eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction

Abdou Rahmane Wade, Harold Duruflé, Leopoldo Sanchez, Vincent Segura

ANR-13-JSV6-0001

eQTLs are key players in the integration of genomic and transcriptomic data for phenotype prediction

Abdou Rahmane WADE[1], Harold Duruflé[1], Leopoldo Sanchez[1]* and Vincent Segura[2]*

[1] INRAE, ONF, BioForA, UMR 0588, F-45075 Orleans, France
[2] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398 Montpellier, France
[*] equal contribution

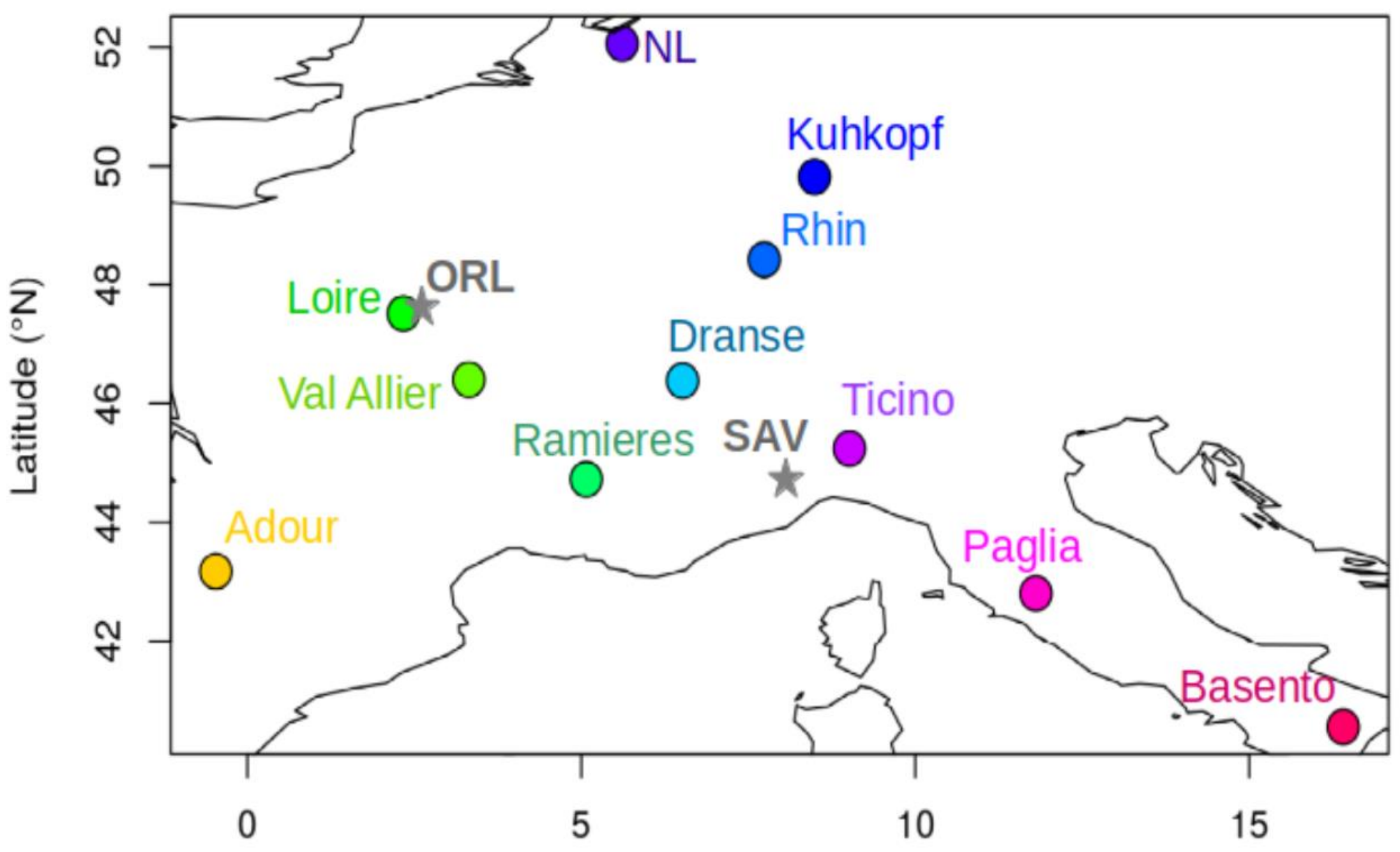
Multi-omics represent the "***missing link***" between phenotypes and genome variation. Few studies yet have addressed their integration to understand genetic architecture and improve predictability.

The mechanisms by which integration is successful when predicting phenotypes are still not known precisely over conditions and species, with **redundancy** between omics being one of the possible explanations. Redundancy reflects the interconnectivity from raw genomic sequence to the organismal phenotype.

Both redundancy and interactivity are key to **understanding genetic architecture** beyond the simple list of effects that is typically provided by genomic approaches.

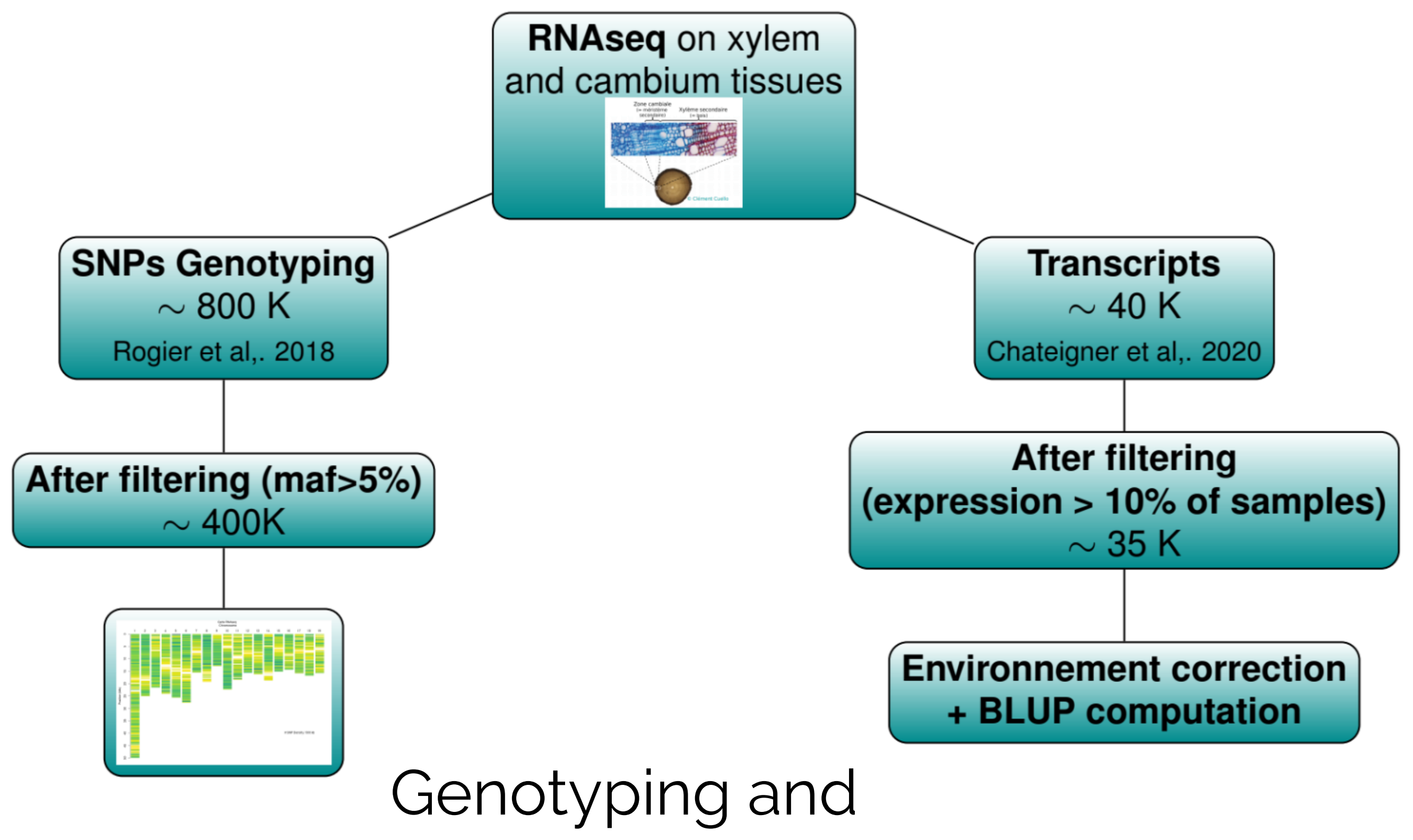
In our study: we propose new insights on data integration for black poplar (*Populus nigra* L.), using one of the simplest integration alternatives (**concatenation**), combined with one of the most popular prediction approaches (ridge regression). We aimed to evaluate the factors affecting prediction accuracy when integrating genomic and transcriptomic data for phenotype prediction using a large number of diverse phenotypes collected in two common gardens.

An original experimental design



241 genotypes representing the genetic diversity of the West Europe

Strategies & Objectives



RNAseq on xylem and cambium tissues

SNPs Genotyping
~ 800 K
Rogier et al., 2018

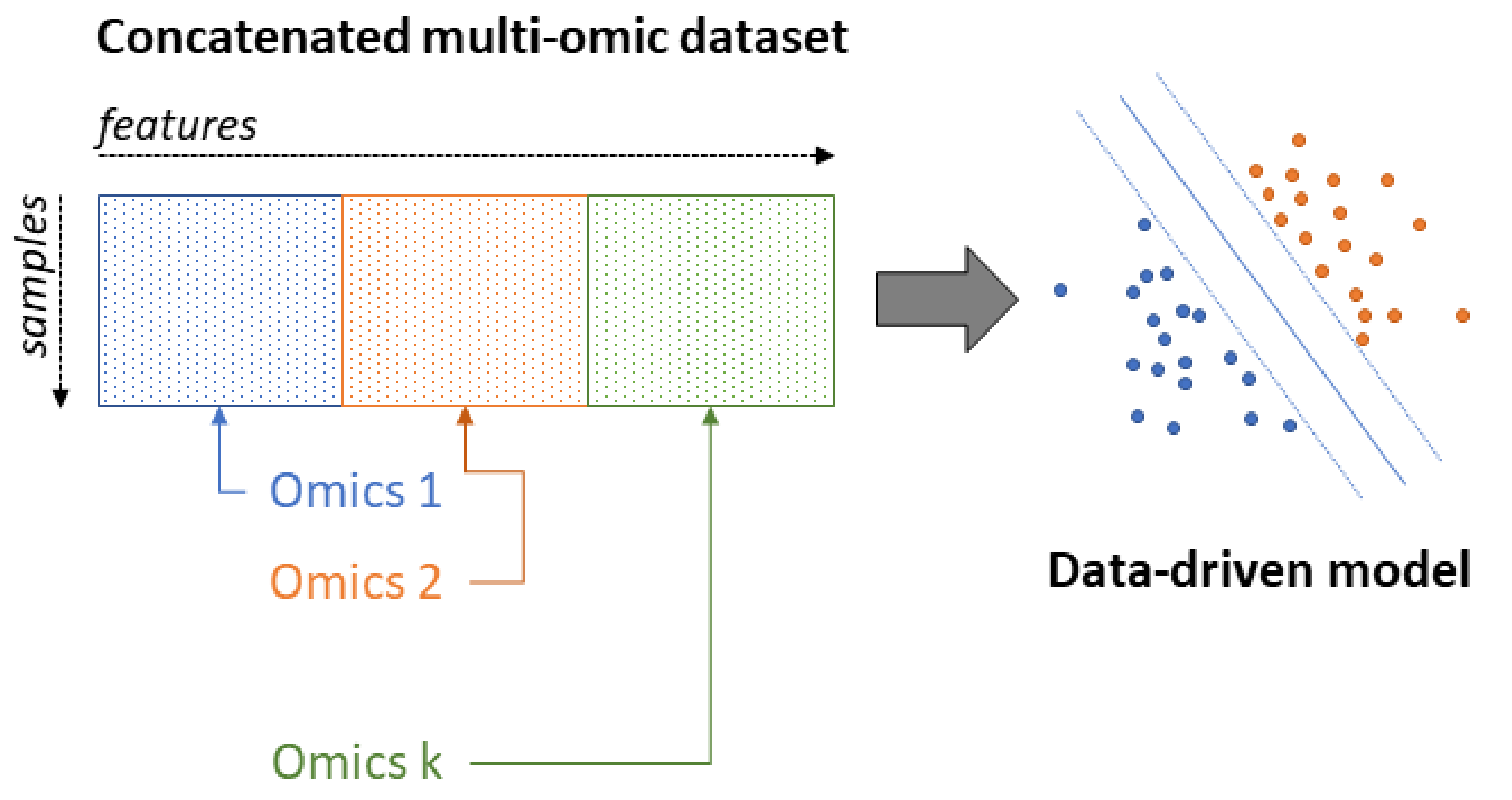
Transcripts
~ 40 K
Chateigner et al., 2020

After filtering (maf>5%)
~ 400K

After filtering (expression > 10% of samples)
~ 35 K

Environnement correction + BLUP computation

Genotyping and RNA sequencing



Concatenated multi-omic dataset

features

samples

Omics 1

Omics 2

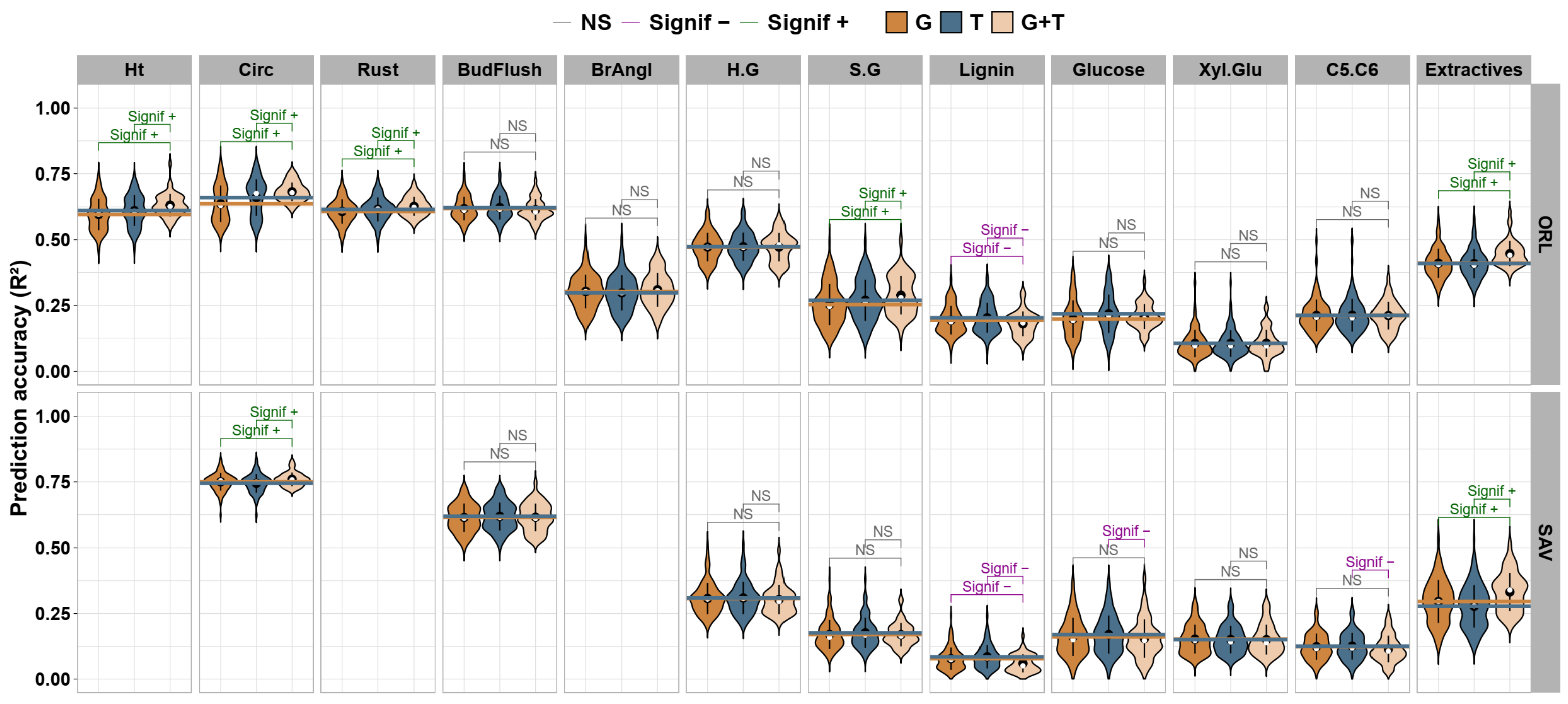
Omics k

Data-driven model

Concatenation of omics

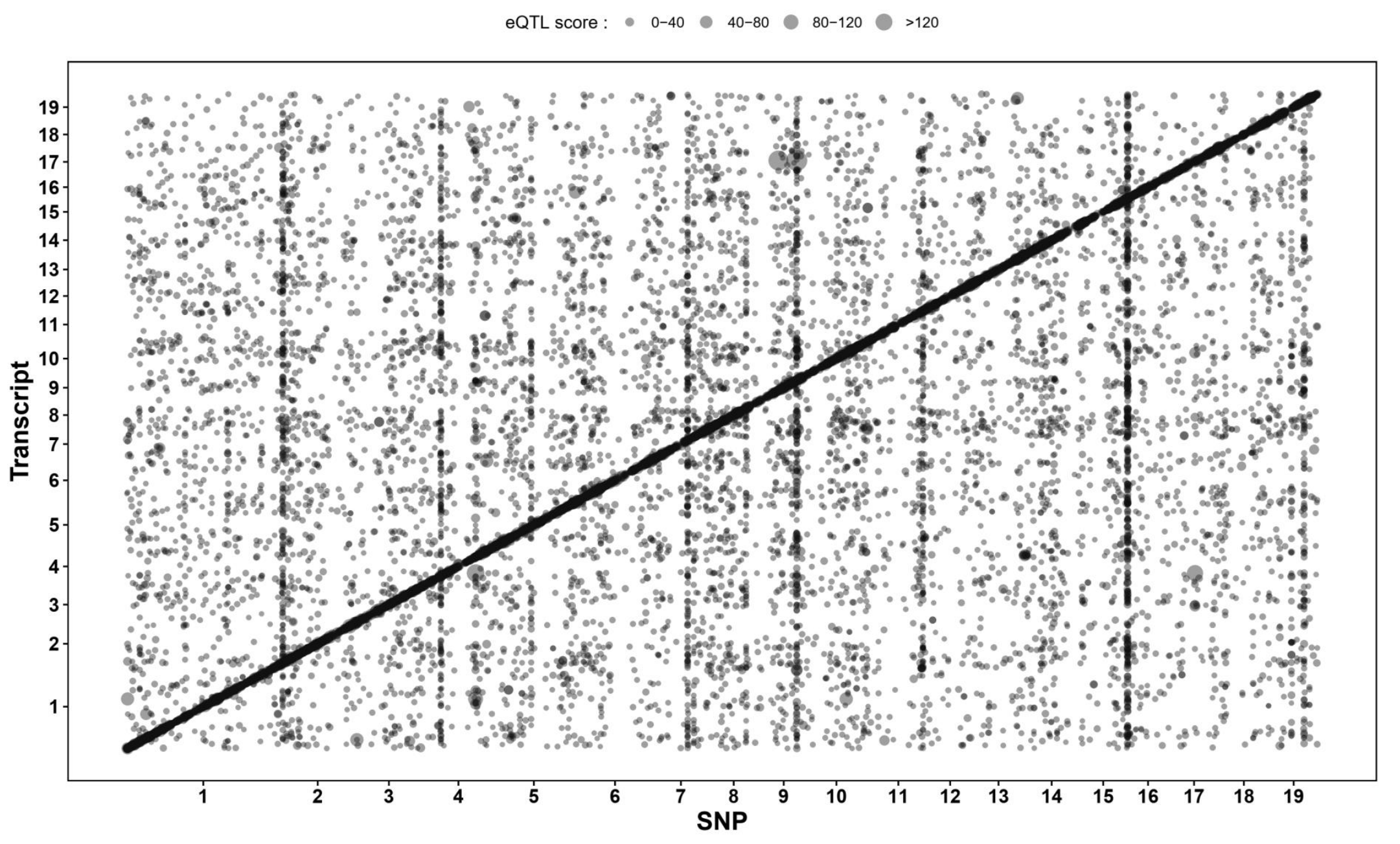
Results

1. Multi-omic model displays performance advantages over the single-omic models for specific functional types of traits.



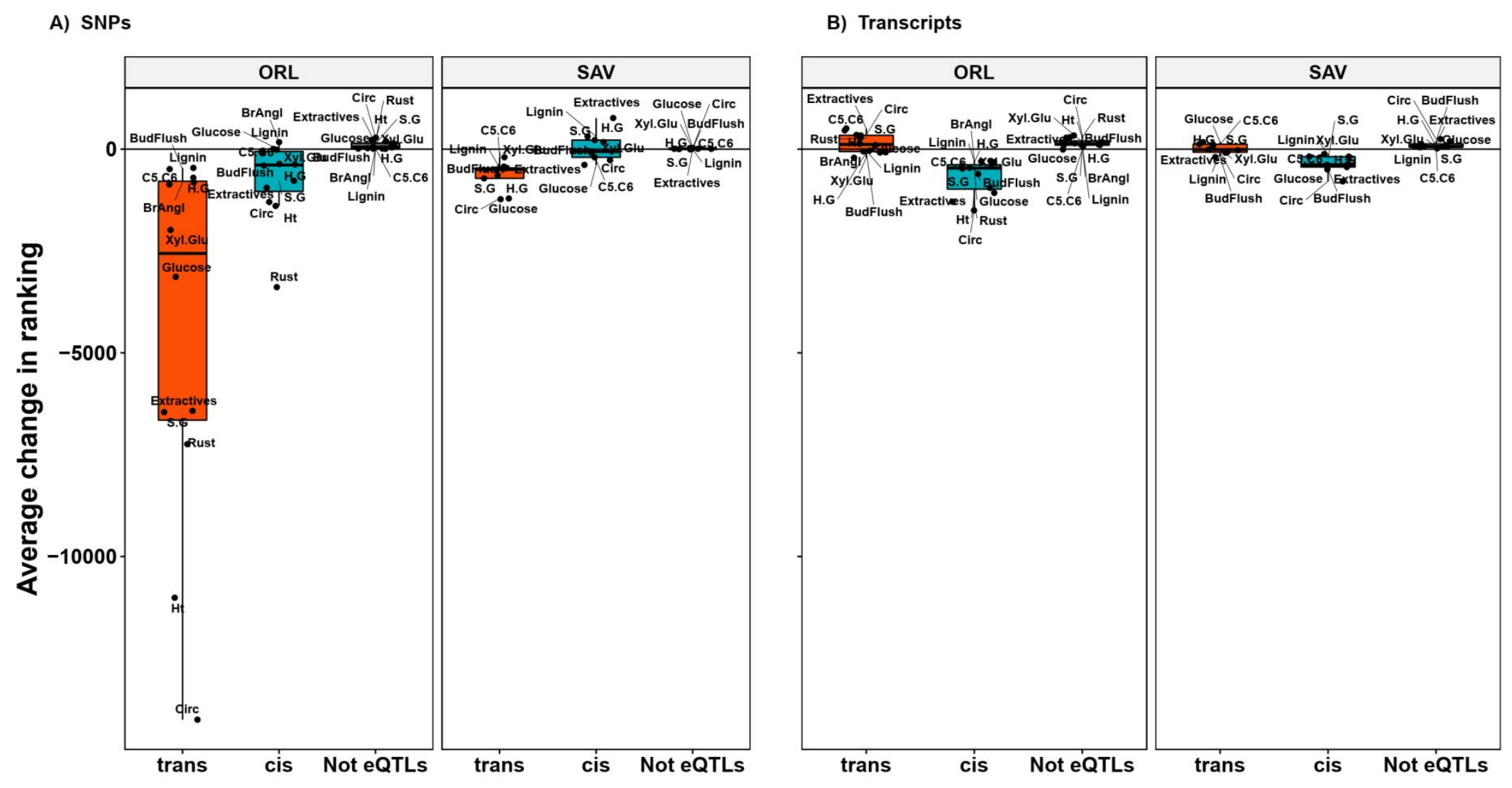
Prediction accuracies (R2) for 21 traits in the poplar dataset according to three models: genotypic data only (G model coloured in dark brown to the left in the panels), transcriptomic data only (T model coloured in dark blue), and concatenating both genotypic and transcriptomic data (G+T model coloured in light brown to the right). Distribution of accuracies resulted from a cross-validation scheme. Significance from paired tests is shown for comparisons between models, with a sign indicating if the accuracy was increased (+) or decreased (-) in the multi-omic model in comparison with the single-omic. The dark brown and dark blue horizontal lines represent the mean of precision distributions of G and T models, respectively.

2. eQTL analysis sheds light into the interplay between the genome and the transcriptome.



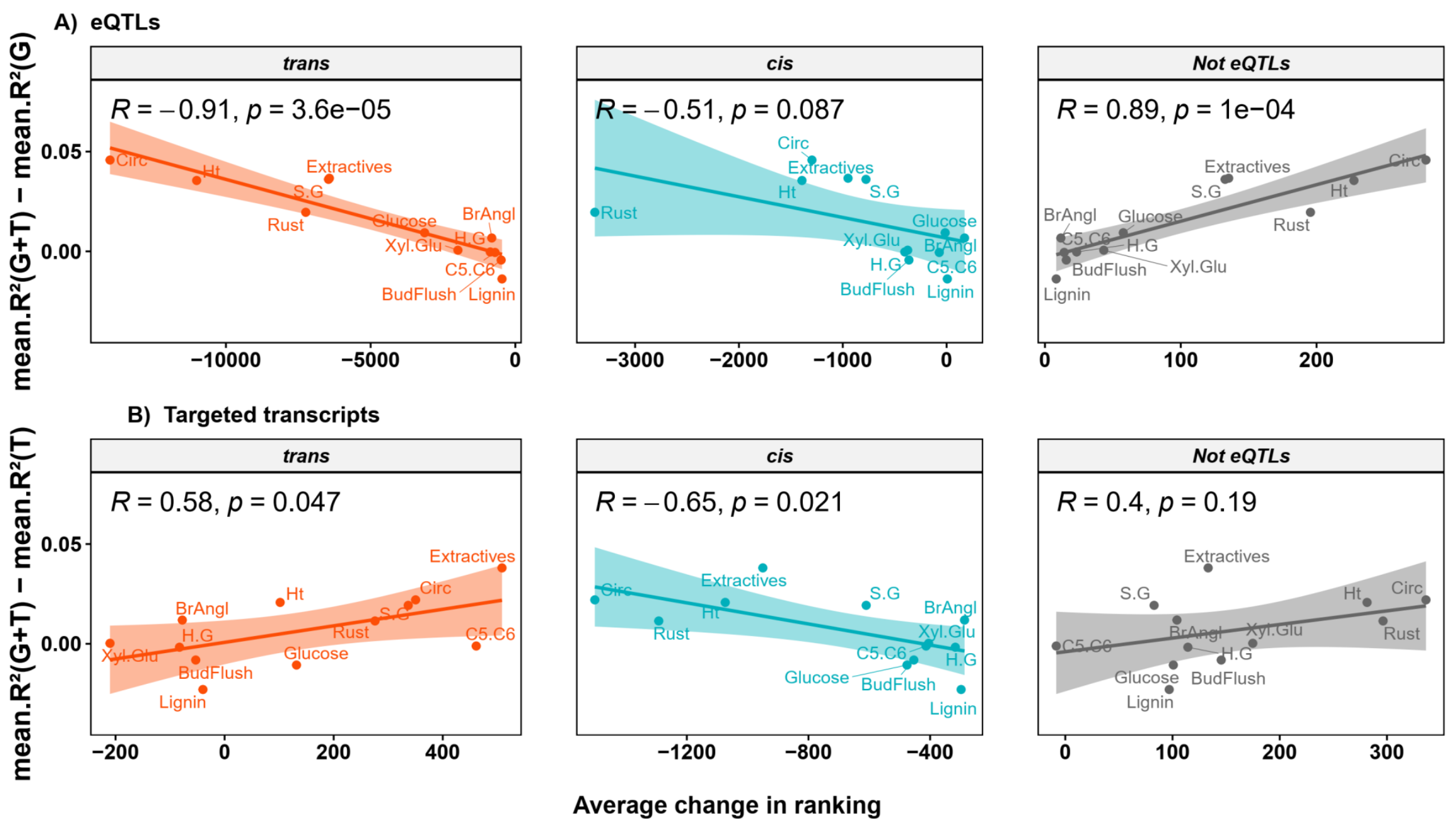
Map of associations (dots) between SNPs (x axis) and transcripts (y axis) through an eQTLs analysis with a multi-locus model. Dot size reflects the association score (-log10 of the p-value of the test) and dot positions correspond to genomic locations of transcripts and SNPs on the 19 chromosomes of the *Populus trichocarpa* reference genome (v3.0). The darkened diagonal includes all *cis* mediated associations, while the off-diagonal dots represent the *trans* associations.

3. Trans-eQTLs show the most important changes of squared effect rank between multi- and single-omic models.



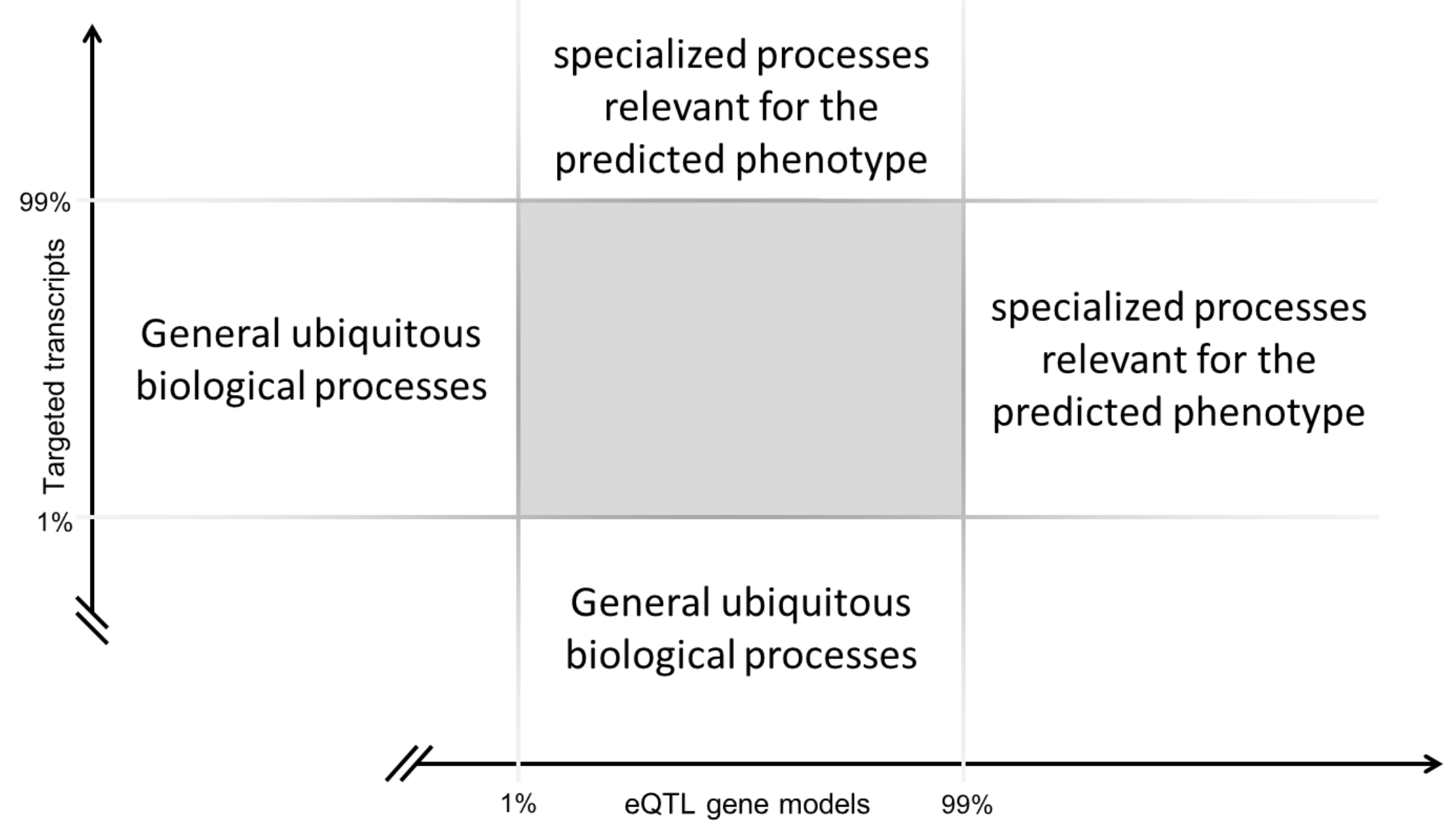
Boxplot of the average change in rank of SNPs (panels A) and transcripts (panels B). Each dot represents the average difference per trait, per site of the predictor ranks between the multi-omic model (G+T) and the single-omic models (G for SNPs and T for transcripts). The red and blue boxplots show the distribution of the average rank change for the trans-eQTLs and cis-eQTLs (A) or trans regulated transcripts and cis regulated transcripts (B), respectively. The boxplot in black shows the distribution for the predictors that have not been found to be associated in the eQTL analysis.

4. A negative relationship exists between the change in ranking of trans-eQTLs and cis-regulated transcripts and the predictive ability of the integrated multi-omic model.



Regression across traits measured at Orleans between average change in rank of predictors and advantage in performance of the multi-omic model (G+T) over the single- omic counterpart (G for SNPs and T for transcripts). The top panel (A) shows the regression obtained with the eQTLs (trans-eQTLs on the left, cis-eQTLs in the middle, and SNPs not detected as eQTL on the right). The bottom panel (B) shows the regression obtained with the regulated transcripts (trans on the left, cis in the middle, and not found to be associated with eQTLs on the right).

5. Gene ontology analysis suggests that top targeted transcripts or eQTLs are trait specific.



specialized processes relevant for the predicted phenotype

General ubiquitous biological processes

specialized processes relevant for the predicted phenotype

General ubiquitous biological processes

Schematic representation of the enriched GO terms among the top targeted transcripts or eQTL gene models list for the circumference of the tree trunk.

Conclusion

Consequently, beneficial integration happens when the redundancy of predictors is decreased, leaving the stage to other less prominent but complementary predictors.
An additional gene ontology enrichment analysis appeared to corroborate such statistical output. These two complementary approaches showed empirically over a series of traits how the best predicting scenarios are built, excluding certain features while promoting others according to their redundancies within the data.
To our knowledge, this is a novel finding delineating a promising method to explore data integration.

INRAE Val de Loire centre

BioForA
2163, Avenue de la Pomme de Pin, CS40001 ARDON, 45075 ORLEANS Cedex 2

This work is part of Abdou Wade PhD thesis and has been accepted for publication in BMC Genomics.
This study was funded by the European Union Horizon 2020 research and innovation program under grant agreement No 773383 (B4EST project).

Contacts :
leopoldo.sanchez-rodriguez@inrae.fr
harold.durufle@inrae.fr
vincent.segura@inrae.fr