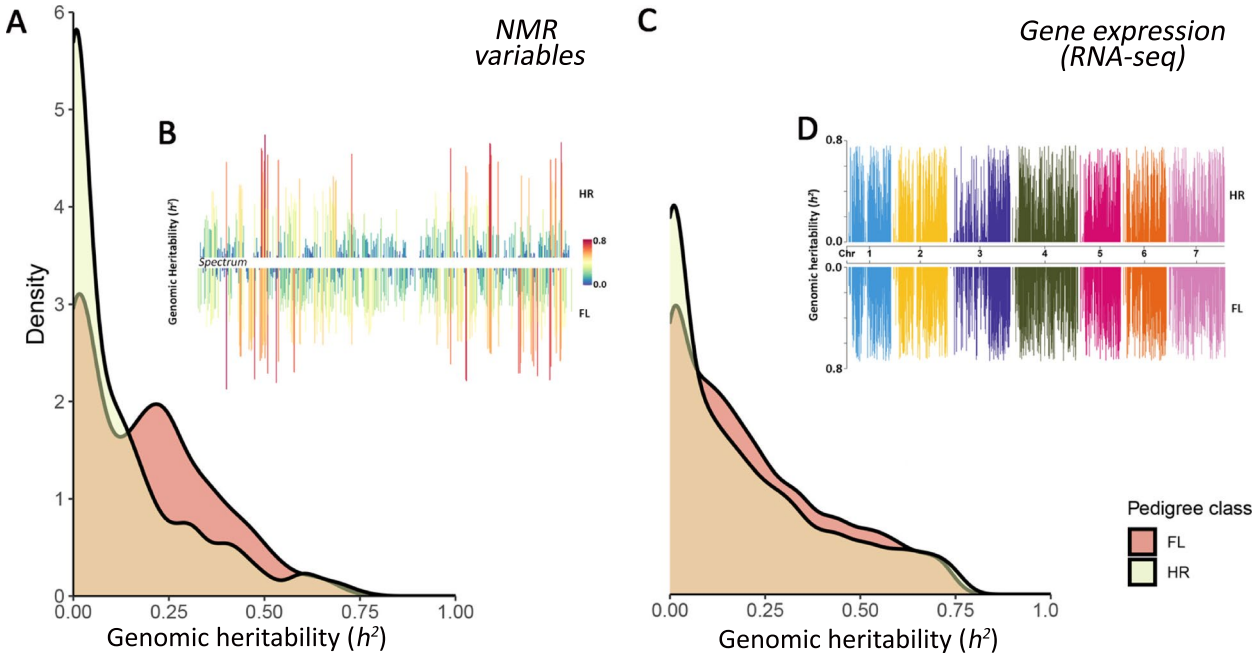

Fig. 2 Genomic heritability of gene expression and NMR features. SNP-based genomic heritability distribution of NMR variables (**A**) and gene expression (**C**) from family pools of two pedigree classes (HR: hybrid ryegrass and FL: *Festulolium loliaceum*) were displayed as density plots. Heritability of transcripts across the genome and NMR variables by position on the spectrum are depicted in subfigures **B** and **D**

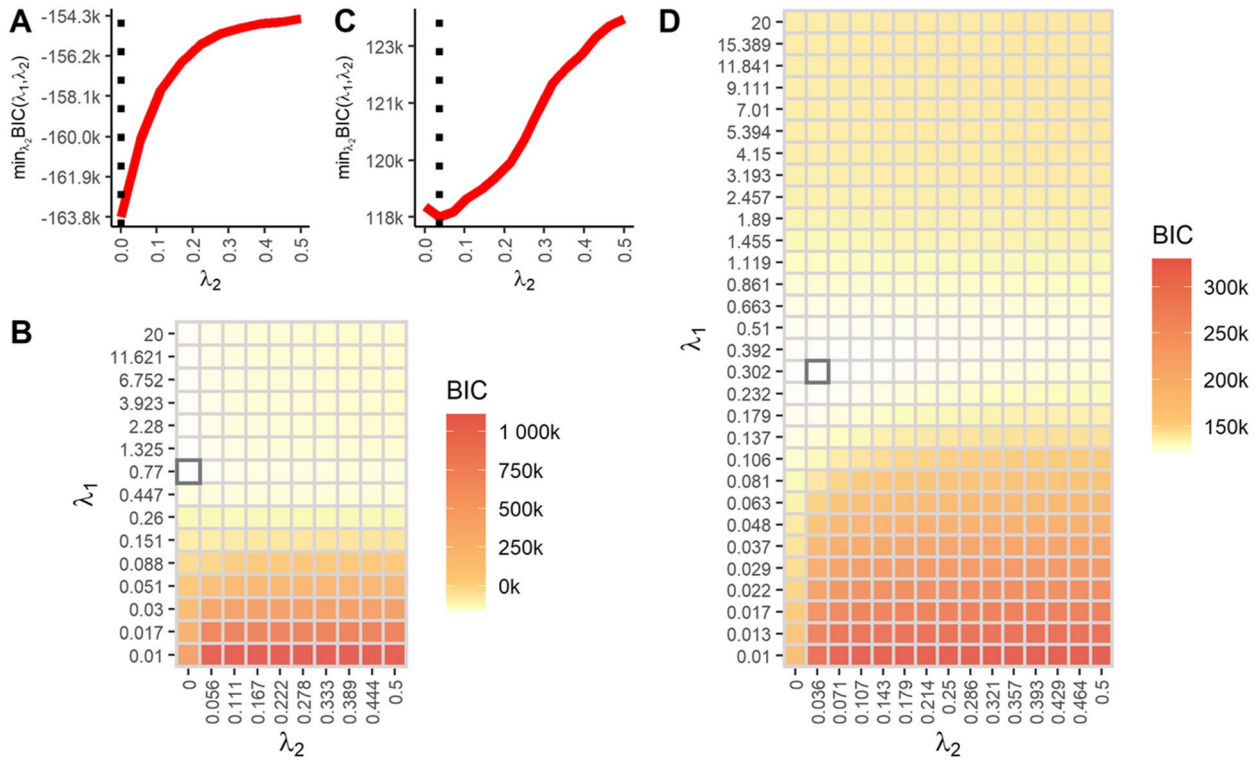

Fig. 3 Grid searching of hyperparameters for graphical lasso model selection with ℓ_1 regularization. **A** and **C** shows the Bayesian information criterion (BIC) as a function of the second (λ_2) penalty for transcriptomic and metabolomic data sets, respectively. **B** and **D** are heatmaps displaying the complete grid search for the values of the tuning parameters λ_1 and λ_2 that minimize BIC, yielding parsimonious models for transcriptomic and metabolomics data sets, respectively

hybridization process generated substantial differences between pedigree classes and it seems to be better captured at the gene expression level.

Exploring lasso penalized precision matrices and network topologies

We detected 14 candidate modules for gene expression and 10 modules for metabolomic for the HR class (Fig. 4). In the FL data sets, it was estimated 16 modules for gene expression and also 10 modules for metabolomics data. The modularity view of the gene-to-gene and metabolite-to-metabolite networks reveals the power-law distribution of node connections, where few vertices are highly connected whereas the majority has only one or few connecting edges. The organization of network structure based on modularity optimization allowed for the selection of intramodular hub nodes that are more likely to be involved in different biological pathways. Out of 70 hubs extracted from HR transcriptomic data (Fig. 4A) and 80 from FL transcriptomic data (Fig. 4C), 30 genes (hubs) were conserved. These high-degree genes are located across all seven chromosomes, varying from two hubs on chromosome three up to 10 on chromosome two. Also, the degree of the hub gene set ranged from 34 to 182 edges. For metabolomic data, we found 32 conserved hub nodes (Fig. 4, B and D), all localized in one half of the NMR spectrum and with degrees ranging from 52 to 357 edges.

Integrative omics

The pairwise fitting of the multivariate genomic model revealed 21 significant edges between traits and omic hub features after FDR correction (Fig. 5). The multi-trait model was fitted 496 times but failed to converge in 54 cases, possibly due to the variance component being close to zero. Therefore, five traits displayed at least one significant edge with hub features in both pedigree classes. More edges can be seen on the left side of the omics-phenotype network relative to the right side, which can be explained by the higher heritability across traits in the FL data set (Supplemental Table S1) as well as overall higher heritability of genomic features (Fig. 2). Additionally, significant connections were found for six out of 30 hub genes and four out of 32 hub NMR variables. Three (hubs 16, 18, and 21) out of the six genes are located distantly apart on chromosome four whereas the remaining hubs 3, 7, and 22 are located on chromosomes one, two, and five, respectively (Supplemental Fig. S2). Genomic heritabilities of hubs displaying significant edges were considerably higher compared to the full feature space, with median h^2 twice as large. A closer look reveals a consistent pattern regarding the direction of the associations. Hub features positively or negatively associated with fiber content traits are also positively or negatively associated, respectively, to dry matter yield. The same holds true for protein content and digestibility traits, where associated hub features are inversely connected to fiber content. Additionally, the majority of hubs associated with phenotypes have more than one significant edge computed from independent analysis and, therefore, confirms the reliability of the estimated omics-phenotype network. We also fitted hub features as covariates in submodel 9 and computed the z-scores and associated p-values, which overall confirmed the results displayed in Fig. 5 (data not shown). Finally, no hub feature had significant edges with traits from both pedigree classes, which can suggest steady genetic differences between classes and/or a lack of power to detect these shared genomic-based associations.

Gene-set enrichment analysis revealed four gene ontology (GO) terms enriched in the set of 30 hub genes displayed in Fig. 5. Overrepresented GO terms were GO:0019438 (aromatic compound biosynthetic process), GO:0018130 (heterocycle biosynthetic process), GO:1901362 (organic cyclic compound biosynthetic process), and GO:0044271 (cellular nitrogen compound biosynthetic process). Bivariate mixed model analysis revealed significant genetic correlations between the expression of gene hubs 18 and 21 and dry matter yield. While hub gene 18 codes for the *atpF* gene (synthase subunit b, chloroplastic) and is associated with energy production (GO:0015986 - proton motive force-driven ATP synthesis), the blast of biological sequences revealed a putative unclassified retrotransposon protein originating from hub gene 21.

Omics-assisted predictions

Using gene expression data as an independent variable performed similarly to SNP-based marker predictions, except for digestibility, protein, and neutral detergent fiber (Fig. 6). Despite the overall poor prediction performance across traits obtained when using NMR features as independent variables, the information contained in this omic layer is useful for protein content prediction, with correlations above 0.4. Prediction accuracy using only hub genes was compared with a second scenario where samples of the same size were drawn from the whole predictor space aiming to check whether hub features carry asymmetrically more (or less) information for prediction purposes. Overall, hub NMR variables appear to be more predictive of nutritive quality traits than random samples of metabolomic features. On the other hand, results suggest a weaker relationship between observed and predicted quality parameters using hub genes as regressors. Finally, using the whole set of available predictors yields predominantly higher accuracies across traits.

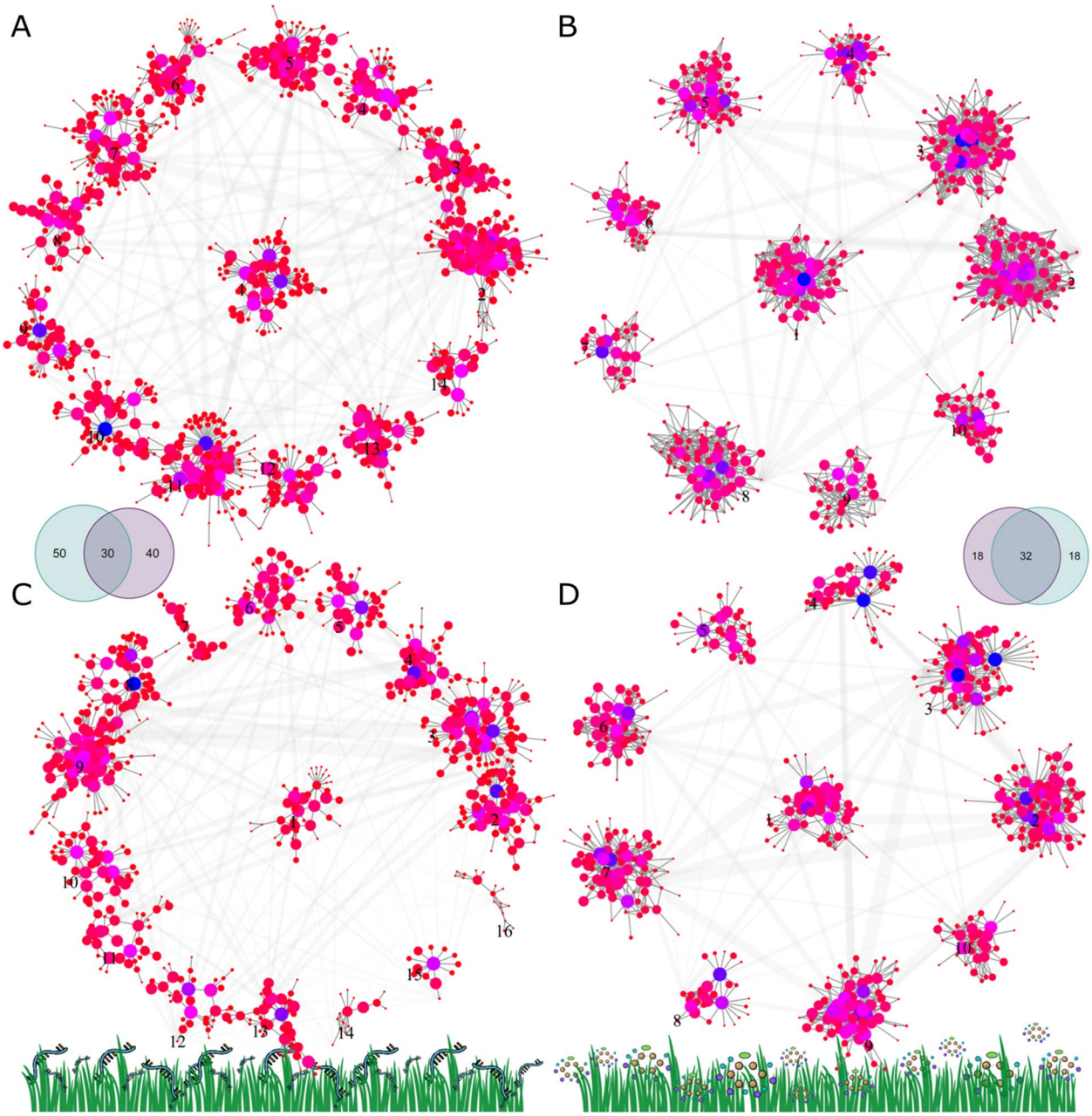

Fig. 4 Abstract modularity view of the gene co-expression and metabolite networks. **A** and **C** networks constructed from gene expression data. **B** and **D** metabolite-metabolite networks built from nuclear magnetic resonance (NMR) spectroscopy. Data of family pools from two pedigree classes were used for network estimation, i.e., **A** and **B** refer to the hybrid ryegrass class while **C** and **D** to the *Festulolium loliaceum*. Both node color and size reflect the hub score, i.e., the principal eigenvector of $\mathbf{Adj} \cdot t(\mathbf{Adj})$ matrix operation, where **Adj** is the adjacency matrix of each graph. The color range goes from red for low-degree nodes to blue for highly connected ones. Edges between modules were collapsed and the width refers to the number of connections shared between any two modules. Venn diagrams show the overlap among sets of top hub features from each data set

Discussion

The study elaborated here explores a network-based approach to combine multi-omic data arising from an $n \ll p$ scenario, inferring associations between biomarker candidates with dry matter yield and nutritive quality traits of polyploid forage grass families. This was accomplished by using a joint graphical lasso model with a fused penalty for network reconstruction, followed by topological property extraction and integration via multivariate mixed modeling. Further, a machine learning-based prediction scheme was explored to verify the extent of information available in hubs and in the

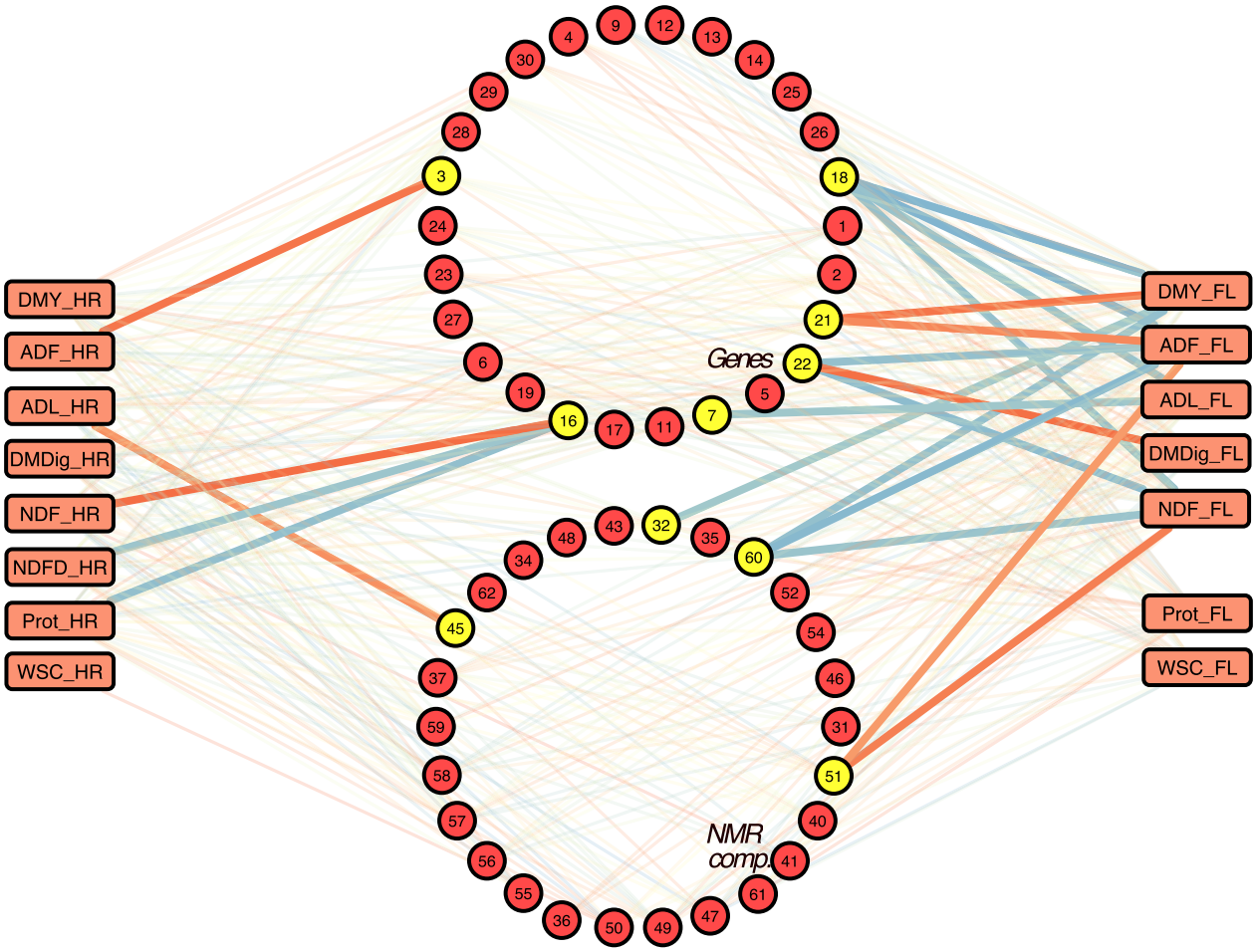

Fig. 5 Weighted network reveals hub features associated with the expression of economic important phenotypes. Traits arranged on the left-hand side were measured on families of hybrid ryegrass while traits arranged on the opposite side were assessed on families of *Festulolium loliaceum*. Edges represent the additive genetic correlation between omic features and traits and were built by the pair-wise fitting of a multivariate genomic model. Stronger edges in a gradient from red (negative) to blue (positive) represent false discovery rate corrected significant correlations at alpha 0.05. Highlighted omic nodes show at least one significant edge. Hub codes from 1 to 62 can be used to gather more information from Supplemental Fig. S2 DMY - dry matter yield; ADF - acid detergent fiber; ADL - acid detergent lignin; DMDig - digestible dry matter; NDF: -neutral detergent fiber; NDFD - digestible NDF; Prot - protein; and WSC - water-soluble carbohydrates

whole feature space for predicting agronomically important phenotypes. The plant material consisted of family pools of inter-specific and- generic grass hybrids from two connected diallels. Crossing different pasture species/genera is not a trivial task; obstacles can emerge. Firstly, out of all initially planned crosses, only a subset generated viable seeds, impacting the sample size. Also, seeds were not abundant for many of the crosses, requiring an additional year of multiplication. Secondly, extraneous offspring patterns were detected, prompting a question of whether normal parental contributions were formed for some of the F_2 families. This inquiry remained unanswered in this manuscript given the complexity of the genetic material (family pools), SNPs called from RNA-seq data, and the unavailability of parental genotypes. Despite the self-incompatibility (SI) ensuring cross-pollination in perennial ryegrass [33], four to eight percent of self-fertilization has been reported [34, 35]. This, in addition to the low success rate of inter-specific and- generic hybridization, might have caused the deviated genomic state of offspring families for crosses that produced a small number of seeds. We did not use the parental information from the diallel structure in the network construction but removed it by regression to control for the kinship among individuals across analyses, a crucial action to avoid spurious results in network reconstruction. Due to the genetic design, correlation among samples is expected, which can lead to the detection of co-expression among features as a result of shared chromosomal segments. Additionally, confounding artifacts

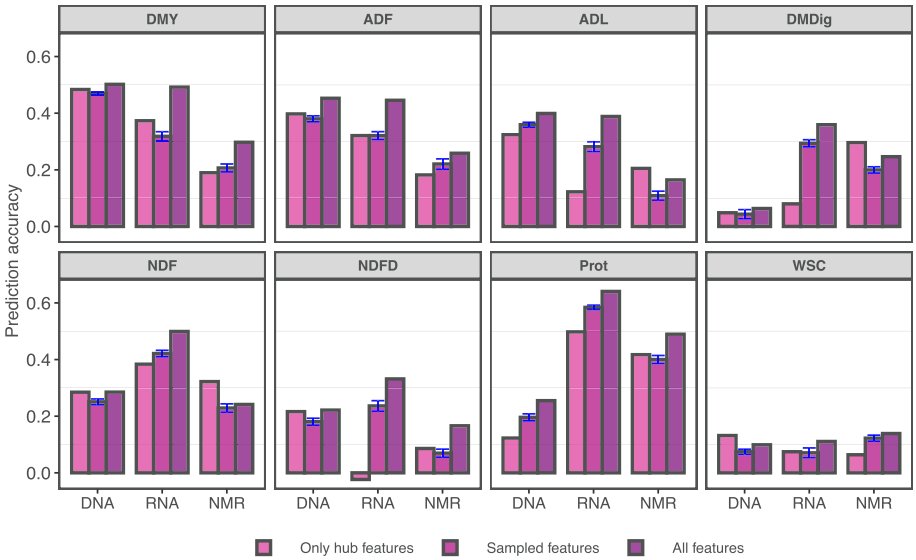

Fig. 6 Random forest-based prediction accuracies reveal strong links between greenhouse and field phenotypes. Accuracies were computed for eight forage grass traits as a function of predictors encompassing three omic layers (DNA: SNP-based markers, RNA: gene expression via RNA-seq, and NMR: variables representing bucketed NMR spectra) and three predictor set configurations as indicated by the color gradient. The standard errors for the mean accuracy of sampled features are depicted in blue color. DMY - dry matter yield; ADF - acid detergent fiber; ADL - acid detergent lignin; DMDig - digestible dry matter; NDF: -neutral detergent fiber; NDFD - digestible NDF; Prot - protein; and WSC - water-soluble carbohydrates

not controlled for can affect groups of genes and NMR variables, which can lead to the detection of spurious correlations. We fitted population structure as covariates by using principal component scores derived from the genetic markers covering the whole genome aiming to alleviate the non-independence among samples, which has been shown to reduce false network discoveries efficiently [36]. An extra layer of precaution to avoid the effect of false-positive edges was deployed by retaining only common hub features between pedigree classes.

The gene co-expression and metabolic networks as the ones we reconstructed in this study (Fig. 4) using RNA-seq and NMR variables, respectively, can contain interesting topological properties e.g., the existence of highly connected nodes and the organization of nodes in modules [12]. We explored these two properties aiming to select, across pedigree classes, conserved hubs extracted at a rate of five per module, therefore, increasing the likelihood of sampling hubs associated with diverse biological processes. Our approach to selecting and associating these features with phenotypic traits is altogether different from the conventional method, which consists of performing a simple correlation-based gene co-expression network analysis followed by thresholding to find modules that can then be summarized into a synthetic (eigen) gene for association with external sample traits [37]. As highlighted by other authors [38, 39], this correlation-based approach cannot distinguish between linear relationships due to directly dependent nodes and those arising from confounding nodes, which might create spurious edges in the graph and, consequently, misleading clustering. In contrast, Gaussian graphical models, as used here, are based on the precision (inverse variance) matrix and express conditional dependence between pairs of features given all the other variables in the data set [40] which, therefore, avoids declaring an edge when no causal relationship exists. Regarding the presence and distribution of edges across reconstructed networks, the proportion of undirected edges given the total available nodes was much higher for the NMR -based metabolic network relative to the gene expression graph. This is a consequence of the lack of independence among bins closely located across the NMR spectrum. Indeed, an average autocorrelation across samples revealed significant spikes up to lag 12 (data not shown). Therefore, a proper feature selection algorithm for spectral data can be implemented to deal with the existence of autocorrelation.

Picturing a biological regulatory cascade, hub genes are usually regulatory factors located upstream, whereas genes represented by low-degree nodes are located on the other end [41]. They can be associated with biological processes from which several others are dependent, yielding the commonly observed power-law degree distribution. The presence of a limited amount of important hub genes, however, does not necessarily imply a simple genetic architecture, because the regulation of the hub gene expression is typically highly polygenic.

Investigating putative hubs can reveal important genes as, for example, the cold-regulated gene *Lolium perenne LIR1* (*LpLIR1*) [42] represented by the hub gene coded as 22 in Fig. 5, which is located at chr5:155166187-155167265 in the *L. perenne* genome and appears to act in the photoperiodic regulation of flowering. The results presented in Fig. 5 suggest that the up-regulation of *LpLIR1* is positively correlated with higher fiber content and reduced digestibility. One possible explanation for this observation could be the effect of heading date on these traits. It is well established that early-flowering genotypes tend to show an early decline in digestibility and higher fiber content, while the opposite is true for late-flowering genotypes. The up-regulation of *LpLIR1* may be influencing heading date, thus leading to changes in fiber content and digestibility. Another example is hub 7, which represents the *PDX1.1* gene, involved in the biosynthesis of vitamin B6 and protection against stresses [43]. Overexpression of PDX proteins was shown to increase seed size and biomass in Arabidopsis [44]. For metabolite-metabolite networks, high-degree nodes may represent signaling molecules or molecules engaged in many reactions. The content and diversity of such molecules have been shown to be shaped by domestication as well as due to crop improvement [45]. Improving biomass output per area is the ultimate breeding goal in a forage breeding program and also implies selection pressure for stress endurance due to animal grazing or mechanical harvesting. In this sense, secondary metabolites are well-known for their role in the plant's response to external disturbances as herbivory [46]. In more general, significant associations can be detected between metabolites and agronomic traits [47] and the whole NMR spectrum can be used for metabolomic-assisted prediction [48]. That being stated, genetic selection for elite grasses might be linked to an altered profile of metabolites, leveraging their usefulness as markers for selection or for prediction purposes. Indeed, great chemical diversity is available in perennial ryegrass [49], not only adding another layer of information for omics-assisted breeding but also enabling target improvement of varieties with a specific profile of key metabolites.

Together, significant additive genetic correlations between omic features and phenotypic traits displayed in Fig. 5 and the presence of over-represented gene ontology (GO) terms in the hub gene set supports the evidence that these features hold fundamental biological properties. We further assessed the predictive power available in the sets of gene and metabolite hubs. This was accomplished by merging the HR and FL data sets for trait prediction aiming to increase the sample size, which even though still below the appropriate size for genomic selection was counterbalanced by a high signal-to-noise ratio given the diallel structure which is expected to boost information for model learning (see Fig. 1). Splitting between training and testing sets would reduce the sample size for training. Therefore, we used the ensemble learning method of random forest with all samples and reported the out-of-bag (OOB) accuracy as a prediction performance metric, eliminating the need to set aside a test set [50]. Despite the crossing scheme, eigenvectors from marker data did not reveal large dissimilarity between pedigree classes (Supplemental Fig. S1), therefore allowing for the joint analysis. Also, random forest is not very sensitive to hyperparameter tunning [51], making it a good option for the designed prediction setup. This can be attested by the magnitude of predictions displayed in Fig. 6. Prediction accuracy for dry matter yield was reported in other studies at 0.31 using diploid ryegrass synthetic populations [52], 0.34 using tetraploid ryegrass [53], and 0.5 investigating diploid perennial ryegrass [2]. Here, we report values of prediction accuracy of dry matter yield that approximate 0.5 (Fig. 6) using both SNP-markers and gene expression, despite the lower sample size but helped by high relatedness among samples, an important component in genomic selection [54]. Also for dry matter yield, surprisingly the most heritable trait (Supplemental Table S1), the set of hub genes and SNPs markers tagging them seem more predictive than features sampled at random. For the remaining traits, mixed results were observed which can be an artifact due to sample size, low heritability, or population structure. Additionally, the signal might be dependent on the genetic background and disappeared as we merged the two data sets for the prediction study. Heritability is an important parameter driving prediction accuracy. If it is low, the error variance will be higher, leading to difficulties in estimating the effect of genome segments accurately [55], especially if the sample size is not sufficiently large. Small values of heritability were primarily observed for quality traits (Supplemental Table S1), which explains the lack of predictive power of the model for digestibility, water-soluble carbohydrates, and digestible NDF, for example. The NIR-based quality parameters are obtained from calibrated models using data of chemical analysis from samples of standard breeding materials and might not translate well into curves of inter-generic and- species hybrids, explaining the lower heritability.

Given that plant tissues were sampled once from pools of seedlings grown in a greenhouse environment at the F_2 generation for transcriptomic and metabolomic analyses, the information carried by the recorded features represents a snapshot of the complex interactome at that particular condition in space, time, and random mating generation. This information was learned by the model and translated into higher prediction accuracy for

protein and digestibility, despite the fact that phenotypes were recorded in later growth stages and in another generation of random mating. Across omic layers, the results also showed that using all available features is almost always a better choice for increased prediction accuracy. Besides more main effects being captured, the random forest model can capture feature-feature interactions [56] as long as the marginal effects are large enough to cause a tree split, therefore, accounting for some of the existing epistasis. Therefore, the existence of significant edges displayed in Fig. 5 and the magnitude of the prediction accuracies presented in Fig. 6 reveals a strong link between field-based phenotypes and heritable omic features assessed from young seedlings in a controlled environment. Altogether, this information brings the question of whether phenotypes from seedlings grown for DNA sampling could be recorded through a low-cost NIR-based method and used to improve the accuracy of genomic selection models, a subject worthy of consideration in future research.

The use of multi-omics in plant breeding-related studies is becoming more popular due to decreasing in cost per data point as a result of modern high-throughput technologies. This has been allowing researchers to reconstruct complex biological networks for inference and mining. Out of the many topological properties that can be retrieved from an interaction network, hub features showing many putative links have been shown to play important biological roles in plants [30]. Our study reveals that narrowing down the high-dimensional feature space generated by high-throughput omic analysis to fewer entities by leveraging properties of the graphical theory can reveal important biomolecules for molecular studies and breeding. Additionally, dimensionality reduction can substantially boost detection power by alleviating the multiple testing problem. Further investigations of candidate features may help elucidate biological processes underlying the expression of phenotypic traits and serve as markers for omics-assisted selection in breeding programs. Even though we did not perform compound identification from the NMR data, this is a feasible task and may reveal metabolites playing important roles in biomass yield and nutritional quality.

Conclusion

The scientific community has seen a sharp increase in publications exploring the usefulness of biological network reconstruction based on high throughput omics data since the 2000s, but studies with forage species remain scarce. Here, we have explored the usefulness of topological properties of gene co-expression and metabolic networks in explaining the phenotypic variance of eight traits assessed in family pools of inter-specific and -generic grass hybrids. Network topology estimated via fused graphical lasso revealed profound network differences between pedigree classes, but a set of 30 high-degree hub genes and 32 hub NMR variables remained conserved across classes given the selection criteria, out of which 10 hubs were found as candidate biomolecules significantly associated with the expression of agronomic phenotypes. Gene set enrichment analysis and weighted omics-phenotype network estimation suggested that sets of hubs are likely to contain essential features modulating interactomes and the expression of economically important phenotypes.

Methods

Plant material and phenotypes

Interspecific hybridization of *L. perenne* × *L. multiflorum* (hybrid ryegrass) and intergeneric crosses of *L. perenne* × *F. pratensis* (*Festulolium loliaceum*), all in tetraploidy forms (2n = 4× = 28), were performed as two connected (by *L. perenne* parents) sparse diallels in the summer of 2017 at the DLF Seeds A/S research station, Store Heddinge - Denmark. Single plants used as parents were extracted from commercial varieties of *L. perenne*, *L. multiflorum*, and *F. pratensis*. A total of 79 and 65 allotetraploid families of hybrid ryegrass and *Festulolium loliaceum*, respectively, were obtained out of several attempts. Hybrid ryegrass (referred to hereinafter as HR) families were obtained after crossing 31 *L. perenne* parents with 79 *L. multiflorum* in a sparse diallel design. For the pedigree class *F. loliaceum* (referred to hereinafter as FL), 24 *L. perenne* parents out of the 31 from the HR diallel were crossed with four *F. pratensis* parents. A sufficient quantity of seeds of F_3 families was obtained after two rounds of multiplication. The field trials were carried out in the autumn of 2020 at two testing sites: i) Denmark (55° 17' 52" N, 12° 24' 58" E) and ii) the Czech Republic (49° 40' 59" N, 17° 58' 05" E). Families from the HR pedigree class were sown in Denmark in plots of 12.5 m^2 with two replicates while families of the FL pedigree class were sown in the Czech Republic in plots of 6.25 m^2, also with two replicates. At each location, entries were assigned to plots arranged in five smaller trials in a randomized complete block design with ~16 entries each. Alongside the described steps, seeds from F_2 families were sown in a greenhouse environment in 2019 at Aarhus University, Research Center Flakkebjerg. One gram of seeds from each family was sown in pots 10 cm in diameter aiming at 120 to 150 emerging individual plants. The total above-ground biomass was harvested as one bulk per family, flash-frozen using liquid nitrogen to stop metabolism, and placed in a -80°C freezer. Frozen tissue ground into a fine powder with liquid nitrogen was used for RNA isolation and sequencing after a quality check. In

addition, aliquots weighing 300 mg from ground tissue were freeze-dried for NMR-based metabolomic profiling.

We collected phenotypes for eight traits at four-time points in Denmark and three-time points in the Czech Republic across the Spring, Summer, and Autumn of the 2021 production year. The following traits were assessed: moisture-corrected dry matter yield standardized by plot size (DMY, g m^{-2}), acid-detergent fiber (ADF), acid-deterged lignin (ADL), dry matter digestibility (DMDig), neutral deterged fiber (NDF), digestible NDF (NDFD), protein (Prot), and water-soluble carbohydrates (WSC). All nutritive quality traits are expressed as a percentage of DMY, except for NDFD, which is a percentage of NDF. Nutritive quality traits were obtained via a near-infrared (NIR) spectrometer onboard the plot combine harvester. Raw NIR data had previously been calibrated and is yearly updated with new wet chemistry analysis, a routine procedure in the breeding company.

Multi-omics data

Gene expression via RNA sequencing

RNASeq libraries were prepared and sequenced at the Beijing Genomics Institute (BGI Hong Kong) using the BGISEQ-500RS sequencing platform technology in 100nt paired-end (PE100) mode. Paired-end reads (20 to 25M sequences per sample) were mapped to pseudo-chromosomes and scaffolds of the Lolium_2.6.1 reference genome [57] using the splice-aware aligner HISAT2 [58]. Alignments were processed by StringTie [59] for transcript reconstruction and gene expression quantification. Normalized read count values in fragments per kilobase of transcript per million (FPKM) were collected for 139,004 transcripts annotated on the Lolium_2.6.1 reference genome. A filter was applied to the expression profile matrix to get rid of transcripts with expression values very low/equal to zero. The threshold for transcription was set to 0.5 median FPKM across all samples, yielding the final filtered gene expression matrix with 18,499 transcripts.

RNASeq-based genetic variants

Variant calling was performed from RNA-seq merged BAM-format alignments using the Bayesian genetic variant detector Freebayes [60]. The initial single-nucleotide polymorphism (SNP) calling resulted in 1,689,206 variants. After retaining only biallelic markers, we filtered variants by the following criteria: i) a maximum missing proportion of 50% at each locus, ii) a minimum mapping quality of 20, iii) a minimum read coverage of five reads per variant position, and iv) minor allele frequency (MAF) greater than 0.05. The final set of SNPs comprises 89,862 variants that were used for downstream analyses.

NMR-based metabolomic data

The metabolomic profiling by proton nuclear magnetic resonance spectroscopy (1H-NMR) was carried out at the Natural Products Laboratory (The Netherlands). Following the sample preparation and spectra acquisition with a 600 MHz Bruker AVANCE III spectrometer (Bruker BioSpin GmbH, Germany), the raw NMR data were processed using the software package NMRProcFlow [61]. After chemical shift calibration and normalization, metabolomic fingerprinting yielded a total of 556 bins with non-zero intensities (referred to hereafter as NMR variables) for 144 plant samples by applying an adaptive Intelligent Binning [AI-Binning, [62]] algorithm. A tab-separated file with samples on rows, NMR variables on columns, and cell-wise intensity values was generated for downstream analysis.

Statistical analysis

Prior exploratory analysis revealed considerable differences between the omics data from the HR class compared to FL class samples. Therefore, downstream analyses were performed considering each of the two classes as distinct but related across layers of omics data. Additionally, this decision was supported by the fact that phenotypes were assessed in different locations, lacking connectedness. Later, these data sets were merged for an omic-assisted prediction study.

Allele frequency-based genomic kernel

The genomic relationship matrix (GRM), which gives the realized genetic similarities among any pair of individuals, was computed for SNP data sets of sizes $p \times n$ equal to 85,283×79 for the HR and 75,299×65 for FL data sets after individually re-filtering by MAF, depth, and missing rate using the same thresholds as described before. The GRM was then used for downstream omics feature corrections due to population stratification and multivariate mixed model analysis. The GRM based on pooled DNA was calculated using method 2 in VanRaden [63] adapted to use allele frequencies instead of discrete genotype calls. First, a column-centered matrix $\mathbf{M}$ was computed as $\mathbf{M} = \mathbf{F} - \overline{\mathbf{F}}$, where $\mathbf{F}_{ij}$ is a matrix of alternative allele frequencies with i indexing samples and j indexing SNP markers. The matrix $\mathbf{G}$ can then be obtained as shown in Eq. 1.

$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}'}{\frac{1}{n}\sum_{i=1}^{m}\hat{p}_j(1-\hat{p}_j)} \tag{1}$$

where n is the ploidy of the breeding material, m is the number of markers, and $\hat{p}_j$ represents the frequency at jth locus simply obtained by taking the column means of the $\mathbf{M}$ matrix. As outbred full-sib F_2 families of tetraploid

plants, the genotype of a family can be described as octoploid [64]. Therefore, the realized relatedness is obtained by scaling the plain genomic relationship matrix from the cross product of **M** by the expected SNP variances, yielding a kernel that is analogous to the traditional numerator relationship matrix, also known as the **A** matrix. Finally, a diagonal correction was applied to **G** considering ploidy number and coverage depth [65].

Adjustment for population stratification

The impute file for the analysis of gene expression data consisted of two subsets of 4,767 features times the number of samples of each pedigree class. The reduced set of genes was obtained after further filtering out transcripts with more than 50% of samples having zero reads and retaining positions with at least 10 or more samples having 10 or more reads. Additionally, a filter on the expressional variance of non-zero elements was performed, selecting features ranked in the top 50th percentile as the variation for genes in the bottom may be largely due to non-biological noise. Finally, we retained only features common to both data sets followed by the addition of a pseudo count to the expression matrix, which was subsequently log(2)-transformed [$\log_2$(x+1)]. The input file for the analysis of NMR data consisted of two subsets of 556 NMR variables for each pedigree class. NMR features were mean-centered and variable intensities were addressed via Pareto scaling, which uses the square root of the standard deviation to reduce the relative importance of high-variance features across the spectrum without much disturbance to the data structure.

Population stratification was detected in an unsupervised manner via the multivariate statistical technique of principal component analysis and corrected via regression modeling. We empirically retained coordinates of the top 10 eigenvectors of each k pedigree class to regress out population stratification as well as possible batch effects among samples. Therefore, the transcriptomic and metabolomic data sets were feature-wise corrected by incorporating principal component scores in the linear model of the form described in Eq. 2.

$$y_i = \mu + \sum_{p=1}^{P} \left(x_{ip}^{PC} \beta_p \right) + \varepsilon_i \tag{2}$$

where, y_i represents the response variable i (omic feature); x_{ip}^{PC} is the entry-specific coordinates of the pth principal component, with $p = 1...P$ where P is equal to 10, β_p is the fixed regression coefficients adjusting for population stratification, and ε_i is the residual which was retained to reconstruct the full corrected omics data sets for network estimation.

Joint graphical lasso analysis for inverse covariance estimation

A joint graphical lasso (JGL) method was used for estimation in a scenario of double-related Gaussian graphical models. The two-class problem of high dimensional features was present in the data set due to the available inter-species/genus crosses. One can expect similar graphical models between the two classes as parents were shared among crosses between them, but also some nuances once the involved species have substantial differences regarding phenotypic traits. Therefore, the joint graphical lasso [40] can handle this situation by estimating two graphical models, one for each pedigree class, and borrowing information across classes. For each pedigree class k ($k = 1, 2$), let a data matrix $\mathbf{X}^{(k)}$ represent column-centered data with p omic features, and $\mathbf{X}^{(k)} \sim N\left(\mu^{(k)}, \Sigma^{(k)}\right)$, where $\Sigma^{(k)}$ is a positive definite $p \times p$ covariance matrix of the omic features. The inverse of $\Sigma^{(k)}$ is the precision matrix $\Theta^{(k)}$ representing the network structure of omic features. By applying an ℓ_1-penalty on $\Theta^{(k)}$ the network is made sparse, where elements will be 0 for conditionally independent pairs of features given the remaining variables. The sparsity condition allows learning graphics even in small sample sizes. The fused graphical lasso formulation in which $\Theta^{(k)}$ are estimated by maximizing the penalized form of the likelihood function for the two classes is shown in Eq. 3.

$$\underset{\{\mathbf{\Theta}\}}{maximize} \left\{ \sum_{k=1}^{2} n_k \left(log\ det\ \mathbf{\Theta}^{(k)} - trace\left(\mathbf{S}^{(k)} \mathbf{\Theta}^{(k)} \right) \right) - P(\{\mathbf{\Theta}\}) \right\} \tag{3}$$

where $P(\{\Theta\})$ is as follows:

$$P(\{\Theta\}) = \lambda_1 \sum_{k=1}^{2} \sum_{i \neq j} \left| \theta_{ij}^{k} \right| - \lambda_2 \sum_{h<k} \sum_{i,j} \left| \theta_{ij}^{h} - \theta_{ij}^{k} \right| \tag{4}$$

here, $\mathbf{S}^{(k)}$ is the empirical covariance matrix of omics features calculated as $\mathbf{S}^{(k)} = n^{-1} \mathbf{X}^{(k)} \mathbf{X}^{(k)T}$. The optimization problem is here solved by the alternating direction method of multipliers (ADMM) algorithm. The solution to the problem of $n \ll p$ in the joint graphical lasso model is based on a penalized log-likelihood approach. In addition, as can be seen in Eq. 4, running JGL requires tuning two nonnegative parameters (λ_1 and λ_2). The λ_1 penalty controls the degree of sparsity while λ_2 determines network similarity. If λ_2 is zero (i.e., no penalty is imposed) then $\Theta^{(k)}$are independent and no information is shared between them. To select the proper hyperparameters, we used a goodness-of-fit approach where a grid search was performed to select values that minimize the Bayesian information criterion [BIC] [66] specified in Eq. 5 [67], yielding values that balance model likelihood and complexity.

$$BIC(\lambda_1, \lambda_2) = \sum_{k=1}^{2} \left[n_k \left\{ tr(\mathbf{S}^{(k)} \hat{\mathbf{\Theta}}^{(k)}) - log\ det\ \hat{\mathbf{\Theta}}^{(k)} \right\} + log\ n_k \sum_{i \leq j} \mathbf{1}_{\left\{ \hat{\mathbf{\Theta}}_{ij}^{k} \neq 0 \right\}} \right] \quad (5)$$

In order to reduce the computational burden, a dense search was performed over λ_1 for each fixed value of λ_2 and a quick search for the former parameter for each fixed value of λ_1 as suggested by [40]. For the metabolomic data set, a uniform log spaced grid starting from 0.01 to 20 with a size equal to 30 was defined for λ_1 whereas a simple sequence equally spaced from 0 to 0.5 (size of 15) was defined for λ_2. The same grid search space was defined for transcriptomic data, however, smaller sizes of 15 for λ_1 and 10 for λ_2 were specified. After selecting the proper hyperparameter values, we run JGL for each omics data set producing four precision matrices $\Theta^{(k)}$. From these matrices, one can compute the partial correlation between pairs of dependent features as $corr_{ij|\mathbf{V}\setminus\{i,j\}} = -\theta_{ij}/\sqrt{\theta_{ii}\theta_{jj}}$. The joint graphical lasso method implemented in the R package JGL [40] was used for network estimation.

Network reconstruction, candidate modules, and hubidentification

Network analyses aiming for complexity reduction were performed in order to prioritize candidate genes and metabolomic features for further integration with phenotypes of interest. Initially, each precision matrix $\Theta^{(k)}$ was converted into a symmetric (graph is undirected) 0-1 matrix of dimensions equal to $p \times p$, referred to as the adjacency matrix $\mathbf{A}^{(k)}$ for each k data set following the definition:

$$\mathbf{A}_{ij}^{(k)} = \begin{cases} 1 \ \ if \ \ \mathbf{\Theta}_{ij}^{(k)} \neq 0, \ i \neq j; \\ 0 \qquad otherwise \end{cases} \quad (6)$$

Four adjacency matrices $\mathbf{A}$ were obtained and from them, we created graphic objects using the R package igraph [68]. Initially, a graph is denoted as $G = (V, E)$ in which each node $v \in V$ represents a biomolecule in this study, whereas each edge $e = (v_i, v_j) \in E$ refers to the interaction between pairs of nodes v_i and v_j. Each graph was organized in modules (communities) via a multi-level modularity optimization algorithm [69], forcing highly connected edges to cluster in modules that are sparsely connected among them. In other words, more edges occur within identified modules than the quantity expected at random. The community structure is essential for finding hub nodes that are more likely to be involved in different biological processes.

Hub features were identified intramodule via maximum Kleinberg's hub centrality score, which is the principal eigenvector of $\mathbf{A}^{(k)} \cdot \left(\mathbf{A}^{(k)}\right)^T$ [70]. By using the hub scores, one can identify the most influential features in the network and explore the biological function of these interacting biomolecules. Therefore, we selected the top five hub features per module and kept only those intersecting across data sets to maximize the probability of selecting true/conserved hubs of genes and metabolites. Aside from boosting power by minimizing the issue of multiple testing, information about conserved hub genes is more likely to possess broad biological significance.

REML variance components and heritability

Single omic features were analyzed by fitting a linear mixed model of the form: $\mathbf{y} = \mathbf{1}\mu + \mathbf{Zu} + \mathbf{e}$, where $\mathbf{y}$ is the response vector (normalized gene expression values or total area of the bin from the bucketed NMR spectrum), $\mathbf{1}$ is a vector of ones linking observations to the constant μ, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$ are vectors of the random additive genetic with covariance structure $\mathbf{G}$ (Equation 1) and independent (identity matrix $\mathbf{I}$ as covariance structure) residual effects, respectively. $\mathbf{Z}$ is the design matrix assigning observations of omic features to the respective F_2 family. The genomic heritability was calculated as $h_g^2 = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$, where h_g^2 measures the proportion of the variance attributed to allele substitution effects captured by the genome-wide markers relative to the total variance.

Phenotypic variance within location was partitioned into the terms defined by the linear mixed model displayed in Equation 7:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zu} + \sum\nolimits_{i=1}^{11} \mathbf{S_i s} + \mathbf{e} \quad (7)$$

where, $\mathbf{y}$, $\boldsymbol{\beta}$, $\mathbf{u}$, $\mathbf{s}$, and $\mathbf{e}$ represent the vectors of the response variable, fixed trial-block effect, random additive genetic effect following $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$, random spatial effect following $\mathbf{s} \sim N(\mathbf{0}, \mathbf{I}\sigma_s^2)$, and random residual effect assumed $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, respectively. Matrices $\mathbf{G}$ and $\mathbf{I}$ are as defined before. Design matrices $\mathbf{X}$, $\mathbf{Z}$, and $\mathbf{S}$ link observations of the response variable to the specific model effect. The spatial effect is a sliding window accounting for 10 neighboring plots in addition to the target experimental unit and works by scanning the field for spatial variation not accounted for by the prior trial design. Genomic heritability was calculated as: $h_g^2 = \sigma_u^2/(\sigma_u^2 + 11\sigma_s^2 + \sigma_e^2)$. Variance components and heritabilities for eight phenotypic traits can be found

in the Supplemental Table S1. Finally, the parameter σ_u^2 was multiplied by the average diagonal of the GRM in both heritability equations presented before.

Phenotypes and omics integration via pairwise fitting of mixed models

The raw phenotypic data were analyzed alongside hub omic features in a multitrait genome-wide fashion via linear mixed models to investigate pair-wise additive genetic correlations. The bivariate model (Eq. 8 and 9) was fitted *lm* times, combining *l* hub nodes and *m* phenotypic traits, for each data set, yielding correlations used to describe the existence of a significant association between the concentration of selected biological molecules and economically important phenotypes.

$$\mathbf{y}_{OME_l} = \mathbf{X_1}\boldsymbol{\beta}_{OME_l} + \mathbf{X_2}\mathbf{b}_{OME_l} + \mathbf{Z}\mathbf{u}_{OME_l} + \mathbf{e}_{OME_l} \tag{8}$$

$$\mathbf{y}_{PHE_m} = \mathbf{X_1}\boldsymbol{\beta}_{PHE_m} + \mathbf{X_2}\mathbf{b}_{PHE_m} + \mathbf{Z}\mathbf{u}_{PHE_m} + \sum_{i=1}^{11}\mathbf{S_i}\mathbf{s}_{PHE_m} + \mathbf{e}_{PHE_m} \tag{9}$$

where $\mathbf{y}_{OME_l}$ and $\mathbf{y}_{PHE_m}$ are vectors of expression/intensities of hub omic features and records of phenotypic traits, respectively; $\boldsymbol{\beta}_{OME_l}$ contains the fixed general mean effect while $\boldsymbol{\beta}_{PHE_m}$ also contains the fixed effect of block within trial; vectors $\mathbf{b}_{OME_l}$and $\mathbf{b}_{PHE_m}$ contains fixed regression coefficients estimated by regressing response variables on principal components' dimensional scores calculated from the genomic kernel; $\mathbf{u}_{OME_l}$ and $\mathbf{u}_{PHE_m}$ are vectors of families' additive genetic effect; $\mathbf{s}_{PHE_m}$ is the vector of random spatial effect with $\mathbf{s}_{PHE_m} \sim N\left(0, \mathbf{I}\sigma^2_{s_{PHE_m}}\right)$; and $\mathbf{e}_{OME_l}$ and $\mathbf{e}_{PHE_m}$ are vectors of random residuals for expression/intensity of hub omic feature *l* and phenotypic trait *m*, respectively. For incidence matrices $\mathbf{X}$ linking fixed effects to response variables, the general mean was the only fixed effect for submodel 8, thus $\mathbf{X_1} = \mathbf{1}$. Matrices $\mathbf{X_2}$ contain scores of the top three principal components computed from the $\mathbf{G}$ matrix (Eq. 1) instead of 1's and 0's, aiming at further accounting for population structure to avoid false-positive associations. The selection of the appropriate number of PC's followed an empirical evaluation of the changes in response variables' heritabilities as they were added. The matrix $\mathbf{Z}$ is the corresponding incidence matrix of additive family effects. Finally, the series of matrices $\mathbf{S_i}$ link the random spatial effect to the surrounding plots and work as a sliding window (cross-shaped format) mapping the field for micro-environmental variations missed by the blocking design. The joint covariance structure of the remaining random terms was assumed as follows:

$$\begin{bmatrix}\mathbf{u}_{OME_l}\\ \mathbf{u}_{PHE_m}\end{bmatrix} \sim N\left(\mathbf{0},\ \mathbf{G}\otimes\begin{bmatrix}\sigma^2_{u_{OME_l}} & \sigma_{u_{OME_l}u_{PHE_m}}\\ \sigma_{u_{OME_l}u_{PHE_m}} & \sigma^2_{u_{PHE_m}}\end{bmatrix}\right) \tag{10}$$

and

$$\begin{bmatrix}\mathbf{e}_{OME_l}\\ \mathbf{e}_{PHE_m}\end{bmatrix} \sim N\left(\mathbf{0},\ \mathbf{I}\otimes\begin{bmatrix}\sigma^2_{e_{OME_l}} & 0\\ 0 & \sigma^2_{e_{PHE_m}}\end{bmatrix}\right) \tag{11}$$

where $\mathbf{I}$ represents an identity matrix and $\otimes$ is the Kronecker product. Besides the scores of the first three principal components, here $\mathbf{G}$ also accounts for the whole-genomic relationship structure of the population. Covariances between response vectors were set to non-existent for residual genetic and error random effects.

For hypothesis testing, we also ran a constrained version of the bivariate model, setting the additive genetic covariance between submodels 8 and 9 (Eq. 10) to zero ($\sigma_{u_{OME_l}u_{PHE_m}} = \sigma_{u_{OME_l}u_{PHE_m}} = 0$). The significance of the additive genetic correlations was tested by comparing the constrained and unconstrained models via a one-tailed log-likelihood ratio test (LRT) with 0.5 degrees of freedom [71, 72]. Multiple testing correction was performed for coefficients across traits within omic features via Benjamini-Hochberg false discovery rate (FDR) [73] procedure at alpha equals 0.05 aiming to control for type I error.

The *lm* additive genetic correlations estimated by fitting the full bivariate model for each data set were retained along with the p-values and FDR-based significant associations and used for constructing the omics-phenotype weighted network graph. A visualization of the network was produced using the software Cytoscape 3.9.1 [74], weighing edges by the magnitude of the trait-omic associations.

Gene ontology enrichment analysis

Transcript protein sequences were subjected to local InterPro analysis using InterProScan v5.28-67.0 [75]. Predictive information concerning conserved protein domains, signal peptides, transmembrane domains, and gene ontology (GO) data was acquired from 14 member databases of InterPro. Per transcript, non-redundant GO information was collected from InterPro outputs using custom scripts. GO-term enrichment analysis was carried out using the Python library GOATOOLS [76] by intersecting the GO-term list of the full perennial ryegrass transcriptome, the GO-term subset of expressed genes, and the GO-term lists of filtered transcript sets (study lists). Significant enrichment was declared via Fisher Exact Test, corrected for false discovery rate [73].

Omics-assisted prediction

Starting from the centered **M** matrix of SNP markers defined before, missing allele frequencies were imputed by chained random forest. This method was selected after comparing the ability in predicting missing allele frequencies against the weighted K-nearest neighbors (KNN) method via cross-validation. The imputations were performed for each pedigree class separately using the R package missRanger [77]. The missRanger function ran using the arguments *num.trees* equal to 100, *sample.fraction* equal to 0.1, *max.depth* of 6, and extratrees for the *splitrule* argument. The imputation was performed by looping over one chromosome at a time within clusters of SNPs created by running a complete-linkage clustering algorithm with $k = 30$ as the desired number of groups.

We used the best linear unbiased estimator (BLUE) of entries as response variables in the prediction study. The adjusted phenotypes were obtained by rearranging the terms and refitting the submodel in Eq. 9 with families as a fixed effect and no PC scores were included. BLUEs within locations were mean-centered to remove differential environmental effects followed by the merging of phenotypes and predictors from HR and FL data sets. The unsupervised machine learning algorithm random forest was used as the engine for the prediction study. Models were fitted using the 'ranger' R package [78] with the hyperparameters minimum node size and a number of randomly drawn candidate features set to five and $\lfloor\sqrt{n}\rfloor$, respectively, where n is the number of variables. Therefore, the random forest model was fitted on the combined data sets, setting the number of decision trees to 2,000. Training out-of-the-bag accuracy (OOB accuracy) was reported as a performance metric. Finally, variable importance was computed via permutation.

Three prediction scenarios were studied. First, we selected a subset of SNPs tagging common hub genes across data sets, the common hub genes, and the common hub NMR variables as three sets of regressors. The second scenario consisted of stochastically sampling 20x sets of 30 genes (then SNPs within these genes) and 32 NMR variables aiming to compare the prediction power contained in hub nodes with randomly sampled features. In the last scenario, we used all common SNPs, genes, and NMR variables as regressors. Besides comparing prediction accuracy with the previous scenarios, here we can assess a common prediction task where the goal is to evaluate the closeness of predicted and observed values using all available predictor variables.

Statistical computing and data visualization

Large-scale computations were performed in the GenomeDK high-performance computing facility located at Aarhus University, Denmark. Mixed model analyses were fitted using DMU package version 6 [79]. Modular network visualizations were produced using the R package NetBioV [80] with the Fruchterman-Reingold layout algorithm to arrange nodes in each module. Finally, miscellaneous plots wore drawn employing the ggplot2 R package [81].

Abbreviations

ADF	acid detergent fiber
ADL	acid deterged lignin
BIC	Bayesian information criterion
BLUE	best linear unbiased estimator
DMDig	dry matter digestibility
DMY	dry matter yield
FDR	false discovery rate
FL	*Festulolium loliaceum*
GBS	genotyping-by-sequence
GO	gene ontology
GRM	genomic relationship matrix
HR	hybrid ryegrass
IQR	interquartile range
JGL	joint graphical lasso
LRT	log-likelihood ratio test
MAF	minor allele frequency
NDF	neutral deterged fiber
NDFD	Digestible NDF
NIR	Near-infrared spectroscopy
NMR	Nuclear magnetic resonance
OOB	Out-of-the-bag accuracy
PC	Principal component
Prot	Protein
REML	Restricted maximum likelihood
RNA-seq	RNA sequencing
SNP	Single nucleotide polymorphism
WSC	Water-soluble carbohydrate

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09292-7.

Additional file 1. Genetic structure via principal component analysis.

Additional file 2. Variance components estimated via Restricted maximum likelihood (REML).

Additional file 3. Exploring the genomic and NMR landscape of hub nodes.

Acknowledgements

We thank all the people involved for their contribution during various stages of the research study and manuscript preparation.

Authors' contributions

E.B. conceptualized the study, wrote the code for analyses and data visualizations, interpreted the results, and drafted the manuscript. D.F. contributed to the conception and design of the B4B project and data curation. I.N. contributed to omics data generation and miscellaneous bioinformatics analyses. I.L. contributed with bioinformatics expertise and project design. M.G. created the populations, carried out field trials, and recorded agronomic phenotypes. T.D. converted NIR spectrums into nutritive quality parameters. C.S.J., T.A., and L.J. contributed to the conception and design of the B4B project and funding acquisition. T.A. acquired funding for and coordinated the B4B project. L.J. supervised the current study and provided valuable comments. All authors critically reviewed the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the Innovation Fund Denmark, through the project Breed4Biomass (grant 6150-00020B).

Availability of data and materials
The omics datasets supporting the conclusions of this article have been made available through an R data package named breed4biomass, which is openly available on GitHub (https://github.com/elesandrobornhofen/breed4biomass).

Declarations

Ethics approval and consent to participate
This study implements all methods in compliance with relevant institutional, national, and international guidelines and regulations. No ethics approval was required for this study. The three species used in this study are not considered as endangered or protected. All plant material belongs to the DLF Seeds A/S breeding company (Store Heddinge, Denmark).

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 10 January 2023 Accepted: 2 April 2023
Published online: 25 April 2023

References

1. Keep T, Sampoux JP, Blanco-Pastor JL, Dehmer KJ, Hegarty MJ, Ledauphin T, et al. High-throughput genome-wide genotyping to optimize the use of natural genetic resources in the grassland species perennial ryegrass (*Lolium perenne* L.). G3 Genes Genomes Genet. 2020;10(9):3347–64. https://doi.org/10.1534/g3.120.401491.
2. Arojju SK, Cao M, Trolove M, Barrett BA, Inch C, Eady C, et al. Multi-trait genomic prediction improves predictive ability for dry matter yield and water-soluble carbohydrates in perennial ryegrass. Front Plant Sci. 2020;11. https://doi.org/10.3389/fpls.2020.01197.
3. Fè D, Cericola F, Byrne S, Lenk I, Ashraf BH, Pedersen MG, et al. Genomic dissection and prediction of heading date in perennial ryegrass. BMC Genomics. 2015;16(1). https://doi.org/10.1186/s12864-015-2163-3.
4. Pignon CP, Fernandes SB, Valluru R, Bandillo N, Lozano R, Buckler E, et al. Phenotyping stomatal closure by thermal imaging for GWAS and TWAS of water use efficiency-related genes. Plant Physiol. 2021;187(4):2544–62. https://doi.org/10.1093/plphys/kiab395.
5. Wen W, Li D, Li X, Gao Y, Li W, Li H, et al. Metabolome-based genome-wide association study of maize kernel leads to novel biochemical insights. Nat Commun. 2014;5(1). https://doi.org/10.1038/ncomms4438.
6. Kopecký D, Šimoníková D, Ghesquière M, Doležel J. Stability of Genome Composition and Recombination between Homoeologous Chromosomes in *Festulolium* (*Festuca* × *Lolium*) Cultivars. Cytogenet Genome Res. 2017;151(2):106–14. https://doi.org/10.1159/000458746.
7. Akiyama Y, Kimura K, Yamada-Akiyama H, Kubota A, Takahara Y, Ueyama Y. Genomic characteristics of a diploid F4 festulolium hybrid (*Lolium multiflorum* × *Festuca arundinacea*). Genome. 2012;55(8):599–603. https://doi.org/10.1139/g2012-048.
8. Glombik M, Copetti D, Bartos J, Stoces S, Zwierzykowski Z, Ruttink T, et al. Reciprocal allopolyploid grasses (*Festuca* × *Lolium*) display stable patterns of genome dominance. Plant J. 2021;107(4):1166–82. https://doi.org/10.1111/tpj.15375.
9. Hu G, Hovav R, Grover CE, Faigenboim-Doron A, Kadmon N, Page JT, et al. Evolutionary Conservation and Divergence of Gene Coexpression Networks in *Gossypium* (Cotton) Seeds. Genome Biol Evol. 2017;evw280. https://doi.org/10.1093/gbe/evw280.
10. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: mathematical aspects. BMC Bioinformatics. 2016;17(S2). https://doi.org/10.1186/s12859-015-0857-9.
11. Naserkheil M, Ghafouri F, Zakizadeh S, Pirany N, Manzari Z, Ghorbani S, et al. Multi-omics integration and network analysis reveal potential hub genes and genetic mechanisms regulating bovine mastitis. Curr Issues Mol Biol. 2022;44(1):309–28. https://doi.org/10.3390/cimb44010023.
12. Li Y, Pearl SA, Jackson SA. Gene networks in plant biology: approaches in reconstruction and analysis. Trends Plant Sci. 2015;20(10):664–75. https://doi.org/10.1016/j.tplants.2015.06.013.
13. Pereira-Leal JB, Audit B, Peregrin-Alvarez JM, Ouzounis CA. An Exponential Core in the Heart of the Yeast Protein Interaction Network. Mol Biol Evol. 2004;22(3):421–5. https://doi.org/10.1093/molbev/msi024.
14. Chen P, Liu P, Zhang Q, Bu C, Lu C, Srivastava S, et al. Gene Coexpression Network Analysis Indicates that Hub Genes Related to Photosynthesis and Starch Synthesis Modulate Salt Stress Tolerance in Ulmus pumila. Int J Mol Sci. 2021;22(9):4410. https://doi.org/10.3390/ijms22094410.
15. Liu W, He G, Deng XW. Biological pathway expression complementation contributes to biomass heterosis in *Arabidopsis*. Proc Natl Acad Sci. 2021;118(16). https://doi.org/10.1073/pnas.2023278118.
16. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2007;9(3):432–41. https://doi.org/10.1093/biostatistics/kxm045.
17. Shahdoust M, Mahjub H, Pezeshk H, Sadeghi M. A Network-based Comparison between Molecular Apocrine Breast Cancer Tumor and Basal and Luminal Tumors by Joint Graphical Lasso. IEEE/ACM Trans Comput Biol Bioinforma. 2019;1. https://doi.org/10.1109/tcbb.2019.2911074.
18. Wu MY, Dai DQ, Zhang XF, Zhu Y. Cancer Subtype Discovery and Biomarker Identification via a New Robust Network Clustering Algorithm. PLoS ONE. 2013;8(6):e66256. https://doi.org/10.1371/journal.pone.0066256.
19. Liu W, Wang Q, Chang J, Bhetuwal A, Bhattarai N, Ni X. Circulatory Metabolomics Reveals the Association of the Metabolites With Clinical Features in the Patients With Intrahepatic Cholestasis of Pregnancy. Front Physiol. 2022;13. https://doi.org/10.3389/fphys.2022.848508.
20. Li Y, Jackson SA. Gene network reconstruction by integration of prior biological knowledge. G3 Genes Genomes Genet. 2015;5(6):1075–9. https://doi.org/10.1534/g3.115.018127.
21. Kapoor R, Datta A, Thomson M. Fused Graphical Lasso Recovers Flowering Time Mutation Genes in Arabidopsis thaliana. Inventions. 2021;6(3):52. https://doi.org/10.3390/inventions6030052.
22. de Abreu e Lima F, Li K, Wen W, Yan J, Nikoloski Z, Willmitzer L, et al. Unraveling lipid metabolism in maize with time-resolved multi-omics data. Plant J. 2018;93(6):1102–15. https://doi.org/10.1111/tpj.13833.
23. Bartzis G, Deelen J, Maia J, Ligterink W, Hilhorst HWM, Houwing-Duistermaat JJ, et al. Estimation of metabolite networks with regard to a specific covariable: applications to plant and human data. Metabolomics. 2017;13(11). https://doi.org/10.1007/s11306-017-1263-2.
24. Zhang X, Pang J, Ma X, Zhang Z, He Y, Hirsch CN, et al. Multivariate analyses of root phenotype and dynamic transcriptome underscore valuable root traits and water-deficit responsive gene networks in maize. Plant Direct. 2019;3(3):e00130. https://doi.org/10.1002/pld3.130.
25. Zheng Y, Wang N, Zhang Z, Liu W, Xie W. Identification of Flowering Regulatory Networks and Hub Genes Expressed in the Leaves of *Elymus sibiricus* L. Using Comparative Transcriptome Analysis. Front Plant Sci. 2022;13. https://doi.org/10.3389/fpls.2022.877908.
26. de Steenhuijsen PW, Heinonen S, Hasrat R, Bunsow E, Smith B, Suarez-Arrabal M, et al. Nasopharyngeal Microbiota, Host Transcriptome, and Disease Severity in Children with Respiratory Syncytial Virus Infection. Am J Respir Crit Care Med. 2016;194:1104–15.
27. Nantongo JS, Potts BM, Davies NW, Fitzgerald H, Rodemann T, O'Reilly-Wapstra JM. Additive genetic variation in Pinus radiata bark chemistry and the chemical traits associated with variation in mammalian bark stripping. Heredity. 2021;127(6):498–509. https://doi.org/10.1038/s41437-021-00476-z.
28. Bo V, Curtis T, Lysenko A, Saqi M, Swift S, Tucker A. Discovering Study-Specific Gene Regulatory Networks. PLoS ONE. 2014;9(9):e106524. https://doi.org/10.1371/journal.pone.0106524.
29. Crombach A, Hogeweg P. Evolution of Evolvability in Gene Regulatory Networks. PLoS Comput Biol. 2008;4(7):e1000112. https://doi.org/10.1371/journal.pcbi.1000112.
30. Tahmasebi A, Ashrafi-Dehkordi E, Shahriari AG, Mazloomi SM, Ebrahimie E. Integrative meta-analysis of transcriptomic responses to abiotic stress in cotton. Prog Biophys Mol Biol. 2019;146:112–22. https://doi.org/10.1016/j.pbiomolbio.2019.02.005.

31. Hollender CA, Kang C, Darwish O, Geretz A, Matthews BF, Slovin J, et al. Floral Transcriptomes in Woodland Strawberry Uncover Developing Receptacle and Anther Gene Networks. Plant Physiol. 2014;165(3):1062–75.
32. Yan Q, Li J, Lu L, Yi X, Yao N, Lai Z, et al. Comparative transcriptome study of the elongating internode in elephant grass (Cenchrus purpureus) seedlings in response to exogenous gibberellin applications. Ind Crop Prod. 2022;178:114653. https://doi.org/10.1016/j.indcrop.2022.114653.
33. Cropano C, Manzanares C, Yates S, Copetti D, Canto JD, Lübberstedt T, et al. Identification of Candidate Genes for Self-Compatibility in Perennial Ryegrass (*Lolium perenne* L.). Front Plant Sci. 2021;12. https://doi.org/10.3389/fpls.2021.707901.
34. Arcioni S, Mariotti D. Selfing and interspecific hybridization in *Lolium perenne* L. and *Lolium multiflorum* Lam. evaluated by phosphoglucoisomerase as isozyme marker. Euphytica. 1983;32(1):33–40. https://doi.org/10.1007/bf00036861.
35. Deniz B, Dogru U. Interspecific hybridisation in Lolium evaluated by morphological genetic markers. New Zealand J Agric Res. 2007;50(3):279–84. https://doi.org/10.1080/00288230709510295.
36. Parsana P, Ruberman C, Jaffe AE, Schatz MC, Battle A, Leek JT. Addressing confounding artifacts in reconstruction of gene co-expression networks. Genome Biol. 2019;20(1). https://doi.org/10.1186/s13059-019-1700-9.
37. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9(1). https://doi.org/10.1186/1471-2105-9-559.
38. Huynh-Thu VA, Sanguinetti G. Gene Regulatory Network Inference: An Introductory Survey. In: Methods in Molecular Biology. New York: Springer New York; 2018. p. 1–23. https://doi.org/10.1007/978-1-4939-8882-2_1.
39. Jiang D, Armour CR, Hu C, Mei M, Tian C, Sharpton TJ, et al. Microbiome Multi-Omics Network Analysis: Statistical Considerations Limitations, and Opportunities. Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.00995.
40. Danaher P, Wang P, Witten D. The joint graphical lasso for inverse covariance estimation across multiple classes. J R Stat Soc Series B Stat Methodol. 2014;76:373–97. https://doi.org/10.1111/rssb.12033.
41. Zeng Z, Zhang S, Li W, Chen B, Li W. Gene-coexpression network analysis identifies specific modules and hub genes related to cold stress in rice. BMC Genom. 2022;23(1). https://doi.org/10.1186/s12864-022-08438-3.
42. Ciannamea S, Jensen CS, Agerskov H, Petersen K, Lenk I, Didion T, et al. A new member of the LIR gene family from perennial ryegrass is cold-responsive and promotes vegetative growth in Arabidopsis. Plant Sci. 2007;172(2):221–7. https://doi.org/10.1016/j.plantsci.2006.08.011.
43. Liu Y, Maniero RA, Giehl RFH, Melzer M, Steensma P, Krouk G, et al. PDX1.1-dependent biosynthesis of vitamin B6 protects roots from ammonium-induced oxidative stress. Molecular Plant. 2022;15(5):820–39. https://doi.org/10.1016/j.molp.2022.01.012.
44. Raschke M, Boycheva S, Crèvecoeur M, Nunes-Nesi A, Witt S, Fernie AR, et al. Enhanced levels of vitamin B6 increase aerial organ size and positively affect stress tolerance in Arabidopsis. Plant J. 2011;66(3):414–32. https://doi.org/10.1111/j.1365-313x.2011.04499.x.
45. Alseekh S, Scossa F, Wen W, Luo J, Yan J, Beleggia R, et al. Domestication of Crop Metabolomes: Desired and Unintended Consequences. Trends Plant Sci. 2021;26(6):650–61. https://doi.org/10.1016/j.tplants.2021.02.005.
46. Degenhardt J. Indirect Defense Responses to Herbivory in Grasses. Plant Physiol. 2009;149(1):96–102. https://doi.org/10.1104/pp.108.128975.
47. Turner MF, Heuberger AL, Kirkwood JS, Collins CC, Wolfrum EJ, Broeckling CD, et al. Non-targeted Metabolomics in Diverse Sorghum Breeding Lines Indicates Primary and Secondary Metabolite Profiles Are Associated with Plant Biomass Accumulation and Photosynthesis. Front Plant Sci. 2016;7. https://doi.org/10.3389/fpls.2016.00953.
48. Guo X, Jahoor A, Jensen J, Sarup P. Metabolomic spectra for phenotypic prediction of malting quality in spring barley. Sci Rep. 2022;12(1). https://doi.org/10.1038/s41598-022-12028-4.
49. Subbaraj AK, Huege J, Fraser K, Cao M, Rasmussen S, Faville M, et al. A large-scale metabolomics study to harness chemical diversity and explore biochemical mechanisms in ryegrass. Commun Biol. 2019;2(1). https://doi.org/10.1038/s42003-019-0289-6.
50. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/a:1010933404324.
51. Probst P, Boulesteix AL, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. J Mach Learn Res. 2019;20(53):1–32.
52. Pembleton LW, Inch C, Baillie RC, Drayton MC, Thakur P, Ogaji YO, et al. Exploitation of data from breeding programs supports rapid implementation of genomic selection for key agronomic traits in perennial ryegrass. Theor Appl Genet. 2018;131(9):1891–902. https://doi.org/10.1007/s00122-018-3121-7.
53. Guo X, Cericola F, Fè D, Pedersen MG, Lenk I, Jensen CS, et al. Genomic Prediction in Tetraploid Ryegrass Using Allele Frequencies Based on Genotyping by Sequencing. Front Plant Sci. 2018;9. https://doi.org/10.3389/fpls.2018.01165.
54. Edwards SM, Buntjer JB, Jackson R, Bentley AR, Lage J, Byrne E, et al. The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet. 2019. https://doi.org/10.1007/s00122-019-03327-y.
55. van der Werf J. Genomic Selection in Animal Breeding Programs. In: Methods in Molecular Biology. Totowa: Humana Press; 2013. p. 543–561. https://doi.org/10.1007/978-1-62703-447-0_26.
56. Yao C, Spurlock DM, Armentano LE, Page CD, VandeHaar MJ, Bickhart DM, et al. Random Forests approach for identifying additive and epistatic single nucleotide polymorphisms associated with residual feed intake in dairy cattle. J Dairy Sci. 2013;96(10):6716–29. https://doi.org/10.3168/jds.2012-6237.
57. Nagy I, Veeckman E, Liu C, Bel MV, Vandepoele K, Jensen CS, et al. Chromosome-scale assembly and annotation of the perennial ryegrass genome. BMC Genom. 2022;23(1). https://doi.org/10.1186/s12864-022-08697-0.
58. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37(8):907–15. https://doi.org/10.1038/s41587-019-0201-4.
59. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5. https://doi.org/10.1038/nbt.3122.
60. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. 2012. arXiv:1207.3907.
61. Jacob D, Deborde C, Lefebvre M, Maucourt M, Moing A. NMRProcFlow: a graphical and interactive tool dedicated to 1D spectra processing for NMR-based metabolomics. Metabolomics. 2017;13(4). https://doi.org/10.1007/s11306-017-1178-y.
62. Meyer TD, Sinnaeve D, Gasse BV, Tsiporkova E, Rietzschel ER, Buyzere MLD, et al. NMR-Based Characterization of Metabolic Alterations in Hypertension Using an Adaptive Intelligent Binning Algorithm. Anal Chem. 2008;80(10):3783–90. https://doi.org/10.1021/ac7025964.
63. VanRaden P. Efficient methods to compute genomic predictions. J Dairy Sci. 2008;91:4414–23. https://doi.org/10.3168/jds.2007-0980.
64. Ashraf BH, Jensen J, Asp T, Janss LL. Association studies using family pools of outcrossing crops based on allele-frequency estimates from DNA sequencing. Theor Appl Genet. 2014;127(6):1331–41. https://doi.org/10.1007/s00122-014-2300-4.
65. Cericola F, Lenk I, Fè D, Byrne S, Jensen CS, Pedersen MG, et al. Optimized Use of Low-Depth Genotyping-by-Sequencing for Genomic Prediction Among Multi-Parental Family Pools and Single Plants in Perennial Ryegrass (*Lolium perenne* L.). Front Plant Sci. 2018;9. https://doi.org/10.3389/fpls.2018.00369.
66. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978;6(2):461-4.
67. Augugliaro L, Mineo AM, Wit EC. ℓ1-Penalized Methods in High-Dimensional Gaussian Markov Random Fields. In: Computational Network Analysis with R. Hoboken: Wiley; 2016.
68. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006;Complex Systems:1695.
69. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Mech. 2008;2008(10):P10008. https://doi.org/10.1088/1742-5468/2008/10/p10008.
70. Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM. 1999;46(5):604–32. https://doi.org/10.1145/324133.324140.
71. Gilmour A, Gogel B, Cullis B, Welham S, Thompson R. ASReml user guide release 4.1 structural specification. Hemel hempstead: VSN international ltd. 2015.

72. Self SG, Liang KY. Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests under Nonstandard Conditions. J Am Stat Assoc. 1987;82(398):605–10. https://doi.org/10.1080/01621459.1987.10478472.
73. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B (Methodol). 1995;57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
74. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13(11):2498–504. https://doi.org/10.1101/gr.1239303.
75. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236–40. https://doi.org/10.1093/bioinformatics/btu031.
76. Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Vesztrocy AW, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep. 2018;8(1). https://doi.org/10.1038/s41598-018-28948-z.
77. Mayer M. missRanger: Fast Imputation of Missing Values. 2021. R package version 2.1.3. https://CRAN.R-project.org/package=missRanger. Accessed 20 Jan 2022.
78. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017;77(1):1–17. https://doi.org/10.18637/jss.v077.i01.
79. Madsen P, Jensen J. A user's guide to DMU-A package for analysing multivariate mixed models. Version 6, Release 5.2. 2013. https://dmu.ghpc.au.dk/dmu/DMU/Doc/Current/. Accessed 20 May 2022.
80. Tripathi S, Dehmer M, Emmert-Streib F. NetBioV: an R package for visualizing large network data in biology and medicine. Bioinformatics. 2014;30(19):2834–6. https://doi.org/10.1093/bioinformatics/btu384.
81. Wickham H. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag New York; 2016. https://ggplot2.tidyverse.org. Accessed 22 Jul 2022.

Publisher's Note