Nitrogen availability during domestication shaped the divergence of tetraploid wheat

Alice Pieri[1], Romina Beleggia[2]*, Tania Gioia[3], Hao Tong[4,5], Valerio di Vittori[1], Giulia Frascarelli[1], Elena Bitocchi[1], Laura Nanni[1], Elisa Bellucci[1], Fabio Fiorani[6], Nicola Pecchioni[2], Stefania Marzario[3], Concetta De Quattro[7], Antonina Rita Limongi[7], Pasquale De Vita[2], Marzia Rossato[7], Ulrich Schurr[6], Jacques L. David[8], Zoran Nikoloski[4,5], Roberto Papa[1]*

[1] Department of Agricultural, Food and Environmental Sciences, Marche Polytechnic University, via Brecce Bianche, 60131, Ancona, Italy

[2] Council for Agricultural Research and Economics (CREA), Research Centre for Cereal and Industrial Crops (CREA-CI), 71122 Foggia, Italy

[3] School of Agricultural, Forestry, Food and Environmental Sciences, University of Basilicata, 85100 Potenza, Italy

[4] Bioinformatics Department, Institute of Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany

[5] Systems Biology and Mathematical Modeling Group, Max Planck Institute of Molecular Plant Physiology, 14476 Potsdam, Germany

[6] Institute of Biosciences and Geosciences (IBG-2): Plant Sciences, Forschungszentrum Julich GmbH, 52428 Julich, Germany

[7] Department of Biotechnology, University of Verona, Strada Le Grazie 15, 37134 Verona, Italy

[8] Montpellier SupAgro, UMR Amelioration Genetique et Adaptation des Plantes, 34060 Montpellier, France

* **Corresponding authors:** r.papa@univpm.it, romina.beleggia@crea.gov.it

Abstract

The domestication of crops is associated with the genome-wide loss of nucleotide diversity caused by selection and genetic drift. Here we investigated 32 genotypes representing key stages in the domestication of tetraploid wheat, i.e. wild emmer, emmer and durum wheat. RNA-Seq analysis was combined with estimates of evolvability, heritability and Q_{ST} to characterize the plasticity of gene expression and its adaptive role in primary and secondary domestication under different nitrogen growth conditions. We confirmed the loss of nucleotide diversity and found that primary and secondary domestication affected gene expression diversity in different ways, influenced by nitrogen availability. We found that nitrogen starvation induced the expression of a larger number of genes in durum wheat compared to emmer and wild emmer. Q_{ST} distributions and Q_{ST}–F_{ST} comparisons revealed distinct selection signatures at each domestication stage. While primary domestication affected the expression of genes involved in biotic interactions, secondary domestication was associated with changes in expression of genes involved in metabolism of amino acids, particularly lysine. Our combined results revealed genes specifically involved in nitrogen metabolism, such as *glutamate dehydrogenase*, with important role in early development, underwent selection during secondary domestication. Therefore, our findings show that nitrogen availability had a pivotal role during the domestication and adaptive responses of one of our major food crops, with varying effects across different traits.

Introduction

Domestication influences the genetic diversity of animals and plants as they adapt to new agroecosystems and undergo selection to meet human preferences and needs. This process is typically associated with the genome-wide loss of nucleotide diversity due to the combined

consequences of selection and genetic drift, which is known as the domestication bottleneck. The loss of genetic diversity has been documented in many domesticated species by comparing them with wild relatives[1]. A parallel effect is the reprogramming of gene expression and the loss of expression diversity, which was first reported in the common bean (*Phaseolus vulgaris*)[2] and subsequently in other domesticated plants and animals[3–5]. Similar observations have been reported at the level of metabolic diversity[6,7].

Changes in nucleotide and gene expression diversity during the domestication of tetraploid wheat (*Triticum turgidum* L., 2n = 4x = 28; AABB genome) are not fully understood. Evidence indicates that domestication occurred in two well-defined phases: Primary domestication from wild emmer (*Triticum turgidum* ssp. *dicoccoides*) to emmer (*Triticum turgidum* ssp. *dicoccum*) started ~12,000 years ago in the Fertile Crescent. This was followed by secondary domestication from emmer to durum wheat (*Triticum turgidum* ssp. *durum*), which started ~2,000 years ago in the Near East and gave rise to durum wheat, the most important form of tetraploid wheat and currently the most widespread Mediterranean crop[8,9].

The molecular mechanisms underlying phenotypic plasticity in crops[10] and their wild relatives must be understood to address the challenges faced by modern agriculture, including the overreliance on nitrogen (N) fertilizers to meet the growing global food demand. N is an essential macronutrient whose availability is directly linked to crop yield and grain quality (protein content)[11–13]. Understanding genetic variations in N acquisition, assimilation and metabolism can therefore provide novel strategies for crop improvement[14,15]. In tetraploid wheat, phenotypic differences related to N availability primarily arose during secondary domestication rather than primary domestication[8], but the relationship between N metabolism and changes in gene expression plasticity during domestication is unclear.

Here we analyzed 32 wild emmer, emmer and durum wheat genotypes by RNA-Seq to determine how contrasting differences in N availability shaped the nucleotide and gene

expression diversity of tetraploid wheat during primary and secondary domestication. Our results provide insight into the pivotal role of N during the domestication and adaptive responses of one of our major food crops.

Results and discussion

A greater loss of nucleotide diversity occurred during the secondary domestication of tetraploid wheat

We prepared 128 RNA-Seq libraries from 4-week-old leaves of 32 tetraploid wheat genotypes representing *T. turgidum* ssp. *dicoccoides*, ssp. *dicoccum* and ssp. *durum* (Supplementary Table S1). On average, 6.8 million of reads per genotype (Supplementary Table S1) were mapped to the A and B reference subgenomes of bread wheat[16]. The mapping frequency exceeded 85% for all the three subspecies and the fraction of reads mapping to gene regions exceeded 72% (Supplementary Table S1).

Variant calling produced 800,996 high-quality single-nucleotide polymorphisms (SNPs). The number of polymorphic sites was similar in wild emmer (617,128) and emmer (613,509), but was much lower in durum wheat (425,513), confirming the higher genetic diversity of the wild population. We identified 190,377 common SNPs shared by all three taxa. As expected, wild emmer and emmer shared the highest percentage of SNPs (33%, 206,578). In contrast, durum wheat shared only 11% (46,352) of its SNPs with wild emmer and 17% (71,147) with emmer.

SNPs principal component analysis (PCA) revealed the broad genetic structure of the three wheat taxa (Figure 1) and confirmed that secondary domestication had a greater impact than primary domestication in differentiating the durum wheat subspecies. The analyzed 12 durum wheat genotypes are genetically very similar, forming a dense cluster that is clearly distinguishable from the wild emmer and emmer genotypes. In contrast, the wild emmer and

emmer genotypes were loosely clustered, indicating a greater genetic admixture. These results are consistent with previous genetic studies on the origins of domesticated wheat and reflect the multiple stages of domestication[17–19], and indicate that the used genotypes are representative.

Nucleotide diversity estimates (Table 1) confirmed the expected substantial loss of nucleotide diversity during domestication. The average nucleotide diversity of durum wheat was ~35% lower than domesticated emmer, which was in turn ~11% lower than wild emmer, confirming the greater impact of secondary domestication. When the cumulative effect of primary and secondary domestication is taken into account, we observed a ~42% reduction in the nucleotide diversity of durum wheat compared to its wild ancestor (Table 1). We observed a parallel increase in Tajima's D from wild emmer (D = 0.26) to emmer (D = 0.46) and durum wheat (D = 1.08), consistent with the bottlenecks of primary and secondary domestication (Table 1)[1].

The variability of gene expression during domestication was influenced by N availability

To quantify the diversity of gene expression in each subspecies, we calculated evolvability scores under high and low N availability conditions. Evolvability was estimated using the additive coefficient of variation (CV_A) in read counts (Supplementary Table S2). In contrast to heritability, CV_A is a standardized measure of additive genetic variation that is not influenced by other sources of variance[20,21], and is therefore well suited for comparative analysis[22]. As for nucleotide diversity, we found that the CV_A decreased during domestication under both N conditions; however, the mean CV_A of all three subspecies was higher under low N conditions (Figure 2a,b; Table 2). Optimal N availability therefore appears to promote a more uniform gene expression pattern, whereas higher variability is observed during N starvation. The association between domestication and declining diversity in gene expression has also been

reported in crops, such as: common bean[2], tomato[3] and sorghum[5] as well as domesticated animal species[4].

We used the contrasting N conditions of our samples to examine whether the loss of expression diversity is associated with the specific aspects of the cultivation environment, causing primary and secondary domestication to have a significantly different impact. Under optimal N conditions, we observed a ~9% loss in expression diversity in emmer compared to wild emmer (effect of primary domestication) and a ~15% loss in durum wheat compared to emmer (effect of secondary domestication). In contrast, these losses were ~18% and 11% under N starvation conditions, revealing twice the loss of expression diversity during primary domestication, but a lower value during secondary domestication (Table 2). All four values differed significantly from each other (Mann–Whitney test, $p < 0.001$). The opposing expression diversity profiles during domestication under high and low N conditions were observed not only for overall gene expression, but also for the subgroup comprising all differentially expressed genes (DEGs) and the subgroup comprising all unmodulated genes (Supplementary Table S3). The loss of expression diversity among the DEGs due to primary domestication was ~9% and ~15% under high and low N conditions, respectively, whereas the loss due to secondary domestication was ~18% and ~14% under high and low N conditions, respectively (Supplementary Table S3). The loss of expression diversity among the unmodulated genes was similar to the values for overall gene expression (Supplementary Table S3).

A phenotypic study of the same accessions used in the present work has already shown that secondary domestication reduced the phenotypic diversity under optimal N conditions, but the reduction was smaller and not significant under N starvation conditions[8]. In the case of durum wheat, selection has apparently enhanced the growth response to N availability, indicating a putative focus on improving N uptake and utilization efficiency. Our expression diversity

results support the conclusions of the earlier study[8], suggesting that selection has favored specific traits and thus led to a more uniform set of cultivars.

Domestication and nitrogen availability shaped the divergence of tetraploid wheats

Genetic differentiation among the three subspecies was estimated by calculating the pairwise fixation index (F_{ST}) for every gene locus in our dataset. As shown in Figure 3a, the lowest genetic differentiation was observed between wild emmer and emmer (mean F_{ST} = 0.09), whereas much higher genetic differentiation was found between emmer and durum wheat (mean F_{ST} = 0.27) and, similarly, between wild emmer and durum wheat (mean F_{ST} = 0.28). These values align with earlier findings that examined broad collections of tetraploid wheat accessions[17,23], and provide additional evidence for the representativeness of the genotypes used.

Divergence at the transcriptomic level was estimated by calculating Q_{ST}, the quantitative analog of F_{ST}, taking N availability into account as an environmental variable. Under both N conditions, we observed the same trend shown for F_{ST} (Figure 3b). Specifically, secondary domestication had a stronger impact on differentiation (emmer *vs* durum wheat, mean $Q_{ST\ LN}$ = 0.23, mean $Q_{ST\ HN}$ = 0.33) than primary domestication (wild emmer *vs* emmer, mean $Q_{ST\ LN}$ = 0.16, mean $Q_{ST\ HN}$ = 0.23). Interestingly, the Q_{ST} distributions of every pairwise comparison showed higher values under optimal N conditions compared to N starvation (Figure 3b), suggesting that N availability during domestication may have contributed to the differentiation of gene expression in the three subspecies.

The Q_{ST} distributions were used to perform a "selection scan", seeking genes whose expression was potentially under selection. Starting from 5,868 genes meeting the heritability criteria ($H^2 \geq 0.7$ or $S \times N \geq 0.2$, that is the species × environment variance component i.e., every species

subgroup × N condition; Supplementary Figure S1), we retained 972 genes having Q_{ST} values in the 5% right tail of the distributions. The Q_{ST} *vs* F_{ST} comparison method[24] was then used to confirm that the divergent expression (high Q_{ST} values) of the filtered genes was caused by directional selection ($Q_{ST} > F_{ST}$) and not by genetic drift ($Q_{ST} \approx F_{ST}$) or stabilizing selection ($Q_{ST} < F_{ST}$)[24]. After removing F_{ST} values < 0.01, we retained 967 genes satisfying the criterion $Q_{ST} > F_{ST}$, indicating that their expression was likely subjected to directional selection in at least one of the evolutionary contexts examined herein (i.e., primary and/or secondary domestication under high and/or low N availability conditions) (Supplementary Table S4).

Gene Ontology (GO) enrichment analysis revealed that selection acted on distinct gene categories during primary and secondary domestication (Supplementary Figure S2). During primary domestication, we found categories associated with "defense-related programmed cell death, modulated by biotic interactions", indicating an enhanced plant hypersensitive response to pathogens. This can be interpreted as a consequence of the transition from the natural growing environment of the wild genotypes to agroecosystems characterized by high-density domesticated crop monocultures. In this context, crops face higher disease pressure from crop-specific pathogens[25] and therefore induce a hypersensitive response, which can lead to programmed cell death and necrosis as a defense mechanism. It is important to note that pathogen defense mechanisms in plants often overlap with the regulation of beneficial symbiotic interactions, therefore, one expects a trade-off between traits associated with symbiosis and innate immunity[26]. Moreover, domesticated crops are less able to fully benefit from microbial interactions than their wild relatives, as observed in a comparative study of bread wheat landraces as well as old and modern varieties[27]. One contributing factor is the widespread use of high-input agricultural practices, because the availability of fertilizers reduces the need for plants to invest in symbiotic relationships[28]. Additionally, certain target

traits in plant breeding, such as phytohormones that regulate flowering time and plant height, can have unintended effects on beneficial symbiosis due to pleiotropy[29].

Among the genes found to be under selection during secondary domestication, we observed the enrichment of categories associated with amino acid metabolism, particularly those related to the "lysine catabolic process" (Supplementary Figure S2). This included genes encoding the bifunctional enzyme lysine ketoglutarate reductase/saccharopine dehydrogenase (LKR/SDH). This enzyme is ubiquitous in plants and animals, and represents the key step in lysine catabolism via the saccharopine pathway (SACPATH). The structure and transcription of the *LKR/SDH* gene has been studied in *T. durum* and compared with other plants, showing species-dependent differences in expression levels including lineage-specific differences between monocots and dicots[30]. Lysine is the first limiting essential amino acid in cereal grains and its catabolic pathway has been targeted to increase the lysine content of maize and rice seed[31–33]. Generally, the quantity of lysine-containing proteins in cereal seeds is much lower than that of storage proteins devoid of lysine, such as prolamins (specifically gliadin in wheat). The SACPATH seems to channel the lysine skeleton into the production of glutamic acid, which is a precursor of proline, one of the most abundant amino acids in glutens[34].

General changes in amino acid metabolism during domestication have been observed in other crops based on nucleotide data, including sunflower[35], maize[36] and common bean[2]. Evolutionary metabolomics has also revealed signatures of selection affecting amino acid metabolism during secondary domestication[6]. In durum wheat, domestication was linked to the selection of a specific protein composition and led to a notable decrease in the diversity of gliadin and glutenin subunits, strongly correlating with grain yield and the technological properties of gluten[13]. The analysis of spring wheat genotypes has shown that the SACPATH is upregulated in response to drought stress, and is significantly more active in drought-tolerant compared to drought-susceptible genotypes[37]. This may reflect the role of proline, which can

be produced from this pathway, as a major constituent of storage proteins and one of the main osmoprotectants produced as a response to stress[38]. These findings suggest that selection for stress-tolerant genotypes as well as seed protein composition during wheat domestication may have influenced the expression of SACPATH genes.

Changes in nitrogen availability trigger the expression of many genes, especially in durum wheat

We identified DEGs in each subspecies that discriminated between optimal N conditions and N starvation using a stringent pipeline and strict thresholds (p-adjust < 0.001) to reduce the number of false positives. We found 3,326 DEGs in wild emmer, 3,305 in emmer and 5,901 in durum wheat, with more upregulated than downregulated genes in all three subspecies. Durum wheat had the highest percentage of private DEGs (~42%, 2,479), whereas similar numbers were found in wild emmer (~14%, 458) and emmer (~15%, 486). Wild emmer and emmer shared ~23% (749) and ~21% (700), respectively, of their DEGs with durum wheat. The percentage of DEGs shared only between wild emmer and emmer was 4% (146), but almost 60% of wild emmer and emmer DEGs and ~33% of durum wheat DEGs were shared by all three taxa (Figure 4a). The proportions of private and shared DEGs were preserved when we separated them into upregulated and downregulated subsets (Figure 4b,c). . In all three taxa, most DEGs were located on chromosomes 2A, 2B, 3A, 3B, 5A and 5B, each carrying > 7.5% of the DEGs; in contrast, chromosomes 6A and 6B each contained only ~5% of the DEGs Supplementary Figure S3.

GO enrichment analysis of the DEGs meeting the threshold FDR < 0.05 revealed 23 macro-categories in wild emmer, 21 in emmer and 25 in durum wheat (Supplementary Figure S4). The main differences between the three subspecies were observed for categories related to

"signaling", "regulation of biological process", "developmental process", and "metabolic process" Supplementary Table S5. We observed the uniform enrichment of GO categories associated with upregulated genes in all three subspecies, including terms linked to N and amino acid metabolism as well as carbon metabolism and photosynthesis (Supplementary Table S5). In contrast, the enrichment of GO categories associated with downregulated genes was more selective, with some GO categories related to N metabolism enriched only in durum wheat, including GO:0006807 and GO:0034641 (N compound and cellular N compound metabolic process, respectively) and GO:0006536 "glutamate metabolic process" (Supplementary Table S5). Functional annotations of the most strongly modulated genes (top 5% $|\log_2 FC|$ values) are reported in Supplementary Table S6.

Our data confirm, on a larger set of samples, earlier observations on the response of wheat to N starvation based on transcriptomics and metabolomics data[36–38]. These earlier studies included one emmer and one durum wheat genotype also present in our sample set[39], but also considered the durum wheat cultivar Svevo[40] and various bread wheat cultivars[41]. As expected, genes involved in N metabolism were modulated during N starvation. Among the key genes for N assimilation, those encoding asparagine synthetase and nitrite reductase were upregulated in every taxon, whereas those encoding glutamate carboxypeptidase and glutamate decarboxylase were downregulated. We observed contrasting profiles for genes encoding ureide permease (encoding a ureide transporter), which were strongly upregulated in all three subspecies in response to N stress, whereas genes encoding nitrate transporters were strongly downregulated. The modulated genes also included transporters of amino acids and other nutrients.

N starvation also influenced other metabolic pathways, revealing many further DEGs involved in C metabolism, especially fatty acid metabolism, glycolysis, photosynthesis and the tricarboxylic acid (TCA) cycle. About 10% of the highest-ranking DEGs represented

transcription factors and protein kinases. The most common functional category (accounting for 17% of annotated DEGs) reflected the general stress response to N starvation, including the mitigation of oxidative stress and detoxification. Examples included genes encoding cytochrome P450s, glutaredoxin family proteins, glutathione S-transferases and peroxidases (Supplementary Table S6).

To compare gene expression between the three taxa while taking the environmental effects into account, we also identified DEGs between each pair of subspecies under all N conditions. Accordingly, we compared emmer *vs* wild emmer (primary domestication, high and low N), durum wheat *vs* emmer (secondary domestication, high and low N) and durum wheat *vs* wild emmer (cumulative effect, high and low N) (Supplementary Figure S5). The wild emmer *vs* emmer comparison revealed few DEGs regardless of N availability (12 and 11 DEGs under high and low N conditions, respectively), whereas the emmer *vs* durum wheat comparison revealed 41 DEGs associated with optimal N and 29 associated with N starvation, and the wild emmer *vs* durum wheat comparison revealed 46 DEGs associated with optimal N and only 10 associated with N starvation. These data indicate that the number of DEGs increases during domestication but only when there is a sufficient N supply (Supplementary Figure S5). Interestingly, there were more upregulated than downregulated genes in all pairwise comparisons under optimal N conditions (~65%) but the proportion increased under N starvation conditions, particularly for the comparison of wild emmer *vs* durum wheat (90%). The preponderance of upregulated genes during domestication has also been observed in maize[42], whereas domestication was shown to increase the proportion of downregulated genes in common bean[2], egg-plant[43] and sorghum[5] landraces compared to wild relatives. The absence of consistent patterns suggests that the evolution of domesticated phenotypes is driven by specific processes that are unique to each crop.

Among the 102 DEGs (Supplementary Table S7) found in at least one of the six pairwise comparisons between subspecies, 35 were also found among DEGs identified between contrasting N conditions and of which 24 were proposed to be under selection. Overall, six genes were identified in all three experiments (i.e. differentially expressed between subspecies and between contrasting N conditions, and showed evidence of selection).

Selection shaped the expression profiles of genes modulated by nitrogen availability

The 6,991 DEGs found in at least one species when comparing the contrasting N conditions included 101 apparently under selection, which are candidates for the adaptive response to N availability. We applied PCA to the normalized read counts in order to investigate if the different genotype groups can be separated based on their gene expression. Initially we incorporated all 6,991 DEGs (Figure 5a,b) before focusing on the subset of 101 DEGs that were also apparently under selection (Figure 5c,d). When considering all DEGs, PC1 did not completely separate the durum wheat genotypes from the other taxa, in contrast to the clear separation observed for the SNP data (Figure 1), and this was particularly evident during N starvation (Figure 5b). There was also a moderate degree of overlap between the wild emmer and emmer genotypes along PC2. However, when we focused on the DEGs under selection, PC1 separated the durum wheat genotypes into a densely clustered group (as observed for the SNP data) under both N conditions, and PC2 separated the wild emmer and emmer genotypes more clearly, especially under high N conditions (Figure 5c,d).

The integration of selection signatures (based on Q_{ST}/F_{ST} values) and differential expression analysis uncovered a set of 101 candidate genes that are interesting due to their potential roles in the domestication and diversification of cultivated wheat, specifically in relation to N availability. Functional annotation (Supplementary Table S8) revealed upregulated genes

involved in carbon (C) metabolism as well as some transcription factors and transporters, as well as both upregulated and downregulated genes responsible for general stress responses and N metabolism, specifically those encoding enzymes involved in amino acid metabolism, such as methionine aminopeptidase, aspartokinase and glutamate dehydrogenase (GDH). The latter is particularly noteworthy because, in addition to its modulation in response to different N conditions and the presence of selection signatures, it was also upregulated in the comparison between wild emmer and durum wheat under optimal N conditions. GDH is a key enzyme involved in N metabolism and N/C balance[44]. This is supported by the co-localization of quantitative trait loci for GDH activity and physiological traits associated with the flag leaf lamina, such as soluble protein and amino acid content, as well as flag leaf area and dry weight[45]. Selection signatures were also identified in the *GDH* gene when comparing landraces with old and modern durum wheat cultivars[9]. Our results confirm that N metabolism has been a key driver during the evolutionary history of wheat, particularly the central role of glutamate in the process of domestication. This was also suggested by a combined transcriptomics and metabolomics study, showing that glutamate and γ-aminobutyric acid (mainly synthetized from glutamate) are central to the genotype-specific response of emmer and durum wheat to N starvation[39].

Conclusion

We have shown that significant changes occurred at the nucleotide and gene expression levels during the domestication of tetraploid wheat, taking into account the environmental variable of N availability. We confirmed that more nucleotide diversity has been lost during secondary domestication compared to primary domestication, and revealed a parallel trend in the loss of gene expression diversity associated to the domestication process, with a stronger effect due to secondary domestication and unveil a parallel different impact of primary and secondary

domestication on the loss of expression diversity, which appears to be related to N availability in the durum wheat selection environment. We present evidence that selection may have operated in different directions during primary and secondary domestication, the former involving changes related to biotic interactions and the latter related to amino acid metabolism. By screening a larger number of genotypes, we reinforce the results of earlier studies[39] indicating a stronger transcriptional response in durum wheat (compared to emmer and wild emmer) to changes in N availability. Finally, through the innovative combination of RNA-Seq analysis and the estimate of quantitative genetics parameters, we developed a pipeline to identify selection signatures and phenotypic plasticity in gene expression data based on evolvability and Q_{ST}/F_{ST} scores. Our findings, elucidating the role of N in tetraploid wheat domestication and adaptive response can guide the development of innovative strategies for crop improvement, resource use efficiency, and environmental sustainability.

Although we demonstrated the validity of our panel of genotypes using leaf tissue, future studies should consider expression data from other tissues to refine the functional analysis of genes in the domestication process. However, pan-genome assembly strategies incorporating a broader range of lines and accounting for more variation[46], could reduce the potential for bias. Future studies on the domestication of wheat may also benefit from the inclusion of additional types of genetic variation (e.g., presence/absence variation) to gain more insight into the genetic factors involved in the adaptation of this important crop.

Materials and methods

Plant material and experimental design

The study included 32 tetraploid wheat genotypes, comprising 10 accessions of wild emmer (*T. turgidum* ssp. *dicoccoides*), 10 accessions of emmer (*T. turgidum* ssp. *dicoccum*), and 12

accessions of durum wheat (*T. turgidum* ssp. *durum*) (Supplementary Table S1). The samples we analyzed were part of a larger experiment, conducted in October 2012 and described elsewhere[8]. Briefly, wheat genotypes were grown for 4 weeks under optimal nitrogen (N+) and nitrogen starvation (N–) conditions in the Phytec Experimental Greenhouse at the Institute of Biosciences and Geosciences (IBG-2), Plant Sciences Institute, Forschungszentrum Jülich GmbH, Germany (50°54′36″ N, 06°24′49″ E). Seeds of uniform size and mass were visually selected, surface sterilized (1% (*w/v*) NaClO for 15 min) and pre-germinated. After germination, seedlings showing uniform growth (seminal root length, 1–2 cm) were transferred to soil-filled rhizoboxes, which were placed in the automated GROWSCREEN-Rhizo phenotyping system available at IBG-2. We used a Type 0 manually sieved peat soil (Nullerde Einheitserde; Balster Einheitserdewerk, Frondenberg, Germany), which provided low nutrient availability (ammonium N and nitrate N concentrations of < 1.0 and < 1.0 mg l^{-1}, respectively). All plants were watered twice daily with 400 ml of tap water and were supplied three times per week with 200 ml of modified Hoagland solution[47] with or without added N. For the N starvation conditions, KNO_3 and $Ca(NO_3)_2$ were replaced with K_2SO_4 and $CaCl_2{\cdot}6(H_2O)$, respectively. The experiment was carried out under natural lighting in the greenhouse, with an air temperature of 18–24 °C and a relative humidity of 40–60%. For each N treatment, we used two replicates of each genotype with two plants per replicate (four plants per genotype in total). After 4 weeks, leaves were pooled from two plants of the same genotype growing in the same rhizobox. Accordingly, four independent biological replicates (two replicates per N condition) were produced for each genotype, with the exception of wild emmer IG 46504, PI 233288, PI 466991, PI 538656, emmer MG 5293/1, and durum wheat Creso, Pedroso and Trinakria, for which only three replicates were available, and emmer Molise Sel. Colli and durum wheat Simeto, for which eight replicates were available. The tissues were immediately frozen in liquid

N_2 and stored at −80 °C. Further details of the experiment and growth conditions are provided elsewhere[8].

RNA extraction and sequencing

RNA was extracted from 100 mg of frozen ground leaves per replicate using the Spectrum Plant Total RNA kit (Sigma-Aldrich, St Louis, MO, USA) followed by treatment with RNase-free DNase using the On-Column DNase I Digestion Set (Sigma-Aldrich). RNA integrity and purity were assessed by agarose gel electrophoresis and a Bioanalyzer 2100, respectively (Agilent/Bonsai Technologies, Santa Clara, CA, USA). Only RNA samples with an RNA integrity number > 8.0 were considered suitable for analysis.

Library construction and RNA sequencing were carried out using the Illumina mRNA-Seq platform at the Montpellier Genomix sequencing facility (http://www.mgx.cnrs.fr) as previously described[48]. Briefly, RNA samples were processed using TruSeq RNA sample preparation kits v2 (Illumina, San Diego, CA, USA). Libraries were quantified by real-time PCR using the KAPA Library Quantification Kit for Illumina Sequencing Platforms (Roche, Basel, Switzerland), followed by quality control using a DNA 100 Chip on a Bioanalyzer 2100. Cluster generation and sequencing were carried out using the Illumina HiSeq 2000 instrument and TruSeq PE Cluster Kit v3, following the Illumina PE_Amp_Lin_Block_V8.0 recipe, and Illumina TruSeq PE Cluster v3-cBot-HS kits with the 2 × 100 cycles, paired-end, indexed protocol, respectively[48].

RNA-Seq library processing and mapping

We pre-processed 128 raw paired-end RNA-Seq libraries[48]. Cutadapt[49] was then used to remove adaptor sequences and trim the end of reads with low quality scores (parameter -q 20) while keeping reads with a minimum length of 35 bp. Reads with a mean quality score < 30

were discarded, and orphan reads (whose mates were discarded in the previous filtering steps) were removed[48]. The final quality of trimmed and filtered reads was assessed using FastQC[50].

The bread wheat (*Triticum aestivum* cv. Chinese Spring) genome assembly IWGSC RefSeq v2.1, along with the corresponding genome annotation, were downloaded from the IWGSC data repository hosted by URGI-INRAE (https://wheat-urgi.versailles.inra.fr/) and used as a reference to map each cleaned library to the A and B sub-genomes. The bread wheat genome was chosen deliberately to ensure the inclusion of an outgroup species that is closely related to the subspecies in the panel. By doing so, we aimed to avoid bias that could arise from selecting only one subspecies among our panel of accessions. We have confidence in this strategy because the *T. aestivum* A and B subgenomes are derived from the tetraploid species included in the study.

STAR v2.7.0e[51] was used for read mapping with the --quantMode TranscriptomeSAM and --quantTranscriptomeBan Singleend options. The output alignments were translated into transcript coordinates (in addition to alignments in genomic coordinates), allowing insertions, deletions and soft-clips in the transcriptomic alignments. The transcriptomic alignments were used as inputs for salmon v1.6.0[52] to quantify gene expression. Raw read counts were computed for all genes in each sample and, to filter out weakly-expressed transcripts, only genes with at least 1 count per million (CPM) in at least 10 samples (of the same subspecies) were retained. This was calculated separately in each of the three subspecies and the raw counts of the filtered genes in each subspecies were then combined for downstream analysis, for a total of 32,358 genes (Supplementary Table S2).

Variant identification

Variants were called by applying BCFtools v1.15 (previously SAMtools)[53] to the alignment bam files. The "*bcftools mpileup*" command was used to determine the genotype likelihoods at

each genomic position, with a minimum alignment quality of 20 and a minimum base quality of 30. The actual calls were obtained using the "*bcftools call*" command. The resulting VCF file was filtered using the "*bcftools view*" command, removing indels and keeping only sites covered by at least three reads in all genotypes. Subsequently, only biallelic SNPs with maximum values of 50% missingness and a 1% minor allele frequency were retained. To identify private and shared SNPs among the different subspecies, every possible comparison of the three subsampled VCF files (wild emmer, emmer and durum wheat) was carried out using the "*bcftools isec*" command.

Population genetics analysis

Variants were filtered (one SNP per 500 kb) using the VCFtools *--thin 500000* option (v0.1.17)[54] and then converted into ped format with PLINK (v1.90p)[55]. PLINK was also used to compute genetic distances between individuals with the *--distance-matrix* flag. The output matrix was used as input for PCA with the *cmdscale* function of R (v4.2.1)[56].

Genetic diversity statistics, including nucleotide diversity (π and θ)[57,58] and Tajima's D (D)[59] were computed on the alignment bam files for each subspecies, from the folded site frequency spectra using ANGSD[60]. First, the *doSaf* function was used to estimate per-site allele frequencies (Saf) then *realSFS* was used to get the site frequency spectra. The statistical loss of diversity[61] was used to test the impact of primary and secondary domestication on the molecular diversity of the three subspecies. For primary domestication, the statistic was computed as $[1 - (x_{emmer}/x_{wild})]$, where x_{emmer} and x_{wild} are the diversities in emmer and wild emmer, respectively, measured using π, θ and D. If x_{emmer} was higher than x_{wild}, then the parameter was calculated as $[(x_{wild}/x_{emmer}) - 1]$. The loss of diversity due to secondary domestication in durum wheat versus emmer was calculated as $[1 - (x_{durum}/x_{emmer})]$, where x_{durum} and x_{emmer} are the diversities in durum wheat and emmer, respectively. If x_{durum} was higher than x_{emmer}, then the parameter was calculated as $[(x_{emmer}/x_{durum}) - 1]$.

We calculated F_{ST} for each pair of populations using ANGSD[60]. Saf and 2D SFS were calculated as for nucleotide diversity, then the *fst index* function was used to obtain the global estimate. To get an F_{ST} value for each gene in our dataset, we used the *fst print* function, which prints the posterior expectation of genetic variance between populations (called A), and total expected variance (called B) for every locus. We then computed the weighted F_{ST} as the ratio of the summed As and summed Bs for every gene region, using an *ad hoc* R script.

Expression profiles, heritability and Q_{ST} analysis

Raw read counts of the 32,358 genes were normalized using the *vst* method allowing the additive coefficient of variation (CV_A) (standard deviation/mean) to be calculated for the two N conditions in every subspecies, averaging the biological replicates of every genotype. The statistical loss approach[61] was then applied to test the loss of expression diversity in the different groups, as previously reported[2]. The statistical significance of the differences between each CV_A value and the percentage loss of expression diversity was determined using the Mann-Whitney test in R (v4.2.1)[56] with the function *wilcox.test*.

To compute heritability, the raw counts of each subspecies under each condition were first normalized using the trimmed mean M-values normalization method in the R package edgeR[62] and the voom normalization method in the R package limma[63]. To determine the variance component of each factor and heritability, the following model was considered:

$$Y_{ijkl} = S_i + G_{j(i)} + N_k + (S \times N)_{ik} + (G \times N)_{jk(i)} + \varepsilon_{l(ijk)},$$

where Y_{ijkl} is the normalized gene expression level, S_i is the species factor, $G_{j(i)}$ is the genotype factor nested in species, N_k is the N level factor, $(S \times N)_{ik}$ is the interaction between species and N levels, $(G \times N)_{jk(i)}$ is the interaction between genotypes and N levels, and $\varepsilon_{l(ijk)}$ is the residual error. All factors were treated as random effects in the model except the intercept, which was a fixed effect. The linear mixed models were fitted using the *lmer* function in R

package lme4 based on the normalized data of each transcript[64]. The heritability (H^2) was calculated as $H^2 = \frac{V_S+V_G}{V_A}$, where $V_A = V_S + V_G + V_N + \frac{V_{S\times N}}{n} + \frac{V_{G\times N}}{n} + \frac{V_\varepsilon}{n}$, V_S is the variance of species, V_G is the variance of genotype, V_N is the variance of N level, $V_{S\times N}$ is the variance of species and N level interaction, $V_{G\times N}$ is the variance of genotype and N level interaction, V_ε is the residual variance, and n is the number of N levels. $V_{S\times N}$ and $V_{G\times N}$ represent the genotype × environment interaction variance components at the species and genotype (nested in species) levels, respectively.

Q_{ST} was calculated between pairs of the three subspecies under low and high N levels separately. The wild emmer *vs* emmer comparison revealed the effects of primary domestication, the emmer *vs* durum wheat comparison revealed the effects of secondary domestication, and the wild emmer *vs* durum wheat comparison revealed the cumulative effect of domestication. To this end, the model can be reduced to $Y_{ijl} = S_i + G_{j(i)} + \varepsilon_{l(ij)}$ at each N level. The Q_{ST} value was calculated as $Q_{ST} = \frac{V_S}{V_S+V_G}$, the ratio of between-species and within-species variance.

Q_{ST} distributions were used to perform a "selection scan" on a restricted number of genes. First, genes were filtered for $H^2 \geq 0.7$ and, in order not to lose genes whose expression was strongly influenced by N availability, also the species × environment ($S\times N$) variance component was evaluated (i.e., every species subgroup × N condition), retaining those genes meeting the threshold $S\times N \geq 0.2$ (Supplementary Figure S1). Successively, we obtained six different Q_{ST} value distributions ($Q_{ST\ \text{WILD EMMER}\ vs\ \text{EMMER}}$, $Q_{ST\ \text{EMMER}\ vs\ \text{DURUM WHEAT}}$ and $Q_{ST\ \text{WILD EMMER}\ vs\ \text{DURUM WHEAT}}$, each for high and low N conditions) and we retained the 5% upper tail of every distribution. Finally, we compared F_{ST} and Q_{ST} values for every gene, discarding F_{ST} values < 0.01. We confirmed that every retained gene satisfied the condition $Q_{ST} > F_{ST}$ allowing it to be classed as undergoing directional selection.

Differential expression analysis

Differential gene expression was assessed by analyzing the pre-processed raw count dataset (32,358 genes). We identified DEGs by comparing (*i*) two conditions (i.e., high and low N levels) within each subspecies, and (*ii*) pairs of the three subspecies under the same N levels, which considered the genotypes nested in species. For the two scenarios, we used three different approaches to detect DEGs: one linear model-based approach implemented in the R package limma[63], and two Poisson model-based approaches implemented in the R packages edgeR[62] and DESeq2[65]. In all approaches, the normalization of raw counts was applied by default in the package before differential analysis. To reduce the number of false positives, the intersection of DEGs resulting from the three approaches was retained[66,67] and the significance threshold was set to an adjusted p-value < 0.001. The DEGs between high and low N levels in at least one subspecies were used for PCA following the *DESeq2* approach[65], first using all the DEGs, then repeating the analysis on the DEGs considered to be under selection. At each step, counts were normalized using the *vst* method before the *plotPCA* function was applied to define principal components 1 and 2 for the two N levels separately.

GO enrichment analysis

Enriched terms in the DEGs and genes under selection were identified using agriGO (v.2.0)[68] with *T. aestivum* reference annotations and the following parameters: hypergeometric test, multiple hypothesis test adjustment according to the Hochberg FDR procedure at significance level < 0.05 and minimum number of mapping entries of 3.

References

1. Bitocchi, E. *et al.* Beans (Phaseolus ssp.) as a model for understanding crop evolution. *Front. Plant Sci.* **8**, 722 (2017).

2. Bellucci, E. *et al.* Decreased Nucleotide and Expression Diversity and Modified Coexpression Patterns Characterize Domestication in the Common Bean. *Plant Cell* **26**, 1901–1912 (2014).

3. Sauvage, C. *et al.* Domestication rewired gene expression and nucleotide diversity patterns in tomato. *Plant J.* **91**, 631–645 (2017).

4. Liu, W. *et al.* Decrease of gene expression diversity during domestication of animals and plants. *BMC Evol. Biol.* **19**, 1–11 (2019).

5. Burgarella, C. *et al.* The Road to Sorghum Domestication: Evidence From Nucleotide Diversity and Gene Expression Patterns. *Front. Plant Sci.* **12**, 666075 (2021).

6. Beleggia, R. *et al.* Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. *Mol. Biol. Evol.* **33**, 1740–1753 (2016).

7. Alseekh, S. *et al.* Domestication of Crop Metabolomes: Desired and Unintended Consequences. *Trends Plant Sci.* **26**, 650–661 (2021).

8. Gioia, T. *et al.* Impact of domestication on the phenotypic architecture of durum wheat under contrasting nitrogen fertilization. *J. Exp. Bot.* **66**, 5519–5530 (2015).

9. Taranto, F. *et al.* Whole Genome Scan Reveals Molecular Signatures of Divergence and Selection Related to Important Traits in Durum Wheat Germplasm. *Front. Genet.* **11**, 217 (2020).

10. Laitinen, R. A. E. & Nikoloski, Z. Genetic basis of plasticity in plants. *J. Exp. Bot.* **70**, 795–804 (2019).

11. Barneix, A. J. Physiology and biochemistry of source-regulated protein accumulation in the wheat grain. *J. Plant Physiol.* **164**, 581–590 (2007).

12. Howarth, J. R. *et al.* Co-ordinated expression of amino acid metabolism in response to

N and S deficiency during wheat grain filling. *J. Exp. Bot.* **59**, 3675–3689 (2008).

13. Laidò, G. *et al.* Genetic Diversity and Population Structure of Tetraploid Wheats (Triticum turgidum L.) Estimated by SSR, DArT and Pedigree Data. *PLoS One* **8**, e67280 (2013).

14. Plett, D. C., Holtham, L. R., Okamoto, M. & Garnett, T. P. Nitrate uptake and its regulation in relation to improving nitrogen use efficiency in cereals. *Seminars in Cell and Developmental Biology* vol. 74 97–104 at https://doi.org/10.1016/j.semcdb.2017.08.027 (2018).

15. Hawkesford, M. J. & Griffiths, S. Exploiting genetic variation in nitrogen use efficiency for cereal crop improvement. *Current Opinion in Plant Biology* vol. 49 35–42 at https://doi.org/10.1016/j.pbi.2019.05.003 (2019).

16. Alaux, M. *et al.* Linking the International Wheat Genome Sequencing Consortium bread wheat reference genome sequence to wheat genetic and phenomic data. *Genome Biol.* **19**, 1–10 (2018).

17. Luo, M. C. *et al.* The structure of wild and domesticated emmer wheat populations, gene flow between them, and the site of emmer domestication. *Theor. Appl. Genet.* **114**, 947–959 (2007).

18. Civáň, P., Ivaničová, Z. & Brown, T. A. Reticulated Origin of Domesticated Emmer Wheat Supports a Dynamic Model for the Emergence of Agriculture in the Fertile Crescent. *PLoS One* **8**, e81955 (2013).

19. Oliveira, H. R., Jacocks, L., Czajkowska, B. I., Kennedy, S. L. & Brown, T. A. Multiregional origins of the domesticated tetraploid wheats. *PLoS One* **15**, e0227148 (2020).

20. Houle, D. Comparing evolvability and variability of quantitative traits. *Genetics* **130**,

195–204 (1992).

21. Hansen, T. F., Pélabon, C. & Houle, D. Heritability is not Evolvability. *Evol. Biol. 2011 383* **38**, 258–277 (2011).

22. Garcia-Gonzalez, F., Simmons, L. W., Tomkins, J. L., Kotiaho, J. S. & Evans, J. P. COMPARING EVOLVABILITIES: COMMON ERRORS SURROUNDING THE CALCULATION AND USE OF COEFFICIENTS OF ADDITIVE GENETIC VARIATION. *Evolution (N. Y).* **66**, 2341–2349 (2012).

23. Mazzucotelli, E. *et al.* The Global Durum Wheat Panel (GDP): An International Platform to Identify and Exchange Beneficial Alleles. *Front. Plant Sci.* **11**, 2036 (2020).

24. Leinonen, T., McCairns, R. J. S., O'Hara, R. B. & Merilä, J. QST -FST comparisons: Evolutionary and ecological insights from genomic heterogeneity. *Nature Reviews Genetics* vol. 14 179–190 at https://doi.org/10.1038/nrg3395 (2013).

25. Savary, S. *et al.* The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol. 2019 33* **3**, 430–439 (2019).

26. Porter, S. S. & Sachs, J. L. Agriculture and the Disruption of Plant–Microbial Symbiosis. *Trends Ecol. Evol.* **35**, 426–439 (2020).

27. Valente, J. *et al.* Symbiotic Variations among Wheat Genotypes and Detection of Quantitative Trait Loci for Molecular Interaction with Auxin-Producing Azospirillum PGPR. *Microorg. 2023, Vol. 11, Page 1615* **11**, 1615 (2023).

28. Martín-Robles, N. *et al.* Impacts of domestication on the arbuscular mycorrhizal symbiosis of 27 crop species. *New Phytol.* **218**, 322–334 (2018).

29. Sawers, R. J. H., Ramírez-Flores, M. R., Olalde-Portugal, V. & Paszkowski, U. The

impact of domestication and crop improvement on arbuscular mycorrhizal symbiosis in cereals: insights from genetics and genomics. *New Phytol.* **220**, 1135–1140 (2018).

30. Anderson, O. D., Coleman-Derr, D., Gu, Y. Q. & Heath, S. Structural and transcriptional analysis of plant genes encoding the bifunctional lysine ketoglutarate reductase saccharopine dehydrogenase enzyme. *BMC Plant Biol.* **10**, 113 (2010).

31. Houmard, N. M. *et al.* High-lysine corn generated by endosperm-specific suppression of lysine catabolism using RNAi. *Plant Biotechnol. J.* **5**, 605–614 (2007).

32. Frizzi, A. *et al.* Modifying lysine biosynthesis and catabolism in corn with a single bifunctional expression/silencing transgene cassette. *Plant Biotechnol. J.* **6**, 13–21 (2008).

33. Long, X., Liu, Q., Chan, M., Wang, Q. & Sun, S. S. M. Metabolic engineering and profiling of rice with increased lysine. *Plant Biotechnol. J.* **11**, 490–501 (2013).

34. Arruda, P., Kemper, E. L., Papes, F. & Leite, A. Regulation of lysine catabolism in higher plants. *Trends Plant Sci.* **5**, 324–330 (2000).

35. Chapman, M. A. *et al.* A Genomic Scan for Selection Reveals Candidates for Genes Involved in the Evolution of Cultivated Sunflower (Helianthus annuus). *Plant Cell* **20**, 2931 (2008).

36. Swanson-Wagner, R. *et al.* Reshaping of the maize transcriptome by domestication. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 11878–11883 (2012).

37. Michaletti, A., Naghavi, M. R., Toorchi, M., Zolla, L. & Rinalducci, S. Metabolomics and proteomics reveal drought-stress responses of leaf tissues from spring-wheat. *Sci. Reports 2018 81* **8**, 1–18 (2018).

38. Kavi Kishor, P. B., Suravajhala, P., Rathnagiri, P. & Sreenivasulu, N. Intriguing Role

of Proline in Redox Potential Conferring High Temperature Stress Tolerance. *Front. Plant Sci.* **13**, 867531 (2022).

39. Beleggia, R. *et al.* Comparative Analysis Based on Transcriptomics and Metabolomics Data Reveal Differences between Emmer and Durum Wheat in Response to Nitrogen Starvation. *Int. J. Mol. Sci.* **22**, 4790 (2021).

40. Curci, P. L. *et al.* Transcriptomic response of durum wheat to nitrogen starvation. *Sci. Rep.* **7**, 1–14 (2017).

41. Sultana, N. *et al.* Transcriptomic Study for Identification of Major Nitrogen Stress Responsive Genes in Australian Bread Wheat Cultivars. *Front. Genet.* **11**, 583785 (2020).

42. Lemmon, Z. H., Bukowski, R., Sun, Q. & Doebley, J. F. The Role of cis Regulatory Evolution in Maize Domestication. *PLOS Genet.* **10**, e1004745 (2014).

43. Page, A., Gibson, J., Meyer, R. S. & Chapman, M. A. Eggplant Domestication: Pervasive Gene Flow, Feralization, and Transcriptomic Divergence. *Mol. Biol. Evol.* **36**, 1359–1372 (2019).

44. Miflin, B. J. & Habash, D. Z. The role of glutamine synthetase and glutamate dehydrogenase in nitrogen assimilation and possibilities for improvement in the nitrogen utilization of crops. *J. Exp. Bot.* **53**, 979–987 (2002).

45. Fontaine, J. X. *et al.* A quantitative genetic study for elucidating the contribution of glutamine synthetase, glutamate dehydrogenase and other nitrogen-related physiological traits to the agronomic performance of common wheat. *Theor. Appl. Genet.* **119**, 645–662 (2009).

46. Walkowiak, S. *et al.* Multiple wheat genomes reveal global variation in modern breeding. *Nat. 2020 5887837* **588**, 277–283 (2020).

47. Hoagland, D. R. & Arnon, D. I. The water-culture method for growing plants without soil. *Circ. Calif. Agric. Exp. Stn.* **347**, (1950).

48. David, J. *et al.* Genotyping by sequencing transcriptomes in an evolutionary pre-breeding durum wheat population. *Mol. Breed.* **34**, 1531–1548 (2014).

49. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10 (2011).

50. Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data - Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2014).

51. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

52. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods 2017 144* **14**, 417–419 (2017).

53. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 1–4 (2021).

54. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156 (2011).

55. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559 (2007).

56. R Core Team. *R: A language and environment for statistical computing. R Foundation Statistical Computing*. https://www.r-project.org/ (2022).

57. Tajima, F. Evolutionary relationship of DNA sequences in finite populations. *Genetics*

105, 437–460 (1983).

58. Watterson, G. A. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* **7**, 256–276 (1975).

59. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).

60. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 1–13 (2014).

61. Vigouroux, Y. *et al.* Identifying genes of agronomic importance in maize by screening microsatellites for evidence of selection during domestication. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 9650–9655 (2002).

62. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–40 (2010).

63. Smyth, G. K. limma: Linear Models for Microarray Data. *Bioinforma. Comput. Biol. Solut. Using R Bioconductor* 397–420 (2005) doi:10.1007/0-387-29362-0_23.

64. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, 1–48 (2015).

65. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

66. Zhang, Z. H. *et al.* A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS One* **9**, e103207 (2014).

67. Rapaport, F. *et al.* Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol* **14**, R95 (2013).

68. Tian, T. *et al.* agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. *Nucleic Acids Res.* **45**, W122–W129 (2017).

Acknowledgements: This work was financially supported by Università Politecnica delle Marche, by the PON a3 PlASS Project. PlASS -Platform for Agrofood Science and Safety and by the Transnational Access capacities of the European Plant Phenotyping Network (EPPN, grant n. 284443) funded by the FP7 Research Infrastructures Program of the European Union.

T.G. thanks TritOmics project, SIR-MIUR grant n. (RBSI14SJ81).

H.T. and Z.N. are funded by the European Union's Horizon 2020 research and innovation program through grant n. 862201.

Figures and Tables

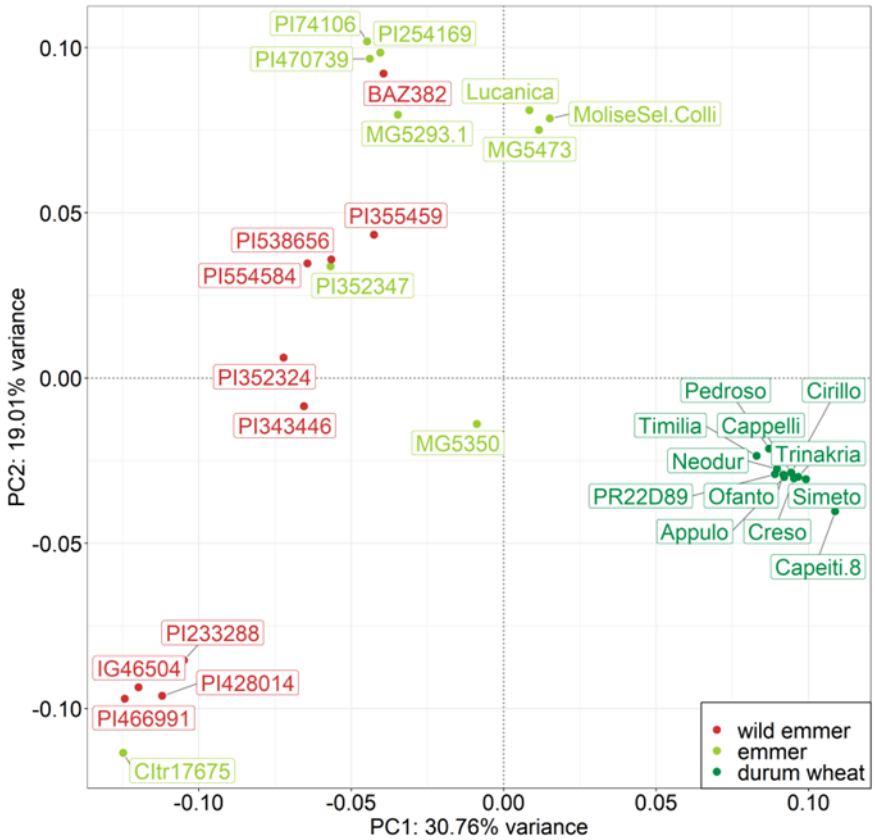


Fig. 1: Principal component analysis of 32 wheat genotypes based on single-nucleotide polymorphisms (SNPs). The first two principal components (PC1 and PC2) are shown. The three colors represent different taxa. Labels show the accession name of each genotype.

				Loss of nucleotide diversity		
	wild emmer	**emmer**	**durum wheat**	Lpd	Lsd	both
π	0.0050	0.0045	0.0029	11.4	34.6	42.1
θ	0.0047	0.0040	0.0029	15.3	27.2	38.3
D	0.2613	0.4605	1.0847			

Table 1: Nucleotide diversity estimates and diversity loss for the three wheat taxa. Diversity loss is shown during primary domestication (wild emmer to emmer, Lpd), secondary domestication (emmer to durum wheat, Lsd) and both processes (wild emmer to durum wheat), based on average π and θ values. The π and θ symbols represent averaged estimates of nucleotide diversity, whereas D represents the neutrality test statistic Tajima's D.

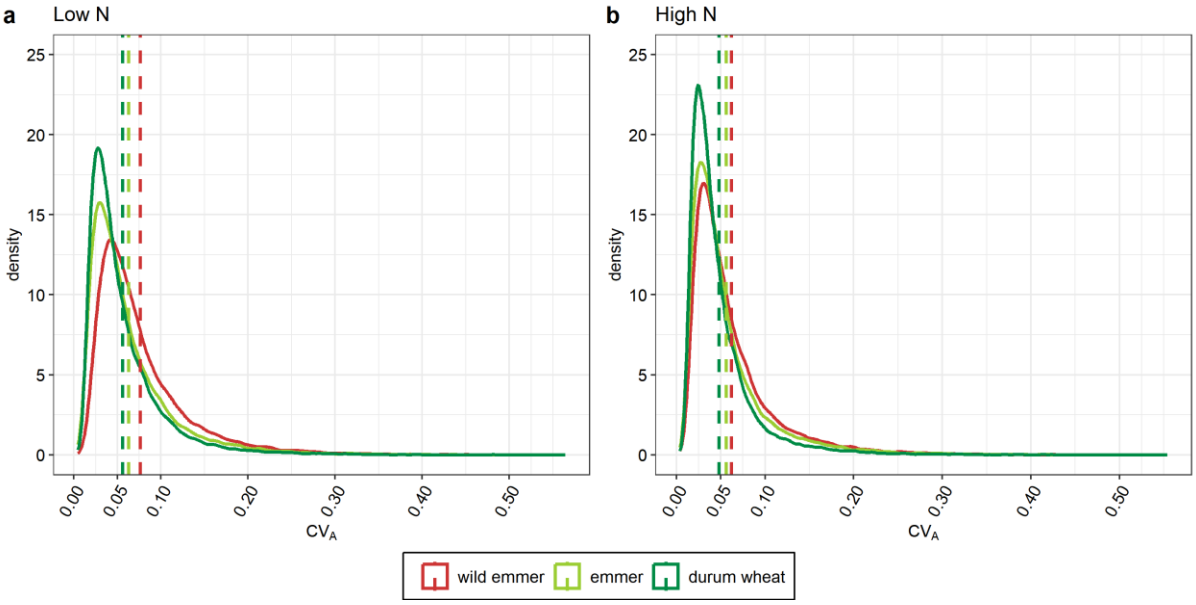


Fig. 2: Density plots of the additive coefficient of variation (CV_A) in the three wheat taxa. Comparison of the estimated density functions of the CV_A in gene expression, calculated using all 32,358 genes. **a** High nitrogen conditions. **b** Low nitrogen conditions. Dashed lines represent the averaged CV_A value, colored according to the different taxa.

				Loss of expression diversity		
	wild emmer	**emmer**	**durum wheat**	Lpd	Lsd	both
CV_A high N	0.062	0.056	0.048	9.1	14.5	22.3
CV_A low N	0.076	0.063	0.056	17.6	11.1	26.7

Table 2: Mean additive coefficient of variation (CV_A) in gene expression and loss of expression diversity for the three wheat taxa. Diversity loss is shown during primary domestication (wild emmer to emmer, Lpd), secondary domestication (emmer to durum wheat, Lsd) and both processes (wild emmer to durum wheat), based on averaged CV_A values calculated for all 32,358 genes.

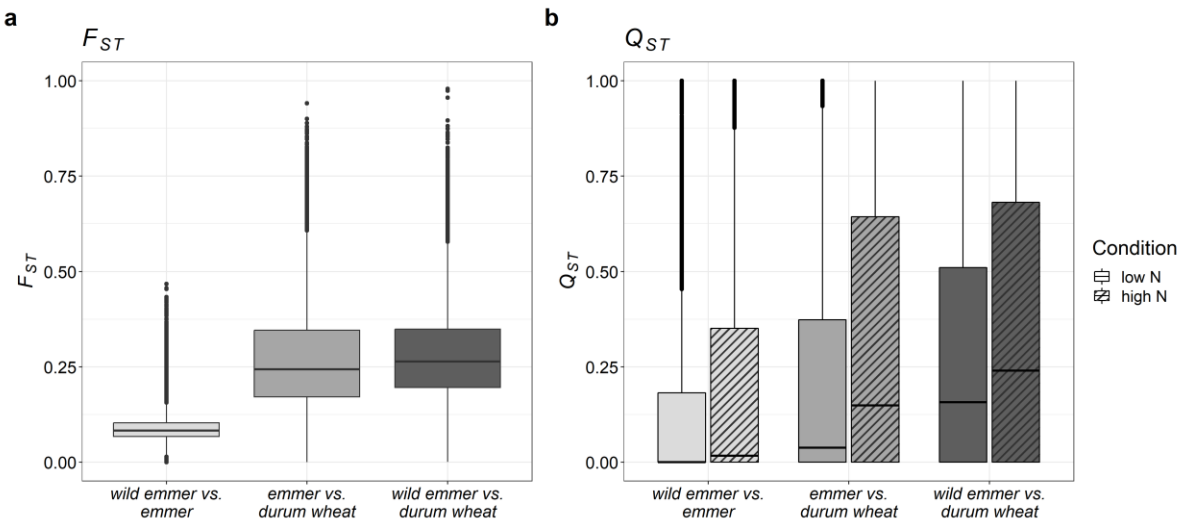
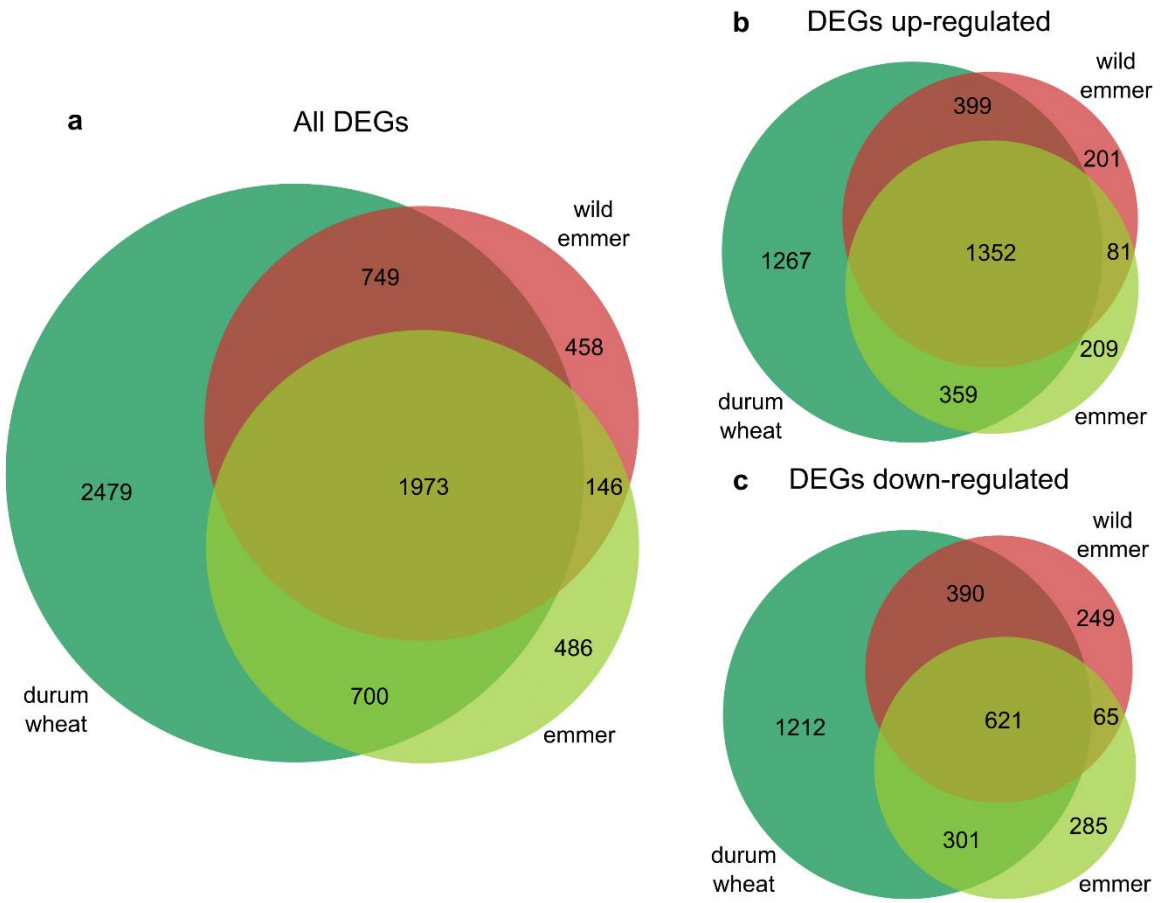


Fig. 3: F_{ST} and Q_{ST} distributions. a Boxplots showing the gene locus F_{ST} distribution for every subspecies pairwise comparison. **b** Boxplots showing the transcript Q_{ST} distribution for every subspecies pairwise comparison under low nitrogen and high nitrogen conditions, represented by empty and hatched grayscale bars, respectively.



744

745 **Fig. 4: Differentially expressed genes (DEGs) when comparing high and low nitrogen**
746 **conditions within each subspecies.** Venn diagrams showing **a** Total set of DEGs; **b**
747 upregulated DEGs only; and **c** downregulated DEGs only.

748

749

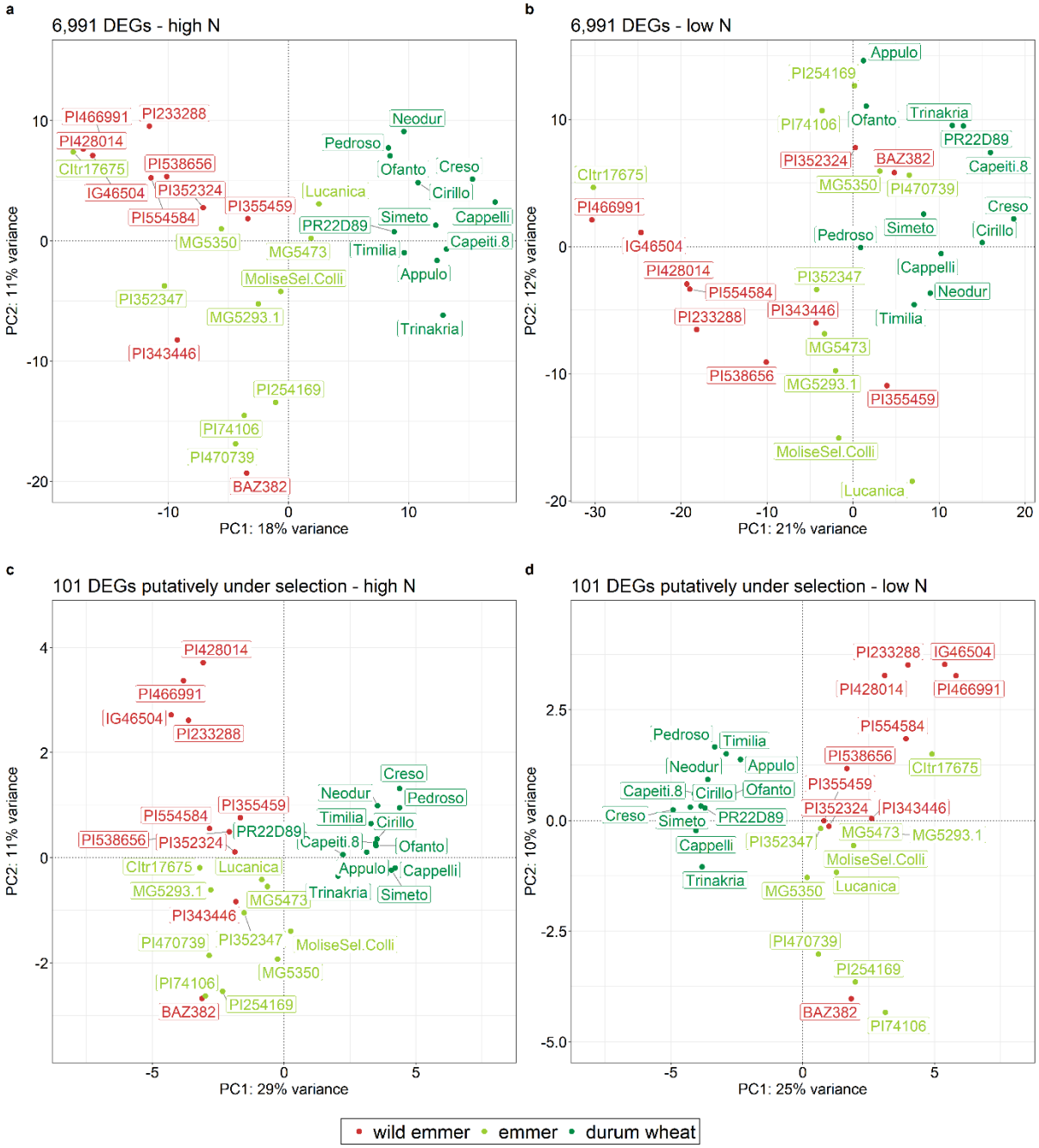
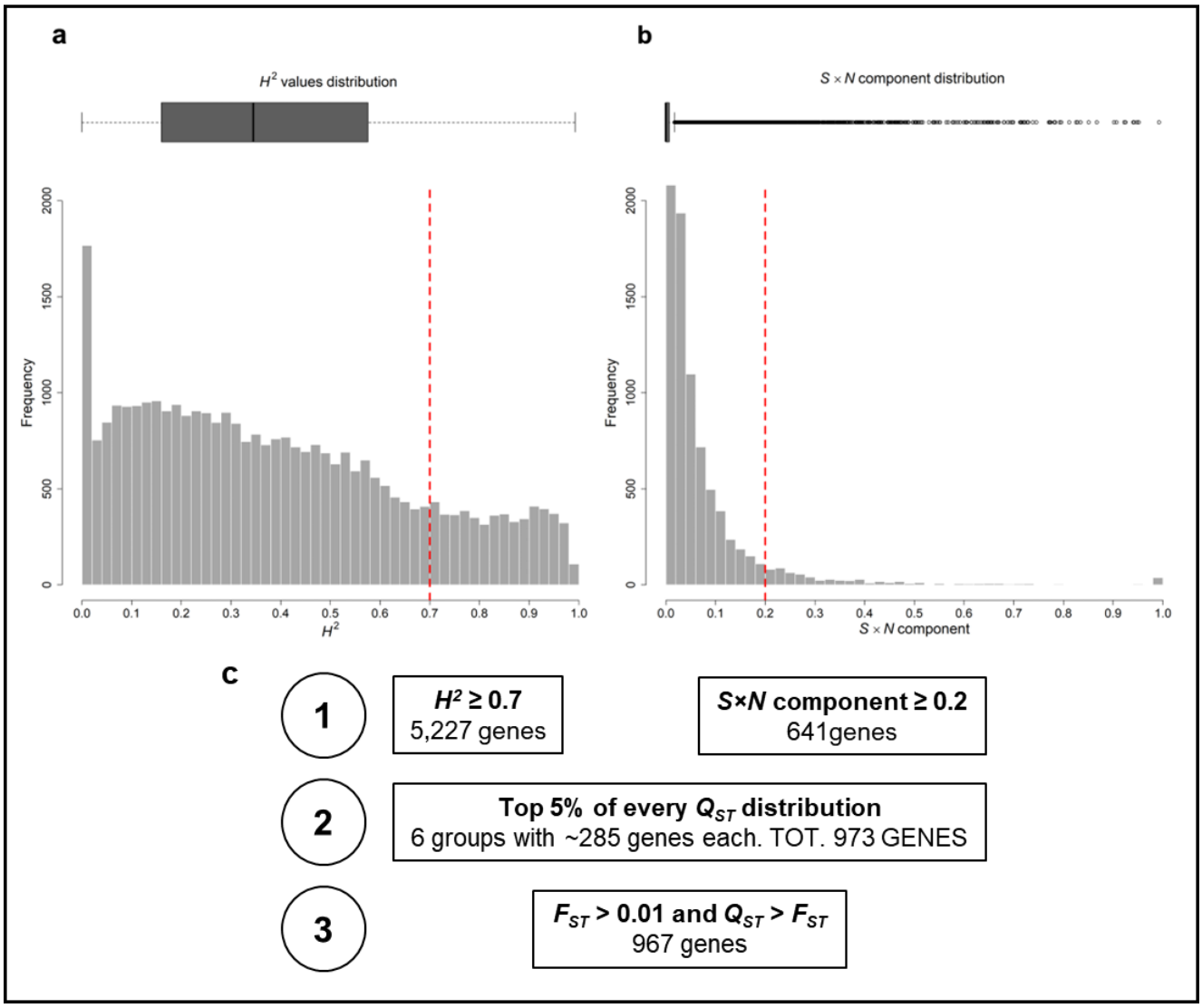
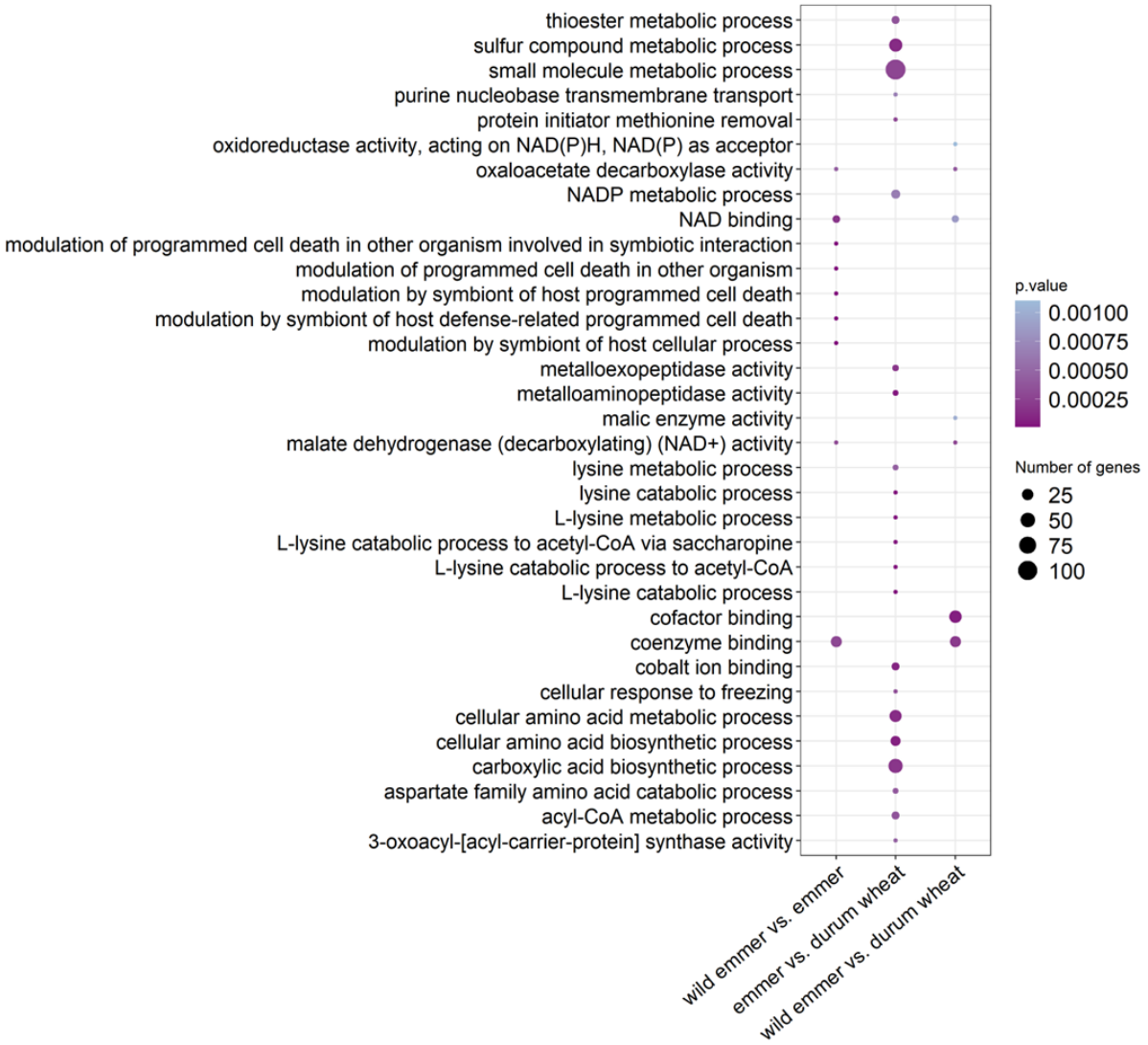


Fig. 5: Principal component analysis of differentially expressed genes when comparing high and low nitrogen conditions within each subspecies. a,b Plots based on all 6,991 DEGs (not filtered): **a** high nitrogen conditions and **b** low nitrogen conditions. **c,d** Plots based on 101 DEGs that are also putatively under selection: **c** high nitrogen conditions and **d** low nitrogen conditions. Samples are represented by taxa-based colored dots. Labels show the accession name of each genotype.

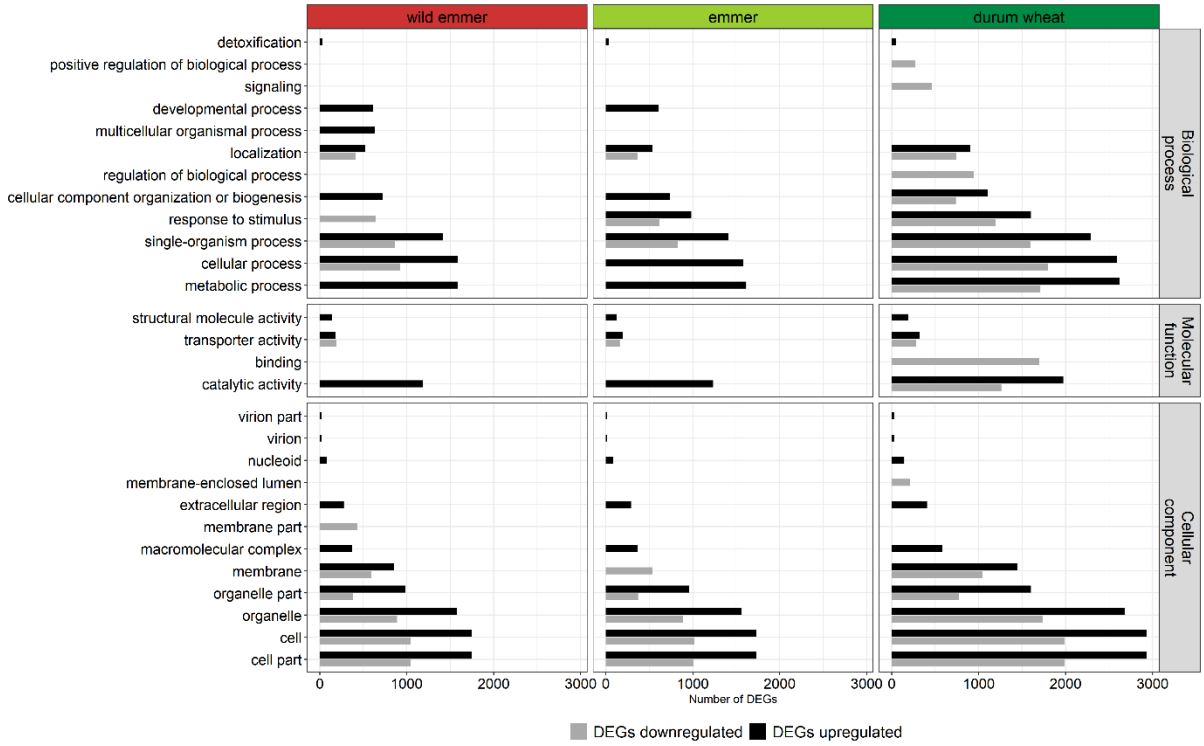


Supplementary Fig. S1: Workflow of gene expression selection scanning. Histograms and boxplots showing distributions of **a** H^2 values and **b** $S \times N$ components calculated using all 32,358 genes. Red dashed lines indicate the arbitrary thresholds chosen for each distribution as the first filtering step of the selection scan. **c** Subsequent steps for the identification of gene expression selection signatures.

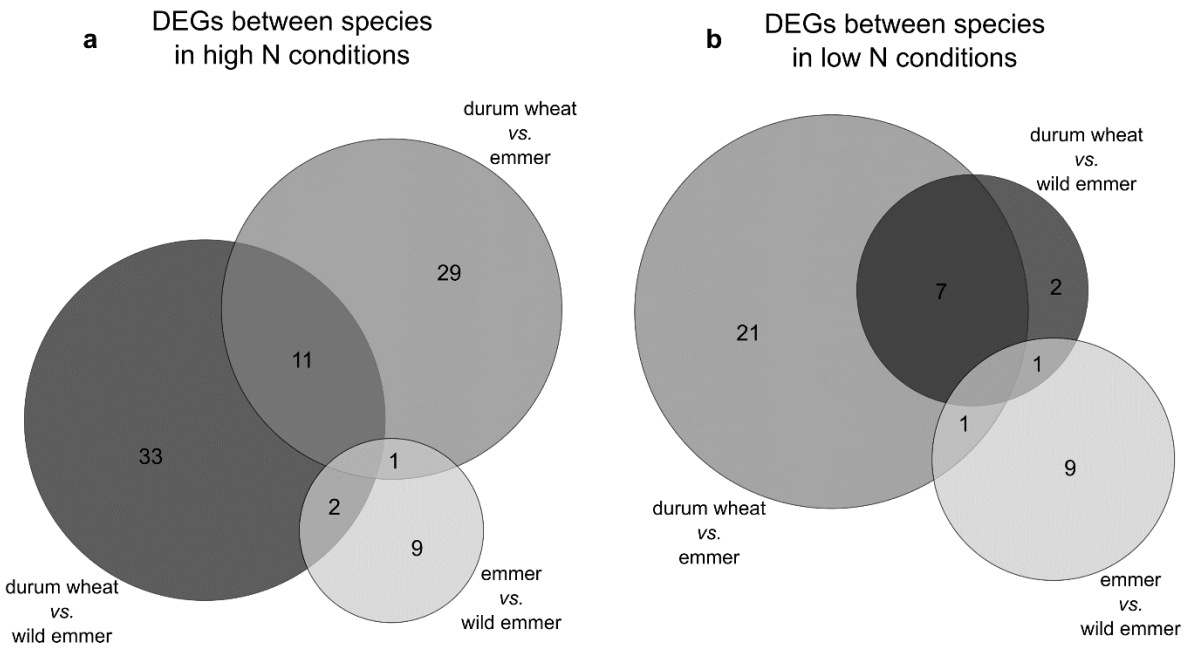


Supplementary Fig. S2: GO categories of genes under selection. Dot plot showing the GO categories enriched in each group of genes putatively under selection during primary domestication (wild emmer to emmer), secondary domestication (emmer to durum wheat) and both processes (wild emmer to durum wheat). The size of the dots represents the number of genes associated with the GO term and colors indicate the p-values. The represented categories passed the multi-test adjustment Hochberg FDR < 0.05.

Supplementary Fig. S3: Genome-wide distribution of differentially expressed genes (DEGs) in the comparison between contrasting nitrogen conditions within each subspecies. Circles represent private DEGs, squares represent DEGs shared between two subspecies, and triangles represent DEGs common to all three subspecies. Full blue shapes represent DEGs satisfying the threshold $|\log_2 FC| \geq 1$. Gray empty shapes represent DEGs satisfying the threshold $|\log_2 FC| \leq 1$. DEGs putatively under selection are shown in orange.



Supplementary Fig. S4: GO classification of differentially expressed genes (DEGs) in the comparison between contrasting nitrogen conditions within each subspecies. Black and gray bars indicate the number of upregulated and downregulated DEGs, respectively. The length of the bars represents the number of DEGs associated with each GO term. The represented categories passed the multi-test adjustment Hochberg FDR < 0.05.



Supplementary Fig. S5: Differentially expressed genes (DEGs) between subspecies. Venn diagrams show the numbers of DEGs between each subspecies in pairwise comparisons under **a** high nitrogen conditions and **b** low nitrogen conditions.

Supplementary Tables

Supplementary Table S1: List of the 128 samples and read mapping results.

Supplementary Table S2: Raw read counts of the 32,358 genes.

Supplementary Table S3: Mean CV_A in gene expression for the three wheat taxa and loss of expression diversity. The loss of expression diversity is shown for two gene subgroups (6,991 DEGs and 25,367 non-DEGs).

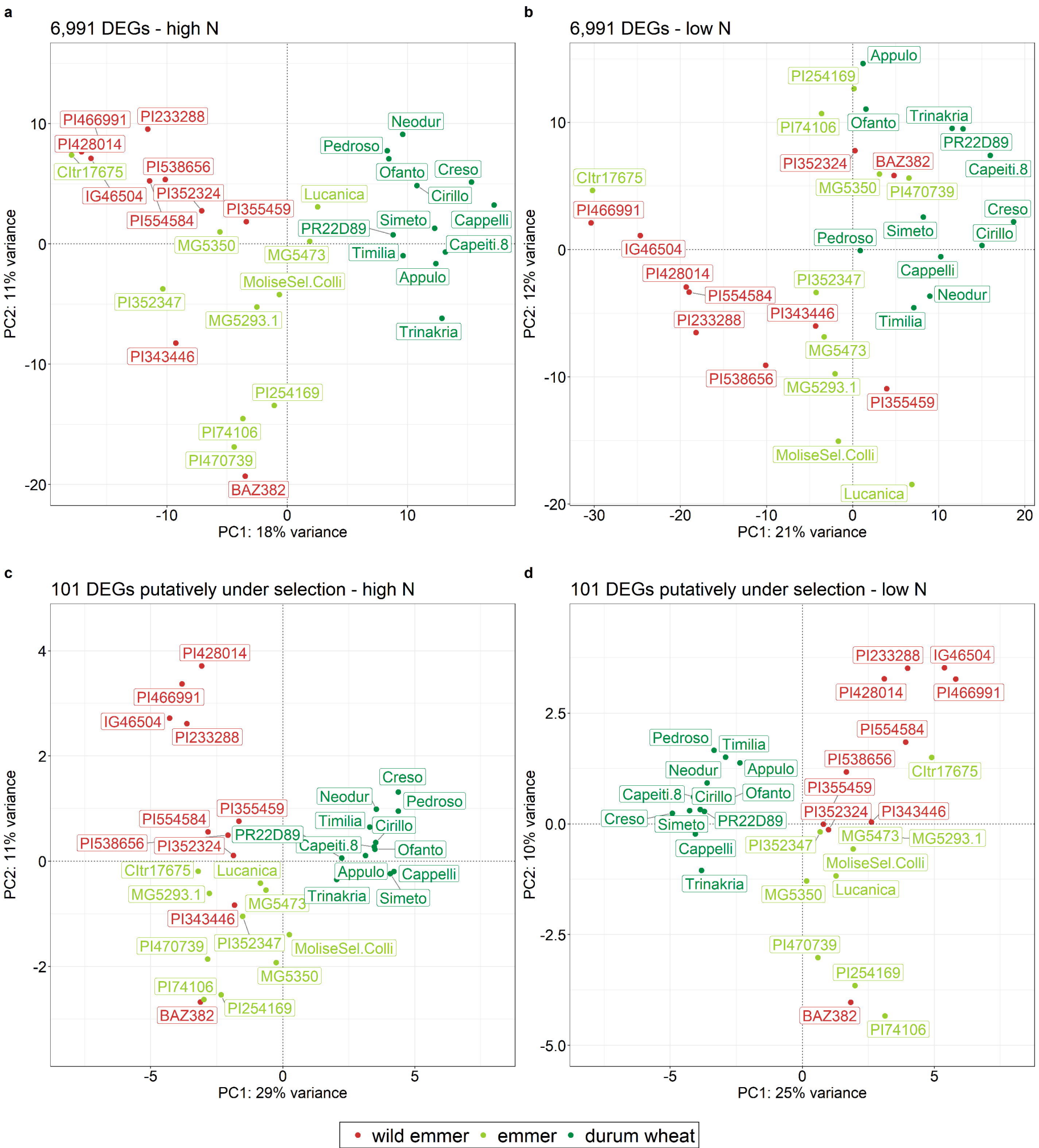
Supplementary Table S4: List of the 967 genes retained from the "selection scan". Each gene is accompanied by its functional annotation and the group in which the selection signal was detected.

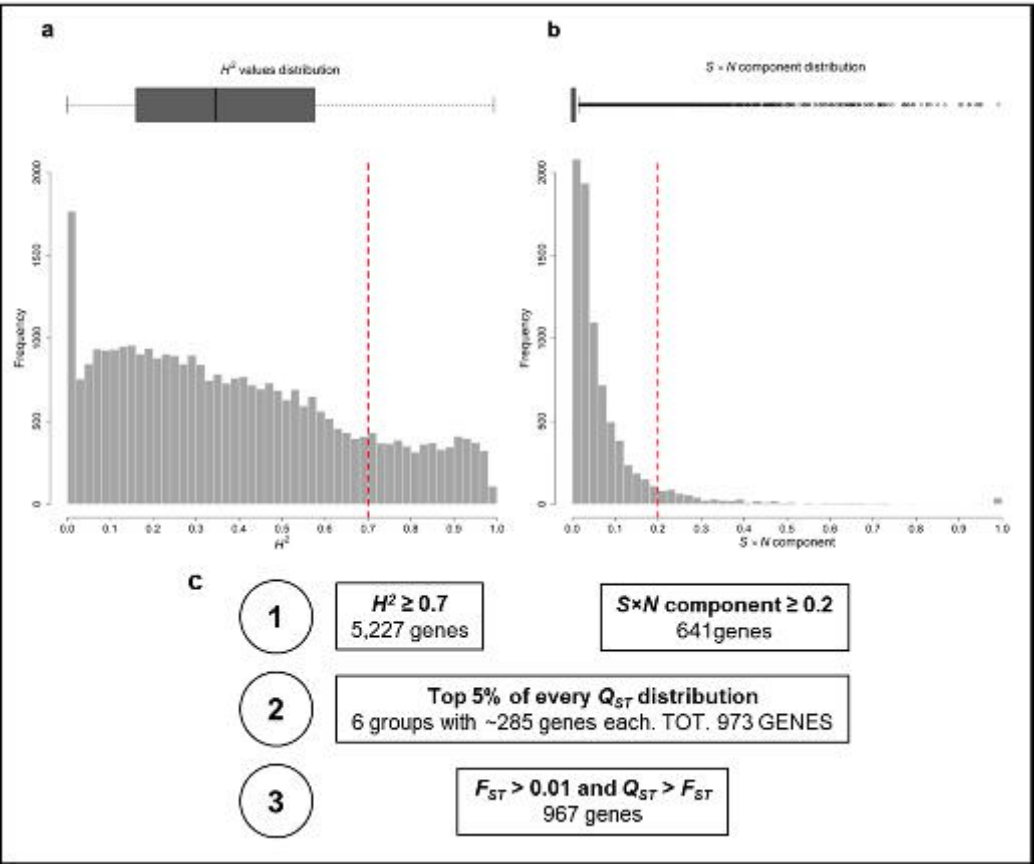
Supplementary Table S5: List of GO "Biological process" and "Molecular function" subcategories for differentially expressed genes (DEGs). GO subcategories are shown for upregulated and downregulated genes under different nitrogen conditions for each subspecies, satisfying the criterion $p \leq 10^{-5}$.

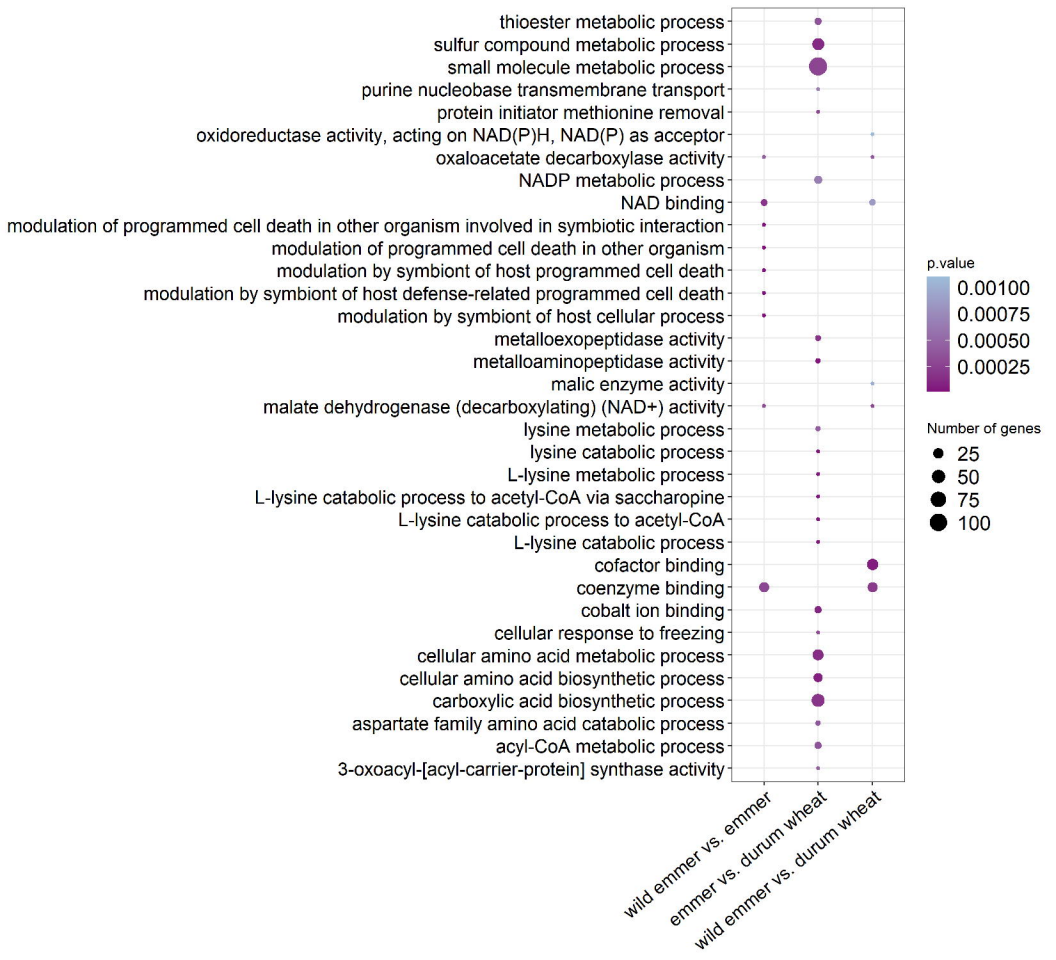
Supplementary Table S6: Functional annotations of the differentially expressed genes (DEGs) between nitrogen conditions in each subspecies. Genes with the top 5% $|\log_2 FC|$ values are shown.

Supplementary Table S7: Functional annotations of the differentially expressed genes (DEGs) between subspecies under all nitrogen conditions. The corresponding $\log_2 FC$ values are shown.

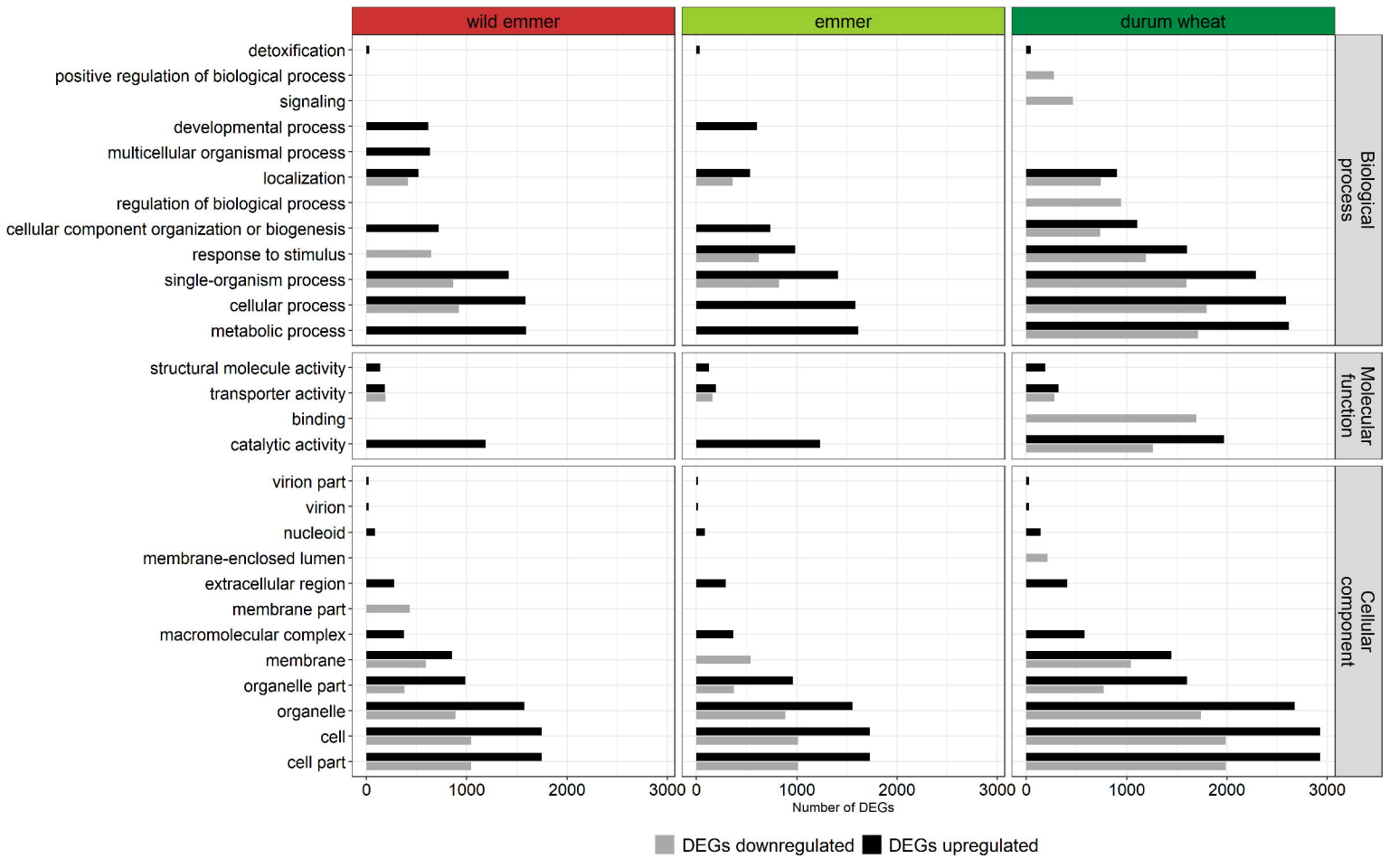
Supplementary Table S8: Functional annotation of the 101 genes selected by the integration of selection signatures and differential expression analysis between nitrogen conditions.

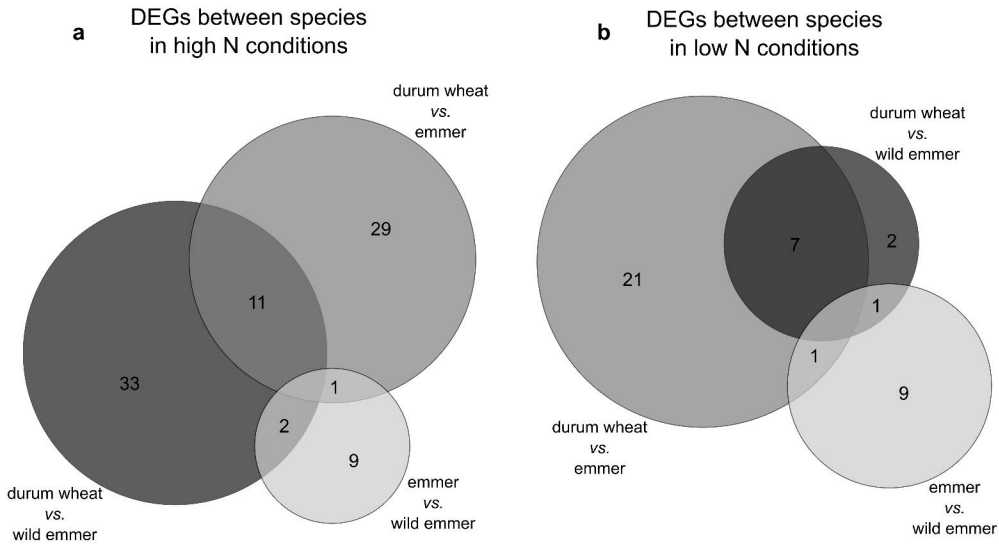

a
6,991 DEGs - high N
PC1: 18% variance
PC2: 11% variance
PI466991
PI233288
PI428014
CItr17675
PI538656
IG46504
PI352324
PI554584
PI355459
MG5350
MG5473
MoliseSel.Colli
PI352347
MG5293.1
PI343446
PI254169
PI74106
PI470739
BAZ382
Neodur
Pedroso
Ofanto
Creso
Lucanica
Cirillo
Cappelli
Simeto
PR22D89
Capeiti.8
Timilia
Appulo
Trinakria
b
6,991 DEGs - low N
PC1: 21% variance
PC2: 12% variance
Appulo
PI254169
PI74106
Ofanto
Trinakria
PR22D89
PI352324
BAZ382
Capeiti.8
CItr17675
MG5350
PI470739
PI466991
Creso
Simeto
Cirillo
IG46504
Pedroso
Cappelli
PI428014
PI352347
PI554584
Neodur
PI343446
PI233288
Timilia
MG5473
PI538656
MG5293.1
PI355459
MoliseSel.Colli
Lucanica
c
101 DEGs putatively under selection - high N
PC1: 29% variance
PC2: 11% variance
PI428014
PI466991
IG46504
PI233288
Creso
Neodur
Pedroso
PI355459
PI554584
Timilia
Cirillo
PI538656
PI352324
PR22D89
Capeiti.8
Ofanto
CItr17675
Lucanica
Appulo
Cappelli
MG5293.1
Trinakria
Simeto
MG5473
PI343446
PI470739
PI352347
MoliseSel.Colli
MG5350
PI74106
PI254169
BAZ382
d
101 DEGs putatively under selection - low N
PC1: 25% variance
PC2: 10% variance
PI233288
IG46504
PI428014
PI466991
PI554584
Pedroso
Timilia
PI538656
Neodur
Appulo
CItr17675
Capeiti.8
Cirillo
Ofanto
PI355459
Creso
Simeto
PR22D89
PI352324
PI343446
MG5473
MG5293.1
Cappelli
PI352347
MoliseSel.Colli
Trinakria
MG5350
Lucanica
PI470739
PI254169
BAZ382
PI74106
wild emmer
emmer
durum wheat


a
H^2 values distribution
Frequency
H^2
b
S × N component distribution
Frequency
S × N component
c
1
$H^2 \geq 0.7$
5,227 genes
S×N component ≥ 0.2
641genes
2
Top 5% of every Q_{ST} distribution
6 groups with ~285 genes each. TOT. 973 GENES
3
$F_{ST} > 0.01$ and $Q_{ST} > F_{ST}$
967 genes


thioester metabolic process
sulfur compound metabolic process
small molecule metabolic process
purine nucleobase transmembrane transport
protein initiator methionine removal
oxidoreductase activity, acting on NAD(P)H, NAD(P) as acceptor
oxaloacetate decarboxylase activity
NADP metabolic process
NAD binding
modulation of programmed cell death in other organism involved in symbiotic interaction
modulation of programmed cell death in other organism
modulation by symbiont of host programmed cell death
modulation by symbiont of host defense-related programmed cell death
modulation by symbiont of host cellular process
metalloexopeptidase activity
metalloaminopeptidase activity
malic enzyme activity
malate dehydrogenase (decarboxylating) (NAD+) activity
lysine metabolic process
lysine catabolic process
L-lysine metabolic process
L-lysine catabolic process to acetyl-CoA via saccharopine
L-lysine catabolic process to acetyl-CoA
L-lysine catabolic process
cofactor binding
coenzyme binding
cobalt ion binding
cellular response to freezing
cellular amino acid metabolic process
cellular amino acid biosynthetic process
carboxylic acid biosynthetic process
aspartate family amino acid catabolic process
acyl-CoA metabolic process
3-oxoacyl-[acyl-carrier-protein] synthase activity
wild emmer vs. emmer
emmer vs. durum wheat
wild emmer vs. durum wheat
p.value
0.00100
0.00075
0.00050
0.00025
Number of genes
25
50
75
100

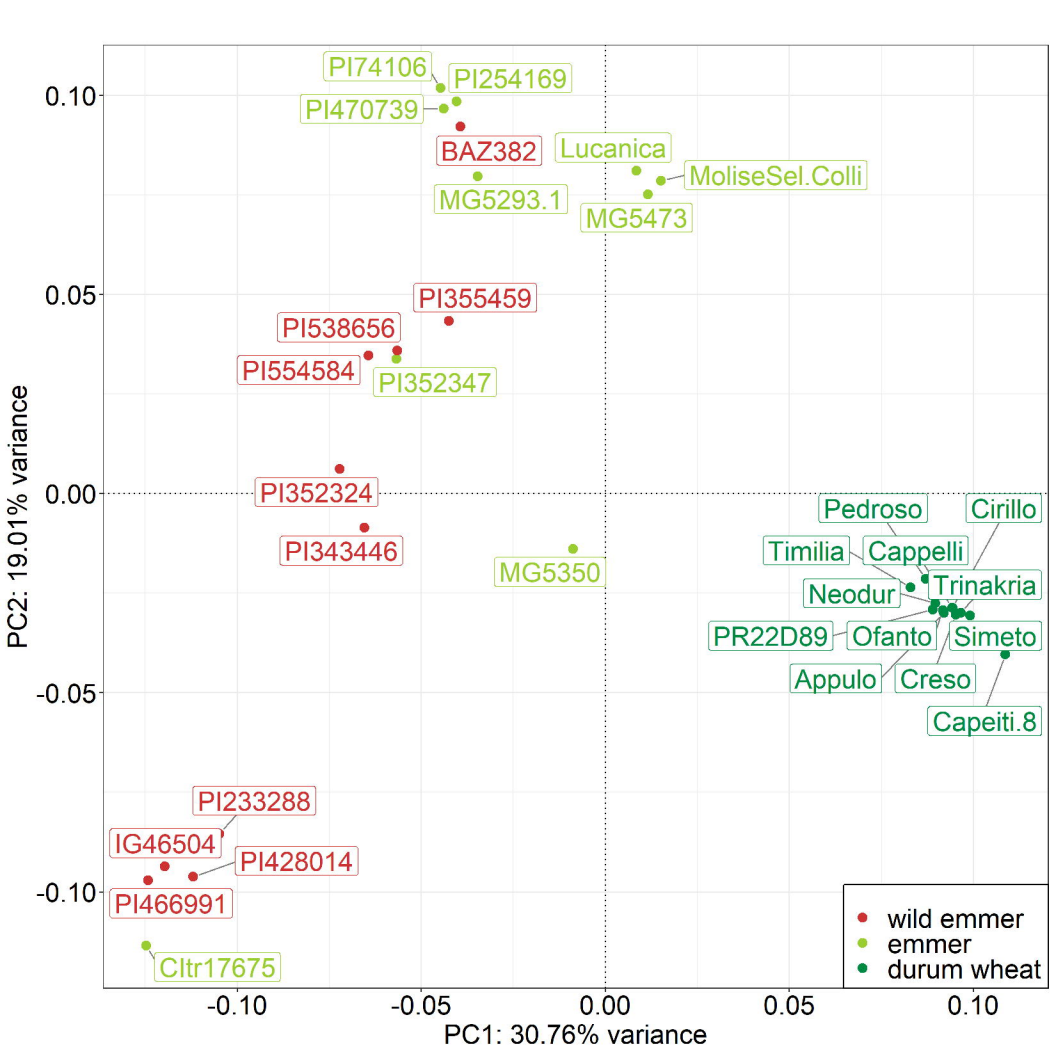
Chr1A
Chr1B
Chr2A
Chr2B
Chr3A
Chr3B
Chr4A
Chr4B
Chr5A
Chr5B
Chr6A
Chr6B
Chr7A
Chr7B
wild emmer
emmer
durum wheat
log2FC
Position (Mb)

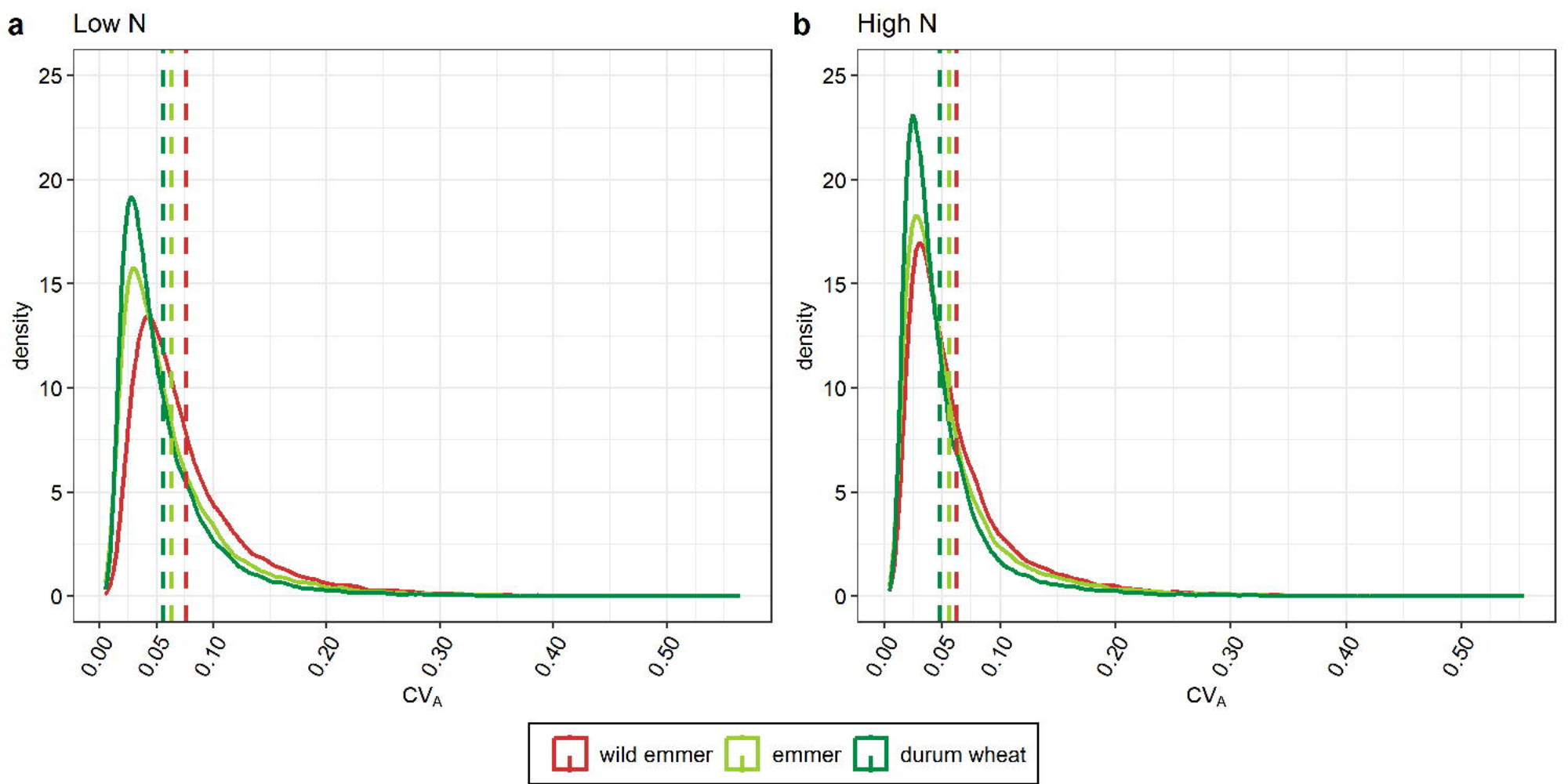

wild emmer
emmer
durum wheat
Biological process
Molecular function
Cellular component
detoxification
positive regulation of biological process
signaling
developmental process
multicellular organismal process
localization
regulation of biological process
cellular component organization or biogenesis
response to stimulus
single-organism process
cellular process
metabolic process
structural molecule activity
transporter activity
binding
catalytic activity
virion part
virion
nucleoid
membrane-enclosed lumen
extracellular region
membrane part
macromolecular complex
membrane
organelle part
organelle
cell
cell part
0
1000
2000
3000
Number of DEGs
DEGs downregulated
DEGs upregulated

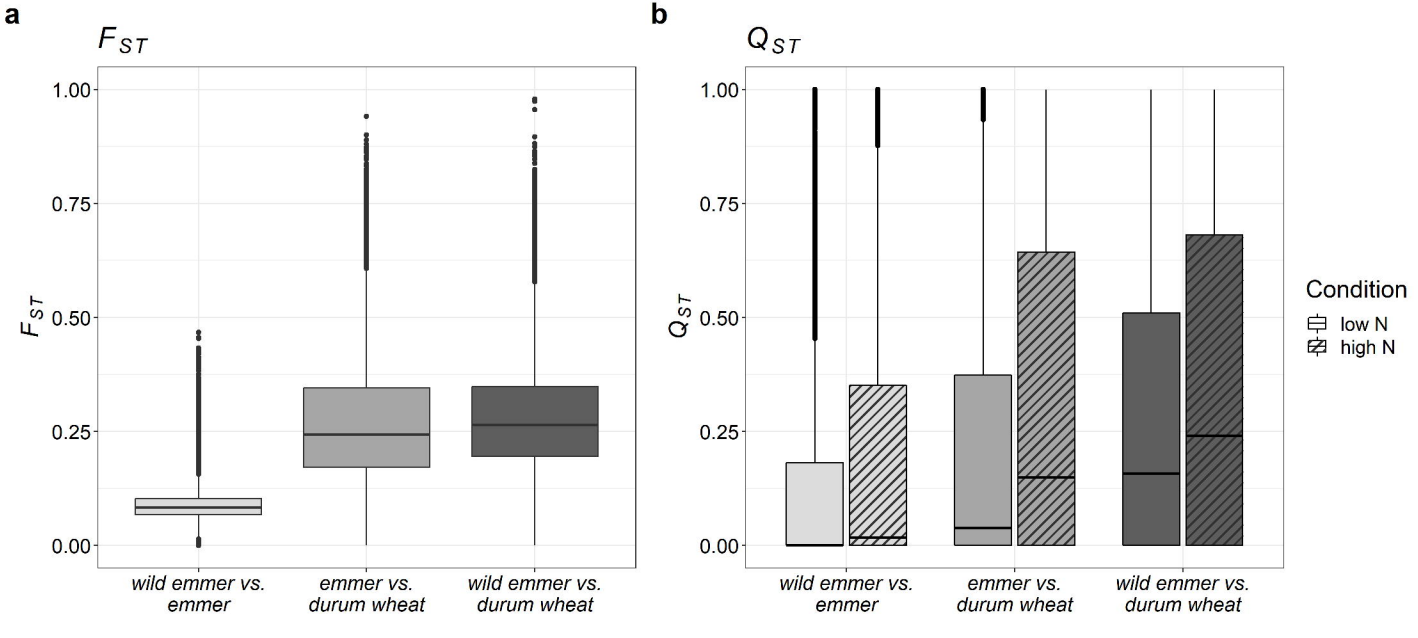

a
DEGs between species
in high N conditions
durum wheat
vs.
emmer
29
11
33
1
2
9
durum wheat
vs.
wild emmer
emmer
vs.
wild emmer
b
DEGs between species
in low N conditions
durum wheat
vs.
wild emmer
7
2
21
1
1
9
durum wheat
vs.
emmer
emmer
vs.
wild emmer

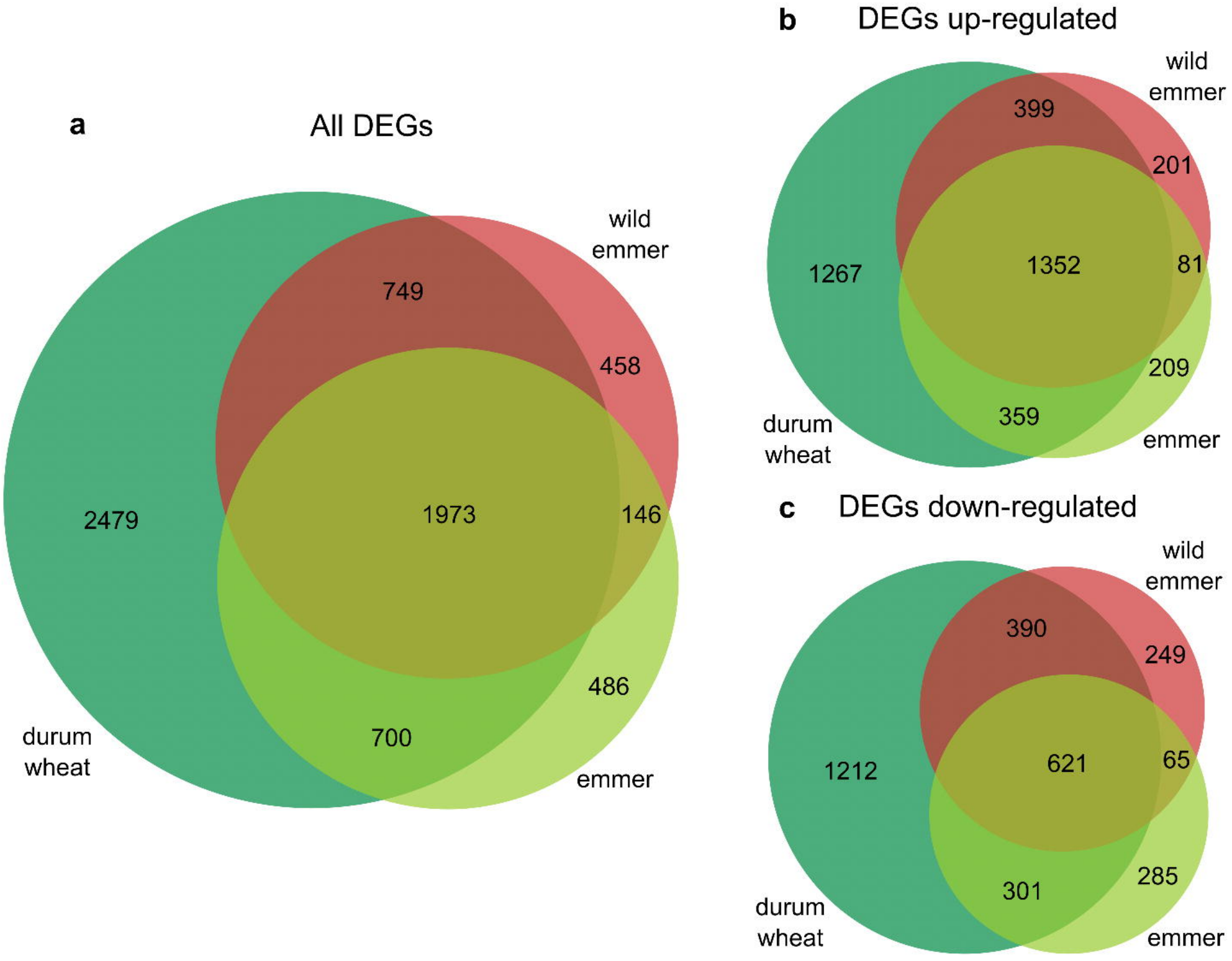
	wild emmer	emmer	durum wheat	Loss of nucleotide diversity		
				Lpd	Lsd	both
π	0.0050	0.0045	0.0029	11.4	34.6	42.1
θ	0.0047	0.0040	0.0029	15.3	27.2	38.3
D	0.2613	0.4605	1.0847			

	wild emmer	emmer	durum wheat	Loss of expression diversity		
				Lpd	Lsd	both
CV_A high N	0.062	0.056	0.048	9.1	14.5	22.3
CV_A low N	0.076	0.063	0.056	17.6	11.1	26.7


PI74106
PI254169
PI470739
BAZ382
Lucanica
MoliseSel.Colli
MG5293.1
MG5473
PI355459
PI538656
PI554584
PI352347
PI352324
PI343446
MG5350
Pedroso
Cirillo
Timilia
Cappelli
Trinakria
Neodur
PR22D89
Ofanto
Simeto
Appulo
Creso
Capeiti.8
PI233288
IG46504
PI428014
PI466991
CItr17675
wild emmer
emmer
durum wheat
0.10
0.05
0.00
-0.05
-0.10
PC2: 19.01% variance
-0.10
-0.05
0.00
0.05
0.10
PC1: 30.76% variance


a
Low N
b
High N
density
25
20
15
10
5
0
0.00
0.05
0.10
0.20
0.30
0.40
0.50
CV_A
wild emmer
emmer
durum wheat


a
F_{ST}
1.00
0.75
0.50
0.25
0.00
wild emmer vs. emmer
emmer vs. durum wheat
wild emmer vs. durum wheat
b
Q_{ST}
Condition
low N
high N


a
All DEGs
wild emmer
749
458
2479
1973
146
486
700
durum wheat
emmer
b
DEGs up-regulated
wild emmer
399
201
1267
1352
81
209
359
durum wheat
emmer
c
DEGs down-regulated
wild emmer
390
249
1212
621
65
301
285
durum wheat
emmer