Genomic predictions and Genome-wide association studies on DNA pools to characterize traditional maize landraces and identify genomic regions associated with agronomic traits and environmental adaptation

Sarah Ben-Sadoun, Agustin Oscar Galaretto, Antoine Roux, Brigitte Gouesnard, Alain Charcosset, Laurence Moreau, Delphine Madur, V. Bourdin, Stephane Nicolas

HAL Id: hal-04807728
https://hal.inrae.fr/hal-04807728v1

Submitted on 27 Nov 2024

Genomic predictions and GWAS on DNA pools to characterize traditional maize landraces and identify genomic regions associated with agronomic traits and environmental adaptation

S. Ben Sadoun[1], A.O. Galaretto[1], A. Roux[1], B. Gouesnard[2], A. Charcosset[1], L. Moreau[1], D. Madur[1], V. Bourdin[1], S.D. Nicolas[1]

1. Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE – Le Moulon, Gif-sur-Yvette, France, 91190
2. AGAP, Univ Montpellier, CIRAD, INRAE, Institut Agro, Montpellier, France, 34398

1. Context and strategy

Traditional maize landraces is a valuable source of genetic diversity to cope with climate change and agroecological transition due to their local adaptation.

DNA pool genotyping is a cost-effective approach to genotype highly heterogeneous maize landraces and identify untapped landraces in modern germplasm (Arca et al., 2023).

We genotyped 262 maize landraces from the French national collection using DNA pools to i) characterize its genetic structure, ii) identify genomic regions associated with agronomic traits and environmental adaptation, and iii) perform genomic predictions.

2. Genetic materials and Phenotypic evaluation

262 traditional maize landraces collected between 1950 and 1970 (Gouesnard et al., 1997) were:

- Genotyped with a 50K SNP array using pooled DNA samples (Arca et al., 2021)
- Evaluated in two field trials for 84 agro-morphological and seed composition traits

Climatic variables of landraces collection sites were retrieved from WorldClim database.

Structure analysis identified 4 genetic groups corresponding to different geographical origins of landraces (Figure 1).
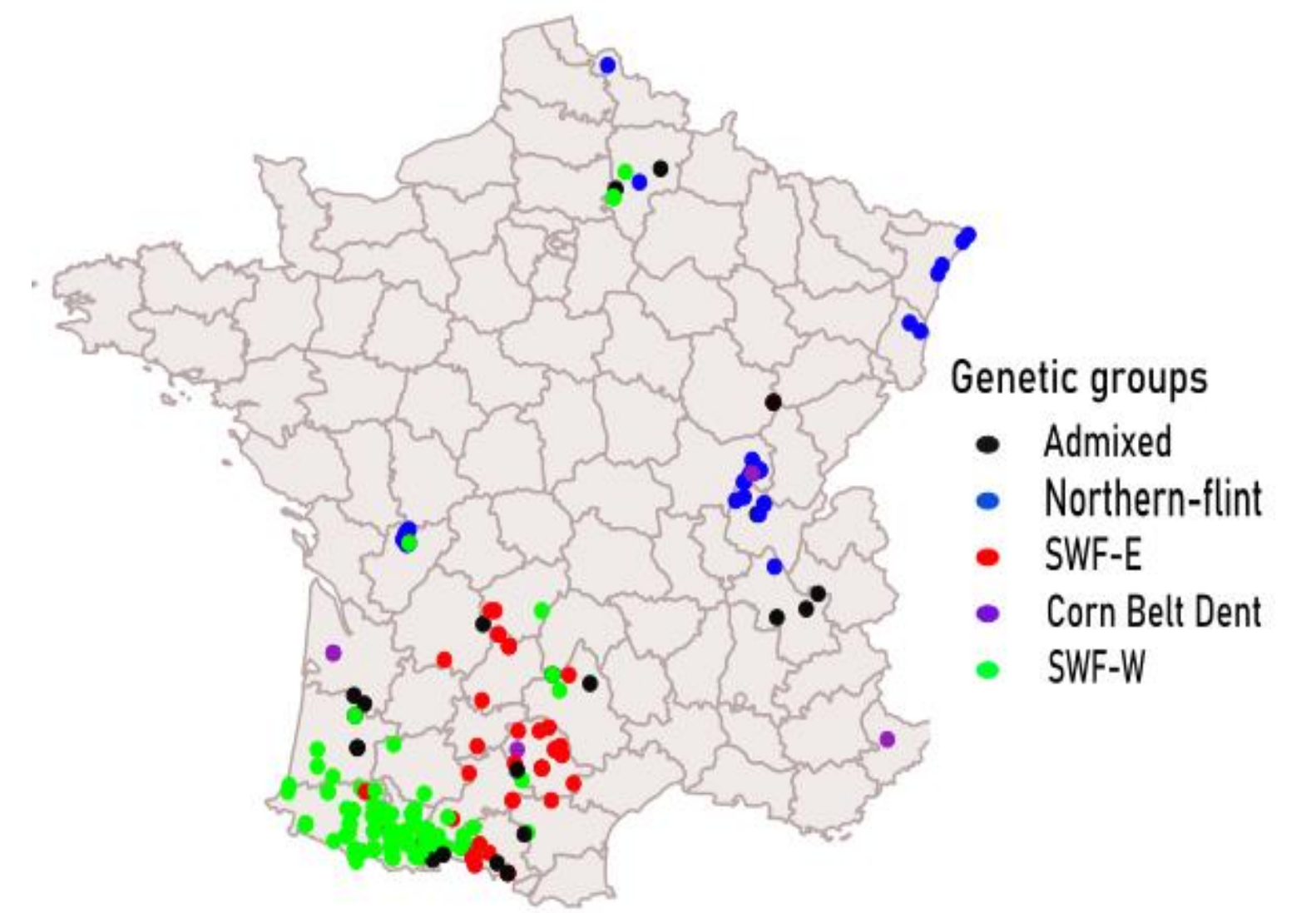

Figure 1: Geographical distribution of the 262 landraces and their assignment to the 4 genetics groups defined by ADMIXTURE (Alexander et al., 2009) Landraces with an assignment probability below 0.5 were considered admixed and colored in black.

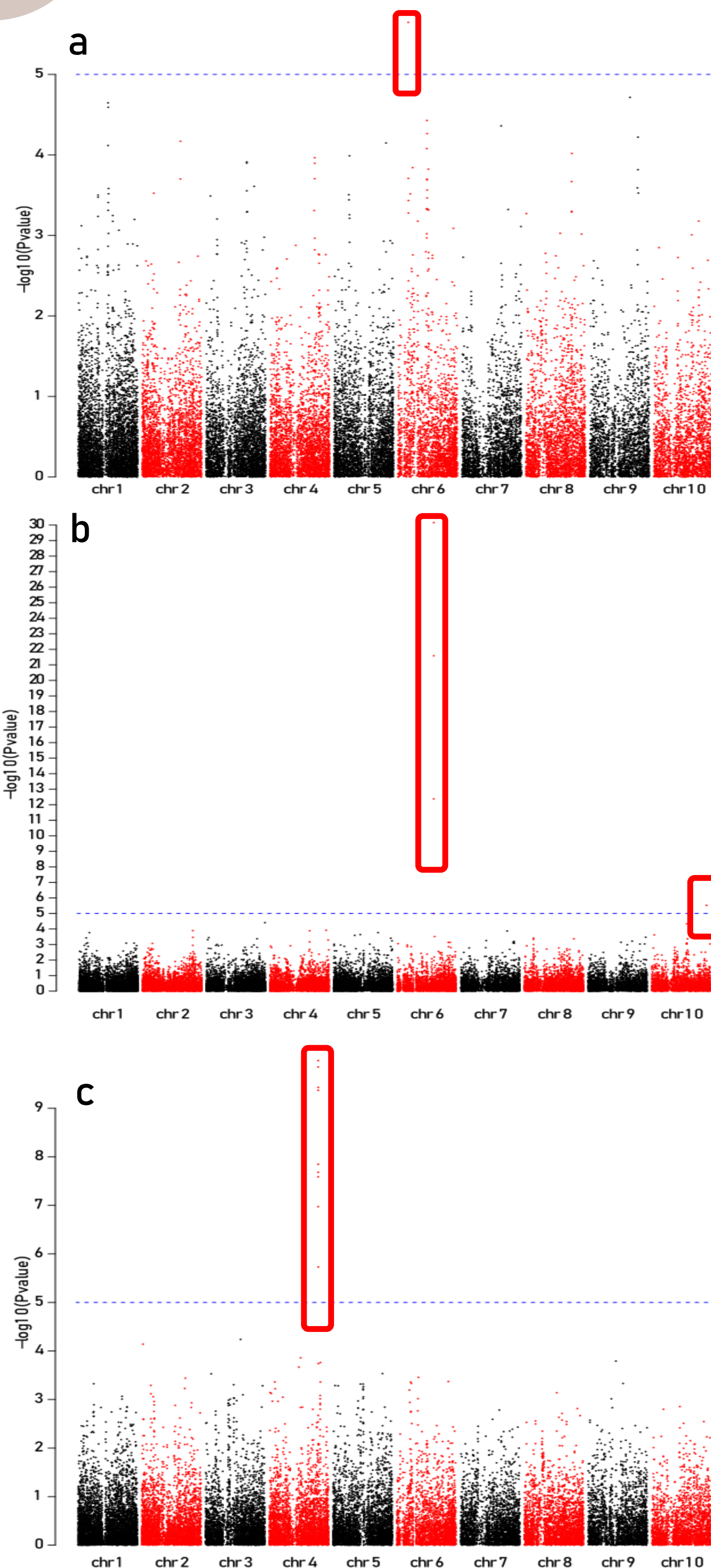## 3. Genome Wide Association Study (GWAS)

Figure 2: Manhattan plots for GWAS of a) flowering time, b) linoleic content, and c) thousand kernel weight.

GWAS performed with allelic frequencies using a mixed model accounting for both relatedness (K) and genetic structure (Q):

$$Y = X\beta + Q\delta + Zu + e$$
$$u \sim N(0, K\sigma_g)$$

K: based on VanRaden (2008), and adapted to use allelic frequencies rather than allele-variants (Cericola et al., 2018)

Q: Admixture matrices for 4 genetic groups

Model implemented in MM4LMM R package (Laporte et al., 2022)

Several Quantitative Trait Loci (QTLs) involved in agro-morphological and seed composition variation were detected (Figure 2).

Some QTLs colocalizes with known genes:

- → QTL on chr6 for linoleic content (Figure 2.b) colocalizes with a gene involved in the pathway of oil biosynthesis in maize (Zheng et al., 2008; Yan et al., 2018)
- → QTL on chr4 (Figure 2c) for Thousand Kernel Weight colocalizes with KRN4 gene that determines Kernel Row Number (Liu et al., 2015)

Thousand Kernel Weight (TKW):

- TKW varies according to genetic structure and geographical origins (Figures 3 and 4)
- Allelic frequency of the most significant SNP on chr4 (Figure 2.c) correlates with genetic structure (Figure 3)

Genome-Environment Association Studies (GEA): detection of several genomic regions associated with environmental and bioclimatic variables
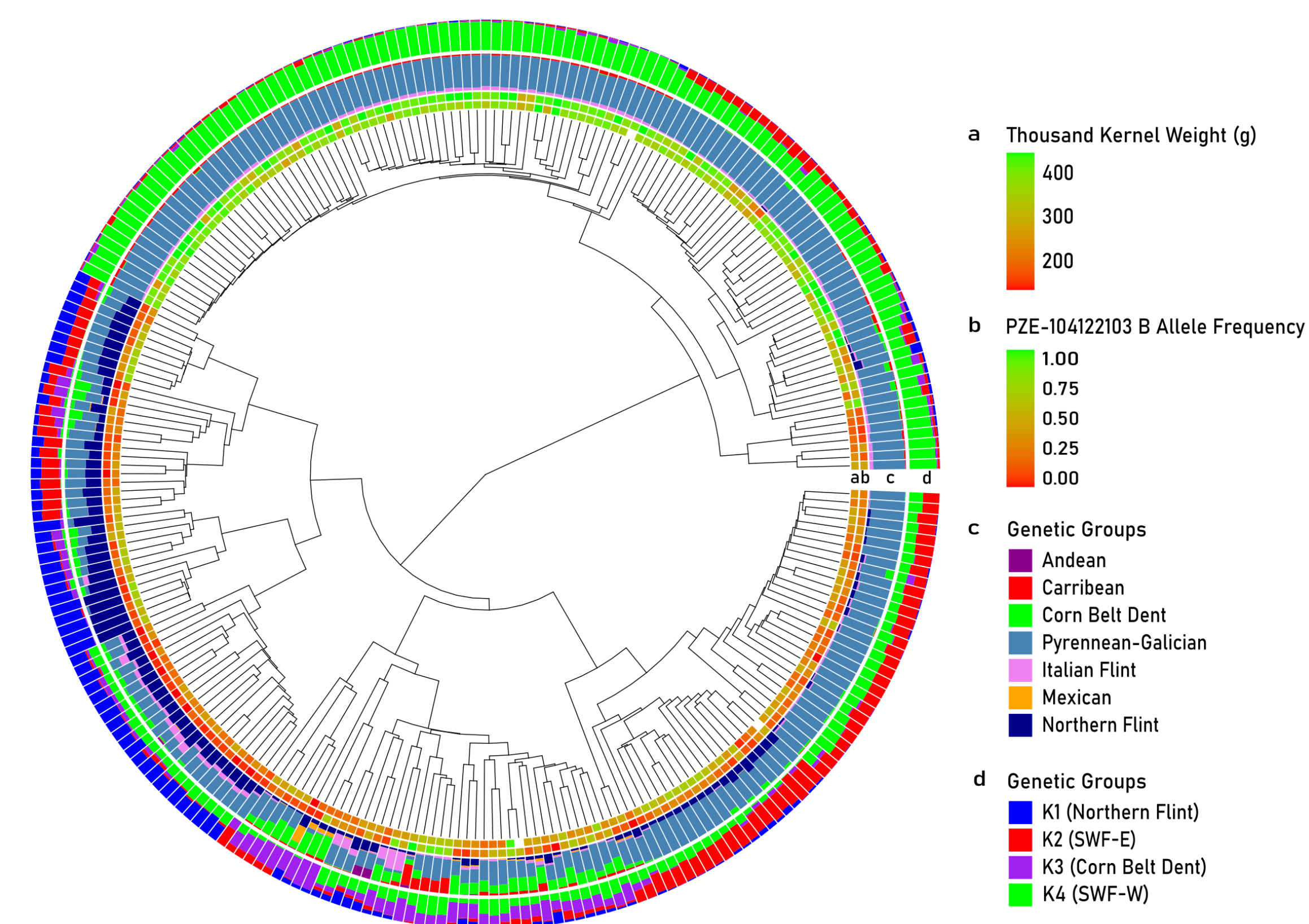

Figure 3: Dendrogram of genetic diversity based of modified Roger's distance (MRD) of the 262 landraces including additional data on (from internal to external layer): a) thousand kernel weight, b) allelic frequency of the most significant SNP associated with thousand kernel weight, c) barplots with the assignment to 7 ancestral groups and d) barplots from ADMIXTURE analysis
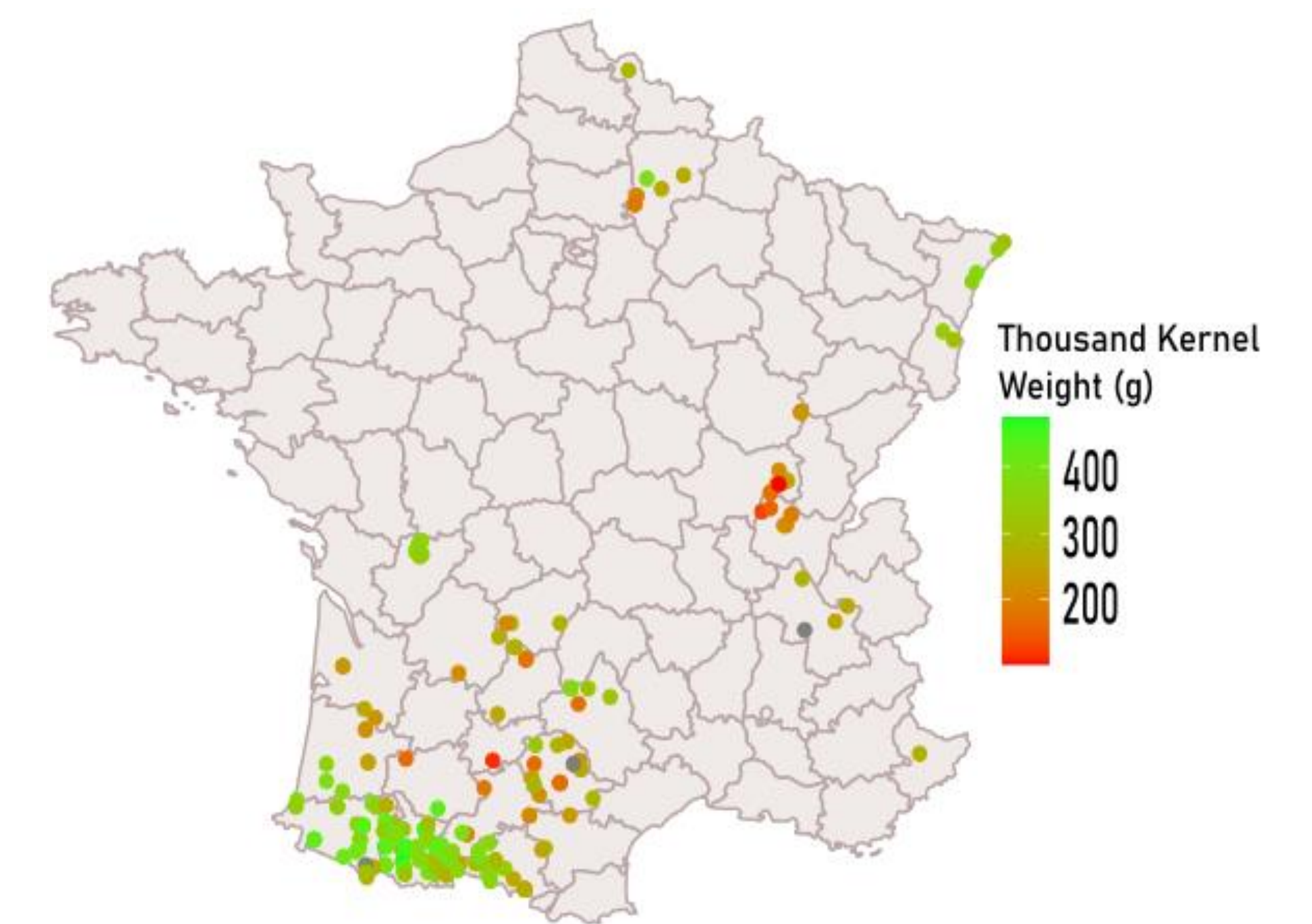

Figure 4: Geographic distribution of the 262 French landraces colored according to their thousand kernel weight

4. Genomic Predictions (GP)

GP performed using a GBLUP model on each trait separately

Precision of predictions evaluated using 2 methods:
- fivefold cross-validations
- leave-one-out approach (LOO)

Highly accurate predictions for agro-morphological traits (Figure 5): it ranges from 0.42 to 0.87 (0.70 on average)

Predictions of traits related to seed composition: quite accurate predictions on average, in particular for grain vitreousness traits (Figure 5)
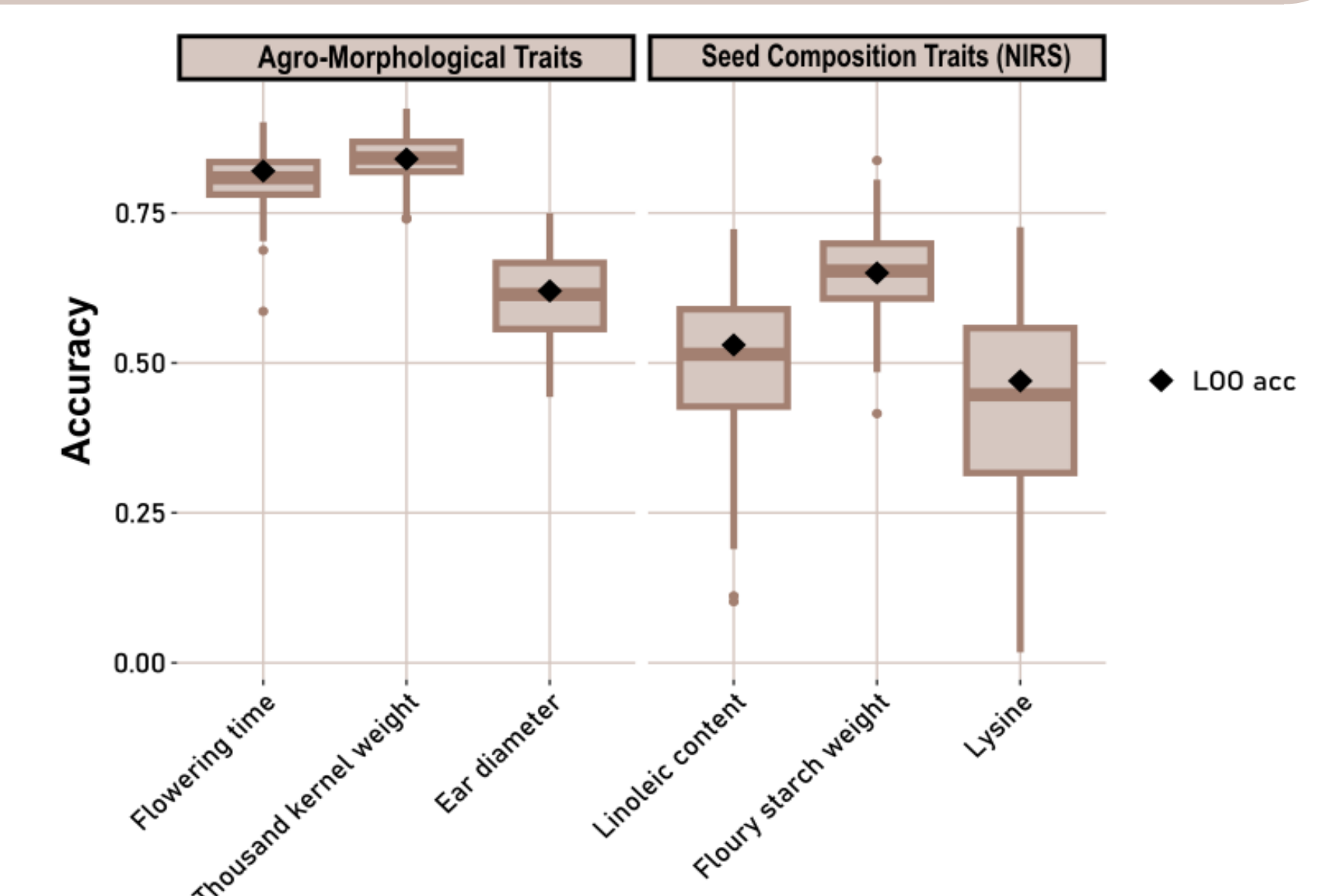

Figure 5: Accuracies (correlations between observed and predicted values) estimated using fivefold CV (boxplots) and LOO approach (dots)

5. Conclusion

GWAS and GEA approaches identified several genomic regions associated with agronomic traits or environmental and bioclimatic variables.

Some of the identified genomic regions colocalize with known genes related to the studied traits.

GP can be used to accurately predict agronomic traits of landraces, which could be a way to characterize large collections of landraces maintained in genebanks.

Our results suggest that DNA pool genotyping of traditional landraces is highly efficient to conduct GWAS and GP

References

Alexander, D. H.; Novembre, J.; Lange, K. Fast Model-Based Estimation of Ancestry in Unrelated Individuals. Genome Res. 2009, 19 (9), 1655–1664. https://doi.org/10.1101/gr.094052.109.

Arca, M.; Gouesnard, B.; Mary-Huard, T.; Le Paslier, M.; Bauland, C.; Combes, V.; Madur, D.; Charcosset, A.; Nicolas, S. D. Genotyping of DNA Pools Identifies Untapped Landraces and Genomic Regions to Develop next Generation Varieties. Plant Biotechnology Journal 2023, pbi.14022. https://doi.org/10.1111/pbi.14022.

Cericola, F.; Lenk, I.; Fè, D.; Byrne, S.; Jensen, C. S.; Pedersen, M. G.; Asp, T.; Jensen, J.; Janss, L. 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, 369. https://doi.org/10.3389/fpls.2018.00369.

Fick, S.E.; Hijmans, R.J. WorldClim 2 : new 1-km spatial resolution climate surfaces for global land areas. International Journal of Climatology. 2017. https://doi.org/10.1002/joc.5086

Gouesnard B; Dallard J.; Panouillé A.; Boyat A. Classification of French maize populations based on morphological traits. Agronomie, 17 9-10 (1997) 491-498. https://doi.org/10.1051/agro:19970906

Laporte F.; Charcosset A.; Mary-Huard T. Efficient ReML inference in variance component mixed models using a Min-Max algorithm. PLoS Comput Biol. 2022 Jan 24;18(1):e1009659. doi: 10.1371/journal.pcbi.1009659

Liu L.; Du Y.; Shen X.; Li M.; Sun W.; et al. KRN4 Controls Quantitative Variation in Maize Kernel Number. PLOS Genetics (2015) 11(11): e1005670. https://doi.org/10.1371/journal.pgen.1005670

VanRaden P.M. Efficient methods to compute genomic predictions. Journal of dairy science (2008) 91(11), 4414-4423. https://doi.org/10.3168/jds.2007-0980

Yan B.; Xu X.; Gu Y.; Zhao Y.; Zhao X.; He L.; et al. Genome-wide characterization and expression profiling of diacylglycerol acyltransferase genes from maize. Genome (2018), 61(10), 735–743. https://doi.org/10.1139/gen-2018-0029

Zheng P.; Allen W.B.; Roesler K.; Williams M.E.; Zhang S.; Li J.; et al. A phenylalanine in DGAT is a key determinant of oil content and composition in maize. Nature genetics (2008), 40(3), 367–372. https://doi.org/10.1038/ng.85