1 **RESEARCH ARTICLE**

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

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

5 [1] INRAE, ONF, BioForA, UMR 0588, F-45075 Orleans, France

6 [2] UMR AGAP Institut, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34398
7 Montpellier, France

8 *Correspondence: leopoldo.sanchez-rodriguez@inrae.fr; vincent.segura@inrae.fr

9

10 **Short title:** Multi-omics phenotype prediction

11 **One-sentence summary:** Successful multi-omics integration when predicting
12 phenotypes makes redundant the predictors that are linked to ubiquitous connections
13 between the omics, according to biological and statistical approaches

14 **Abstract**

15 Multi-omics represent a promising link between phenotypes and genome variation. Few
16 studies yet address their integration to understand genetic architecture and improve
17 predictability. Our study used 241 poplar genotypes, phenotyped in two common
18 gardens, with their xylem and cambium RNA sequenced at one site, yielding large
19 phenotypic, genomic and transcriptomic datasets. For each trait, prediction models were
20 built with genotypic or transcriptomic data and compared to concatenation integrating
21 both omics. The advantage of integration varied across traits and, to understand such
22 differences, we made an eQTL analysis to characterize the interplay between the
23 genome and the transcriptome and classify the predicting features into CIS or TRANS
24 relationships. A strong and significant negative correlation was found between the
25 change in predictability and the change in predictor importance for eQTLs (both TRANS
26 and CIS effects) and CIS regulated transcripts, and mostly for traits showing beneficial
27 integration and evaluated in the site of transcriptomic sampling. Consequently,
28 beneficial integration happens when redundancy of predictors is decreased, leaving the
29 stage to other less prominent but complementary predictors. An additional GO
30 enrichment analysis appeared to corroborate such statistical output. To our knowledge,
31 this is a novel finding delineating a promising way to explore data integration.

32 **Keywords**: Genomic Prediction, omics, Multi-omics integration, eQTL, *Populus nigra*

Introduction

Genomic prediction, the prediction of phenotypes with genome-wide polymorphisms, has become a key tool to plant and animal breeders. This approach relies on statistical modelling to infer the effect of genomic variants, with many different modeling alternatives proposed in the literature (de los Campos et al., 2013; Varona et al., 2018). These models are mostly devised to predict the additive and transmissible contribution to individual genetic values, although dominance and epistatic interactions can also be accounted for Varona et al. (2018). Despite their success in identifying relevant effects and predicting phenotypes accurately, even in their most complex formulations, these models do not capture *per se* the genetic architecture of complex traits (Gianola, 2021). Beyond the statistics, it is the use of biological and functional information from the different organizational layers lying between the raw sequence and the organismal phenotype that will likely provide the required insights to reveal genetic architectures. Layers such as DNA methylation (Epigenome), transcripts (Transcriptome), proteins (Proteomics) or metabolites (Metabolites), are nowadays becoming increasingly accessible for many species, opening prospects towards a better understanding of the genetic architecture of complex traits.

In order to simultaneously account for these different layers of data in phenotype prediction, several integration approaches have been proposed (Ritchie et al., 2015). Among those, the most frequently used approach is the transformation or kernel-based integration, which consists in transforming each omics data into an intermediate form, usually taking the shape of a relationship matrix between the individuals (Guo et al., 2016; Schrag et al., 2018; Li et al., 2019; Morgante et al., 2020). Effects owing to different omics can then be integrated into a single analytical model, each effect being associated to a given kernel. Eventually, different kernels can be further combined by Hadamard product to add extra interaction terms between effects (Guo et al., 2016; Morgante et al., 2020). Integration can also be carried out across models, in what is known as model-based integration (Ritchie et al., 2015). Such integration can happen for a given omic type over different datasets or populations, each one summarized by its own model, with a final global model feeding on the top features contributed by each of the initial models. Another variant of the same model-based integration proceeds through a multistage approach, combining sequentially different omics for a given population (Ye et al., 2020). One of the simplest integration approaches, however, remains data concatenation (Azodi et al., 2020), by which multiple omics are placed side by side into a single large input matrix. Unlike kernels, whose results are produced at the individual level, the concatenation approach allows for the effects of multiple features at each omic to be estimated, whether they are SNPs, transcripts or any other omic. Another advantage, derived from that atomization of effects, is the fact that

71 interactions between omics can be more easily captured, without the risk of being lost
72 by intermediate transformations.

73 Most of the studies dealing with omics integration for phenotypic prediction have
74 focused on gauging predictive abilities. To that level, the reported benefits of integration
75 are context dependent across studies and, in general, amounting to small differences
76 when compared to single omics counterparts. A series of published comparisons in
77 maize illustrates this point. Using kernels to integrate genomic and transcriptomic data,
78 Guo et al. (2016) found improved accuracies over single omic counterparts for most of
79 the 11 economically important traits under study. Schrag et al. (2018), on the contrary,
80 found no benefit in integration following a similar approach and on a similar set of
81 production-related traits. For Azodi et al. (2020), however, using concatenation of
82 genomic and transcriptomic data for three maize traits yielded benefits only for one of
83 the traits. Studies on other biological models also showed similar context dependent
84 results. Based on the Drosophila melanogaster Reference Panel and different
85 transcriptomic datasets, Li et al. (2019) and Morgante et al. (2020) found subsequently
86 no benefit of integration following a multiple kernel approach in terms of predictive
87 abilities, and over different sets of fitness-related traits. When the integration included a
88 gene ontology (GO) category as an additional layer of information, accuracies were
89 surprisingly improved (Morgante et al., 2020). Using the same Drosophila panel,
90 however, Ye et al. (2020) found some benefits by following a model-based integration
91 approach, with a first modeling stage aiming at detecting SNP associated with
92 transcripts (eQTLs), and a subsequent prediction model focused on resulting eQTLs.
93 The number of studies, however, is not yet high enough to draw general conclusions.
94 Benefits might depend jointly on methods of integration and targeted traits, reflecting the
95 complexity of underlying architectures and conditions of studied populations.

96 Beyond the reported differences in prediction performance, there is still a scarcer
97 number of studies available that were able to pinpoint some of the possible causes
98 underlying the changes brought by integration to prediction. Already, at statistical level,
99 omics like sequence polymorphisms and transcriptomics are likely non-orthogonal to
100 some extent. If such redundancy is not conveniently handled at the model level, one
101 can expect inaccurate estimation of effects and impaired prediction accuracy as a result
102 (Farrar and Glauber, 1967; Ritchie et al., 2015). Redundancy between genomic and
103 transcriptomic data was addressed in a few studies, typically by gauging the amount of
104 extra variance captured by the different integration models compared to single omic
105 counterparts. For instance, the successful integration described by Guo et al. (2016)
106 was systematically accompanied by extra levels of captured variance, suggesting that
107 each extra layer added to the model contributed to some extent with non-redundant
108 information, and thus improving the prediction. The opposite behavior is described in
109 Morgante et al. (2020), with integrative models showing similar levels of captured

110   variance to those of single omics, indicating high levels of redundancy. It is interesting
111   to note here that for this latter study, redundancy was not found between GO terms, the
112   only layer bringing benefits to integration in the study. The most explanatory GO terms
113   with genomic data were different from those detected for transcriptomic data. The
114   second, more biological, approach is to look to what extent the most important features
115   in both omics show at the same time mutual associations, in other words, if relevant
116   SNPs are associated or not to relevant transcripts for a given phenotype. Azodi et al.
117   (2020) showed, in maize, that the transcriptome brings information on the phenotype
118   that is different from the one brought by genomic polymorphisms, by highlighting that
119   the information carried by the most important transcripts to predict flowering time is not
120   redundant with that carried out by the most important SNPs. In mice, two independent
121   studies used a Bayesian approach to model the phenotype with both genomic and
122   transcriptomic data (Ehsani et al., 2012; Takagi et al., 2014). These studies showed that
123   specific SNPs (eQTLs) associated with gene expression profiles can contribute to the
124   observed redundancy between the two data sources, which is reflected by the fact that
125   their importance for phenotype prediction was substantially affected under the
126   integrative approach.

127   Further research is needed to enrich the number of studies in data integration. It is clear
128   that the mechanisms by which integration is successful when predicting phenotypes are
129   still not known precisely and over a wide range of conditions and species, with the
130   hypothesis of redundancy being one of the possible explanations. To some extent,
131   redundancy reflects interactivity in the highly integrative space going from the raw
132   genomic sequence to the organismal phenotype. Both redundancy and interactivity are
133   key features to understand genetic architecture beyond the simple list of effects that is
134   typically provided by genomic approaches. Most of the available studies on data
135   integration involve model species like drosophila, maize, and notably humans. In the
136   present study, we proposed new insights on data integration for a species not frequently
137   found as subject of these approaches, black poplar, and using one of the simplest
138   integration alternatives (concatenation) combined to one of the most popular prediction
139   approaches (ridge regression). Here, we aimed at evaluating the factors affecting
140   prediction accuracy when integrating genomic and transcriptomic data for phenotype
141   prediction. Using a fairly large number of diverse phenotypes collected in two common
142   gardens for a collection of black poplars, we specifically analyzed the change in
143   importance of each of the potentially redundant sources (eQTLs and their target genes)
144   between a multi-omics model and the single omics counterparts, together with the
145   evolution of prediction accuracy. Under a more functional point of view, we further
146   studied the redundancy using a Gene ontology (GO) enrichment analysis.

Results

Multi-omics model displays performance advantages over the single omic ones for specific functional types of traits

Twenty-one traits of different types were phenotyped (**Table 1**) from 241 poplars grown in two common garden experiments located at contrasting sites (Orleans, France, and Savigliano, Italy). RNA sequencing analysis was also performed on young differentiating xylem and cambium tissues of the entire set of genotypes sampled in the common garden located at Orleans, resulting in large genomic (428,836 SNPs) and transcriptomic (34,229 transcripts) datasets. For each phenotypic trait, three ridge regression models were built: the first one with genotypic data as predictors (denoted G), the second one with transcriptomic data as predictors (T), and the third one with integration by concatenation of both omics data (G+T). The prediction accuracies for the three models varied across trait types, with growth, pathogen tolerance and phenology traits having average performances above 0.5 on both testing sites, while biochemical and architectural traits having average performances below 0.5 (**Figure 1**).

We compared for each trait the prediction accuracies of both the single omic models with the multi-omics model, and tested if they significantly differed with a paired Wilcoxon signed-rank test. Over all traits, the differences between the average accuracy of multi-omics model compared to the single omic models ranged from -0.025 to 0.054. Seven of the 21 traits showed a significant gain with the multi-omics model over both the single omic models. These 7 traits included all the growth traits, the pathogen resistance trait, as well as 3 of the 14 biochemical traits (S.G_ORL, Extractives_ORL and Extractives_SAV). It is noteworthy that most of these traits (5/7) were measured in Orleans, the site where transcriptomics data were also collected. The only 2 traits presenting an advantage for the multi-omics model at the Italian site (Circ and Extractives) were also advantageous on the French site. Some traits showed a significant loss of accuracy with the multi-omics model, two (Lignin_ORL, Lignin_SAV) when the comparison was against the G counterpart, and four (Lignin_ORL, Lignin_SAV, Glucose_SAV and C5.C6_SAV) when it was with T model. Of note, all these traits displaying a decrease in accuracy with the multi-omics model were biochemical traits, they had low prediction accuracies and were more often than not measured in Savigliano (3 in Italy versus 1 in France).

In summary, the multi-omics model showed performance advantages over the single omic models in 7 of the 21 traits, more frequently on traits measured in Orleans where transcriptomic data were collected than in the Italian site. The multi-omics model underperformed the single omic models on 4 occasions, corresponding to 3 traits

5

183 measured in Italy and one in Orleans. For the 10 remaining traits, no differences
184 between models were detected (**Figure 1 and Supplemental Table S1**).

185 **eQTL analysis sheds light into the interplay between the genome and the**
186 **transcriptome**

187 To further gain insight into the interplay between the two omics layers for phenotype
188 prediction, we studied their relationships through an eQTL analysis. Such analysis was
189 performed with two specific detection steps, the first ignoring linkage disequilibrium
190 between SNPs (called Step_0) and the second detecting multi-locus eQTLs (called
191 Step_opt). The resulting eQTLs at both steps were further classified into CIS and
192 TRANS regulatory elements according to their genomic proximity with the transcripts
193 there were associated with (for more details see the method section). **Figure 2** and
194 **Supplemental Figure S1** presents a map of these associations, respectively for
195 step_opt and Step_0, with dot size reflecting the eQTL score. The darkened diagonal
196 includes all CIS mediated associations, while the off-diagonal dots represent TRANS
197 eQTLs. It is important to note that some positions at the marker axis present highly
198 populated vertical trails across the genome, corresponding to important regulatory hubs.

199 For both detection stages, we found eQTLs for 10,242 out of the 34,229 transcripts
200 available in the transcriptomic dataset. Step_0 detected a total of 119,022 eQTLs on the
201 marker dataset, including 72,841 (61.2%) CIS regulatory elements and 46,181 (38.8%)
202 TRANS regulatory elements. At the optimal step of the eQTLs analysis (Step_opt),
203 which accounted for linkage disequilibrium between SNPs, we detected a total of
204 18,248 eQTLs, of which 7,845 (43%) were CIS regulatory elements, and 10,403 (57%)
205 were TRANS regulatory elements (**Supplemental Figure S2A**). CIS eQTLs displayed
206 on average a larger effect than TRANS eQTLs (**Supplemental Figure S2B**). This
207 point explains why there were so many at step_0, while their number drastically
208 decreased at step_opt when LD was accounted for by the multi-locus approach.
209 Whatever the step considered, the maximum distance between CIS eQTLs and their
210 associated genes ranged between 12 and 14 kb.

211 **The importance of predictors is less preserved for traits displaying a predictive**
212 **advantage with integration**

213 To get insights into the factors explaining the gain or loss in predictive ability when using
214 two omics by comparison with a single omic, we further looked at the variation in
215 importance of the individual predictors over the two types of predictive models. For each
216 of the three models, the importance of the different predictors was estimated as the rank
217 of their squared effects from the ridge regression models.

218 We looked at correlations between the importance of predictors across single and multi-
219 omics models, splitting the predictors into three categories determined from the eQTL
220 analysis: TRANS eQTLs or regulated transcripts, CIS eQTLS or regulated transcripts
221 and no eQTL. For SNPs, the correlation between the importances ranged from 0.62 to
222 0.99 across traits and predictor typologies. They were generally lower for the traits that
223 also showed advantages with the G+T model over single omic models, and for those
224 measured in the French site. Rust resistance, for instance, had the lowest correlations
225 across the different categories among all measured traits (0.62, 0.63 and 0.65,
226 respectively for TRANS eQTLs, CIS eQTLs and non eQTLS). Also, growth traits
227 showed relatively low correlations compared to most of the traits, although this
228 happened only for measurements in the French site (Ht_ORL, Circ_ORL), with those in
229 the Italian site (Circ_SAV) being much higher and comparable to the top correlations.
230 For the remaining traits, correlations between importances remained high, generally
231 above 0.9 but with a few values close to 0.8 (**Supplemental Figure S3A**). The
232 correlations between transcript importances (**Supplemental Figure S3B**) were
233 generally lower than those for SNPs, varying between 0.52 and 0.89 across traits and
234 predictor categories. Following a similar pattern as for SNPs, the traits showing the
235 lowest correlations were also those for which the multi-omics displayed a predictive
236 advantage over single omic models, as well as those measured in the French site.
237 Growth and rust resistance traits were those showing the lowest correlations. Although
238 with small differences, CIS-regulated transcripts showed lower correlations than those
239 from TRANS-regulated counterparts, across traits and sites.

240 **TRANS-eQTLs show the most important changes of squared effect rank between**
241 **multi-omics and single omic models**

242 Previous correlations pinpointed to changes in importance in some of the categories of
243 predictors. Such changes can be illustrated by the difference in importance (rank of
244 squared effect) between the multi-omics model and that of the single omic model (either
245 T or G, for transcripts and SNPs respectively) (see Methods for details).

246 When looking at the variation of the differences in importance (**Supplemental Figure**
247 **S4, Figure S4**), the amounts were much larger for eQTLs (G+T versus G) than for
248 targeted transcripts (G+T versus T). Higher variations were also found for TRANS
249 eQTLs than for CIS counterparts, and for traits phenotyped at Orleans than for those in
250 the Italian site. Thus, changes in importance occurred with more intensity for eQTLs,
251 with a TRANS regulation, and linked to traits measured where the transcripts were
252 sampled.

253 An alternative way of visualizing those changes is represented in **Figure 3**. This time,
254 changes were averaged for a given trait and the resulting distribution of averages

255 represented by predictor category and site. Patterns were very different between eQTLs
256 and targeted transcripts, and also between sites. The most important changes in
257 ranking happened at the Orleans site. With respect to predictor typologies, it was
258 TRANS-eQTLs that showed the most important changes, with an overall loss of
259 importance when switching to the G+T model, notably for the traits benefiting the most
260 in performance from concatenation (growth and rust resistance). Less conspicuous
261 were the changes for CIS-eQTLs, overall of negative sign but of lesser magnitude. Non-
262 eQTLs showed generally small changes across traits. For targeted transcripts, the most
263 impacted typology was CIS regulated genes, with an overall loss in ranking across
264 traits, notably for growth and rust resistance traits.

265 A negative relationship exits between the change in importance of eQTLs and CIS
266 regulated transcripts and the predictive ability of the integrated multi-omic model

267 **Figure 4** represents the link across traits between average change in importance of
268 predictors and advantage in performance of multi-omics over the single omic
269 counterpart. In the case of eQTLs-TRANS, generally the most affected predictors
270 following concatenation, a significant relationship (r=-0.81, p=0.0015) can be drawn
271 where gains in prediction occurred at the expense of losses in ranking of predictors. A
272 similar pattern, although of lesser magnitude, is to be found for eQTLs-CIS and CIS
273 regulated genes (r=-0.6, p=0.037 and r=-0.64, p=0.024, respectively). No significant link
274 was found for TRANS regulated genes. An equivalent representation for the Savigliano
275 site showed no significant links across categories of predictors (**Supplemental Figure**
276 **S6**).

277 Gene ontology analysis suggests that top targeted transcripts or eQTLs are trait
278 specific

279 We further selected the transcripts or eQTLs whose importance was most affected
280 through data integration, by focusing on the 1% percentile at each extreme of the
281 distributions, and carried an enrichment analysis in GO terms on the resulting features
282 (**Figure 5 and Supplemental Table S2**).

283 For all type of traits, the analysis of the GOs showed enrichment of terms from general
284 cell cycle process (e.g. "regulation of RNA export", "regulation of nucleobase", "positive
285 regulation of RNA, vesicle-mediated transport") in the lists of eQTL gene models
286 selected as having the most negative impact on their importance during integration. The
287 same results were visible for all traits with the lists of targeted transcripts selected with a
288 negative effect, the GO terms enrichment were associated to ubiquity process like
289 "protein targeting to chloroplast" or "protein localization to chloroplast" in the

290   circumference of the tree trunk study (**Figure 5A**) or "phosphorelay signal transduction"
291   and "cellular response to ethylene" for the lignin content study (**Figure 5B**).

292   For traits that showed significant gain with integration, this analysis suggests that the list
293   of targeted genes with the most positive effects under integration are enriched with
294   different terms specific to the traits. For example, we found cell wall related terms like
295   "cell wall polysaccharide catabolic process", "xylan metabolic process" for the
296   circumference of the tree (**Figure 5A**). The same results were found for the eQTL gene
297   models selected with the most positive effects with specific GO terms for tree
298   circumference like "formation of plant organ" or "formation of anatomical boundary".

299   On the contrary, in traits that showed significant predictive losses with multi-omics
300   model over single omic ones, like the lignin content (**Figure 5B**), showed also
301   enrichment in GO terms of general process for the most positive effect targeted
302   transcripts or eQTL gene models (e.g. "nucleic acid metabolic process", "gene
303   expression" for the targeted transcripts and "cellular lipid catabolic", "fatty acid catabolic
304   process" for the eQTL gene models).

Discussion

306   In this study, we used 21 traits to compare the relative advantages of integrating
307   genomic and transcriptomic data for phenotype prediction versus using each omic
308   separately. This relative advantage of integration over single omic varied across traits.
309   For traits such as growth and pathogen resistance, integration yielded more accurate
310   predictions than the single omic counterparts, while for most of the others, basically
311   biochemical traits, no difference was detected, with still a few cases of underperforming
312   concatenation. By using a simple modeling approach like ridge regression, we showed
313   that gains in the traits benefited by integration were associated with systematic changes
314   in importance for some specific predictors, and that those predictors were involved in
315   the interplay between SNP polymorphisms and transcripts, pinpointing at adjustments in
316   effects due to redundancies. Such findings at the statistical level were also backed up
317   by a subsequent biological analysis of GO terms.

318   In order to better understand the reasons underlying trait differences in the benefits of
319   concatenation, we sought to evaluate the interplay between the genomic and
320   transcriptomic data, by making use of an eQTL analysis. Such analysis allowed us firstly
321   to categorize the predictors into CIS eQTLs, TRANS eQTLs, non eQTLs, CIS regulated
322   transcripts, TRANS regulated transcripts and transcripts with no eQTL detected.
323   Secondly, based on such categorization, we could quantify the changes in predictor
324   importance for each of these categories when using the multi-omics model by
325   comparison with single omic ones. Over all the traits under study, we found a strong

326 negative and significant correlation between the relative advantage of the multi-omics
327 model compared to the single omic ones and the drop in importance of the predictors
328 for eQTLs (R=-0.81 for TRANS eQTLs and R=-0.6 for CIS eQTLs) and transcripts
329 regulated in CIS (R=-0.64). Such a relationship could be interpreted in terms of
330 redundancy between predictors coming from different omics. Indeed, the traits that
331 benefited the most from concatenation were also those for which TRANS eQTLs and
332 CIS-regulated transcripts lost the most in terms of importance in the combined
333 predictive model compared to the single omic counterparts, and thus for which the
334 combined model decreased the redundancy between predictors by down weighting
335 those that specifically matter for the eQTL versus transcript covariation. Redundancy,
336 *per se*, would not necessarily explain gains or loss in performance, but down weighting
337 redundant predictors could allow other minor predictors, otherwise silenced, climb in
338 importance in such a way that the concatenation model improved in predictability.

339 To our knowledge, this study is the first to establish such a relationship between
340 integration success and redundancy between omics layers, pinpointing eQTLs as key
341 players in such interplay. It is worth mentioning that we could establish such a
342 relationship because of the relatively large number of traits under study, compared to
343 previous works (Ehsani et al., 2012; Guo et al., 2016; Morgante et al., 2020; Azodi et
344 al., 2020; Li et al., 2019; Takagi et al., 2014; Schrag et al., 2018). The relative gains
345 from integration ranged from -0.02 to +0.05 R² across all 21 traits. These gains are
346 indeed small, but are consistent with the state of the art. However, our objective here
347 was to attempt to understand the factors that underlie this gain in order to produce new
348 knowledge that will allow us to improve in more consequent ways the advantage of
349 integration with other methods.

350 **A gain of integration was mainly found for traits related to growth and for traits**
351 **evaluated in the same location as transcriptomic evaluation**

352 We observed a significant advantage of multi-omics over single omic for all growth-
353 related traits (Ht_ORL, Circ_ORL and Circ_SAV). Since growth results from cell division
354 and expansion in the apical and cambial meristems (Xylem and cambium) (Chaffey et
355 al., 2002) this relationship between the tissues from which we extracted transcripts and
356 the growth traits (circumference and height) may explain the significant integration
357 advantage observed for these traits.

358 The advantage of integration over the single omic models was also observed for leaf
359 rust resistance. Although xylem and cambium, the tissues sampled for RNA
360 sequencing, seem disconnected to a phenomenon occurring at the leaf level, the
361 relationship here is likely indirect since links between resistance and growth have been
362 reported by other studies (Wang and Kamp, 1992; Steenackers et al., 1996; Newcombe

363 et al., 2001). Following the same reasoning, phenology and architectural traits
364 considered here do not show clear relationships with cambial meristems or with growth-
365 related traits, and therefore support the lack of benefits observed for them in the
366 concatenation models.

367 For the majority of wood biochemical traits, the multi-omics integration model performed
368 similarly or worse than the single omic model. It is noteworthy that these traits are
369 overall not well predicted with single or multi-omics, and they originally come from near
370 infrared spectroscopy prediction which may have included some noise to their variation.
371 We could hypothesize that this factor underlies the observed poor performance for this
372 category of traits during integration.

373 These series of observations across traits also point to the idea that transcripts capture
374 some new information not necessarily available for SNPs, such as that associated to
375 the genic interplays occurring at the specific tissue sampled for transcripts, which could
376 be of a non-additive nature (gene-gene interactions), or even genotype-by-environment
377 interaction effects, which are both not explicitly modelled when using exclusively SNPs
378 as predictors.

379 Among the 21 analysed traits, we observed that the benefits of the concatenation model
380 happened more often for the traits measured in Orleans where the transcriptomic data
381 were also collected than for those in the Italian site. This advantage when phenotypic
382 and transcriptomic data evaluation are carried out at the same location can be
383 interpreted in terms of genotype-environment interactions effectively captured by the
384 transcripts (Buil et al., 2015; Idaghdour and Awadalla, 2013). Conversely, for
385 phenotypes evaluated at the Italian site, the transcriptomic data more likely bring
386 redundant information to that of SNPs, which in turn do not result in any advantage in
387 the multi-omic integration.

388 **Negative change in rank between the two models implies an increase in**
389 **importance**

390 Our goal was to identify factors that influence the success of genomic and
391 transcriptomic data integration for phenotype prediction. To this end, we chose a simple
392 integration method that allows us to track changes in the importance of each variable
393 between the integration model and the single omic models. As described in Zampieri et
394 al. (2019) and Ritchie et al. (2015), there are several ways to integrate multi-omics data,
395 the simplest being the concatenation method. Using a ridge regression model,
396 concatenation allows to directly estimate each predictor effect, accounting for all other
397 variables (SNPs and transcripts), unlike LASSO and elastic-net where some degree of
398 variable selection is applied, while trying to minimize the covariation between the
399 predictors' effects. This method allowed us to track the evolution of the relative

400 importance of the predictors in the multi-omics model compared to the single models,
401 and therefore infer potential redundancies by the changes in importance. Since the
402 effects of SNPs or transcripts between the two models are not at the same scale to
403 gauge importances, we had to bring them to a common scale. A simple and efficient
404 way to do this is to work with ranks of the squared effects of predictors.

405 Comparing the changes in ranks between the multi-omics model and the single omic
406 ones informed us about the gain or loss in importance of each predictor. A predictor will
407 have a positive change in rank when it has high importance (low rank) under the single
408 omic models and ends up with low importance (high rank) in the multi-omics model.
409 Conversely, a negative change in rank between the two models implies an increase in
410 importance. A zero rank change corresponds to a predictor that keeps the same
411 importance between the two models.

**Integration success is driven by the loss in importance of covariation sources
between genomics and transcriptomics**

414 Our main hypothesis was that sources of redundancy between SNPs and transcripts
415 play an important role in the success of the integration. The ideal candidates as a
416 source of redundancy between SNPs and transcripts are eQTLs and the genes they
417 regulate, so we performed an eQTL analysis to identify eQTLs (CIS and TRANS) and
418 their regulated genes among our dataset. In order to remain in the same framework as
419 in the ridge prediction models, we used the results of the eQTL analysis for which
420 linkage disequilibrium was not taken into account, which enabled us to get information
421 at the SNP level rather than at the locus level. However, the SNPs in our dataset are
422 derived from RNAseq and are representative of the functional space of the genome,
423 thus capturing few SNPs in the intergenic spaces. This might have affected our ability to
424 detect some TRANS-eQTLs. Nevertheless, our multi-locus analysis showed that
425 TRANS-eQTLs remained the majority, with some hotspot or hub loci associated with a
426 fairly large number of transcripts. Such behavior has previously been reported in other
427 species such as yeast (Albert et al., 2018) or maize (Liu et al., 2017; Swanson-Wagner
428 et al., 2009).

429 The main results of our study is the strong negative and significant correlation found
430 between the relative advantage of the multi-omic model over the single omic ones and
431 the average importance losses of the eQTLs (more pronounced for TRANS) and the
432 genes regulated in CIS. It is important to note here that CIS-regulated genes are on
433 average regulated by more eQTLs than their TRANS-regulated counterparts, 10.15 and
434 6.63 eQTLs respectively (**Supplemental Figure S7**), suggesting that CIS-regulated
435 genes are the source of more redundancies than TRANS-regulated genes in our
436 dataset. To our knowledge, this study is the first to establish this direct relationship

437   between integration success and losses in the importance of eQTLs and regulated
438   genes. This relationship was possible to establish due to the relatively large number of
439   diverse traits that we used.

440   There are results obtained in other studies that indirectly suggest the importance of
441   eQTLs for multi-omics integration between genomics and transcriptomics. Ehsani et al.
442   (2012) observed in mice losses in importance of eQTLs in the combined genomics and
443   transcriptomics model versus the model with only genomics for the phenotype that
444   shows an advantage of integration (body mass). Such behavior of eQTLs in this study
445   was observed only for a single phenotype with low resolution genotyping data. Also, Ye
446   et al. (2020) were successful in improving the performance of phenotype prediction in
447   Drosophila using genotypes of eQTLs regulating genes that are important for the
448   phenotype. They proceeded with successive selection steps involving a transcriptome
449   wide association study (TWAS) with an eQTLs analysis for the TWAS significant genes,
450   while optimizing the detection thresholds of these two analyses. Their results indirectly
451   suggest the importance of eQTLs for the integration between genomics and
452   transcriptomics. The negative correlation between the relative advantage of the multi-
453   omics model over the single omic ones and the average losses in importance of
454   covariation sources suggests that the integration success is driven by a minimization of
455   the redundancy between genomics and transcriptomics. Azodi et al. (2020) observed in
456   maize that concatenation between genomic and transcriptomic data improves the
457   prediction of one of their 3 studied phenotypes. For this phenotype, they showed that
458   the most important SNPs and transcripts were not redundant in the sense that they
459   were not located in the same genomic regions, nor were they regulators of important
460   transcripts.

461   **The observed redundancy may be explained by biological processes**

462   Up to now, we have shown statistically that the sources of redundancy were penalized
463   with a weaker importance under integration. Our GO enrichment analysis provided a
464   more biological point of view to bring extra evidence of the role of redundancy.
465   Generally speaking, genes pointing at general ubiquitous biological processes were
466   more likely the source of redundancy, while those associated with specific processes
467   could more easily bring extra useful information to the prediction process. The GO
468   analysis showed that genes gaining in prediction importance under integration were
469   generally associated with specialized processes of relevance for the predicted
470   phenotype. This pattern was observed notably for traits related to wood production. On
471   the contrary, the genes that were most heavily affected in their importance under
472   integration showed a characteristic enrichment of terms linked to the general cell cycle
473   processes. As the transcriptomic data came from young differentiating xylem and
474   cambium tissues, the redundancy (and complementarity) that we observed is strongly

13

475 associated with phenotypes related to the production of wood, like the trunk
476 circumference. This interpretation of our results might also apply to the loss of prediction
477 for traits whose genes are not likely to be represented in our transcriptome, such as
478 those related to rust resistance for example. One eventual validation could be to
479 complement the transcriptomic data with alternative collection on tissues other than
480 those closely connected to xylem and cambium, like leaves, and work on traits more
481 specifically expressed on the collected tissue (i.e. rust resistance). If the genes
482 associated with the general biological processes are found to be sources of redundancy
483 through GO analysis, one strategy to improve prediction could be to reduce or minimize
484 their contribution to the models.

Perspectives

486 One of the main findings of this study is the fact that certain predictors with ubiquitous
487 connections seem to be made redundant when integration takes place, leaving the
488 stage for other features to be picked up, eventually less prominent but bringing true
489 complementarity to the integrative prediction. For the sake of simplicity, our study could
490 not take the extra step to devise a novel alternative to account for such redundancies.
491 However, it would be quite straightforward to outline a basic strategy where the
492 importance of predictors going into the model is penalized according to some function
493 describing their redundancy in the data. Under kernel-based integration, for instance,
494 some kind of optimization of composition in features included in the relatedness
495 matrices could be devised so that the resulting kernels bring complementary
496 information. Under a model-based integration, a multistage approach could be devised
497 where associations between all involved omics are firstly carried out, so that the
498 features contributing the most to the associations can be subsequently penalized to
499 some degree or filtered out when it comes to construct a consensual model. More
500 research would be required to devise and test a strategy to derive robust weightings.

501 It would be essential to gather extra information on the beneficial role of multiple omics,
502 collected at different development stages or distinct tissues, so that links to different
503 traits can be drawn. This is certainly a costly endeavor, which could be focused on
504 specific training populations. Ideally, integration studies on those training sets would
505 allow us to identify important hubs in the genetic architecture of traits, and use that
506 information for differential weighting on other related populations with no or basic
507 access to extra omics layers.

Material and methods

Plant material, experimental design and phenotypic evaluation

14

510 We studied 241 genotypes of *Populus nigra* originated from 11 major river catchments
511 across 4 countries and representative of the species range in Western Europe. These
512 poplars were evaluated in common garden experiments located on 2 contrasting sites
513 (Orleans noted ORL and Savigliano noted SAV) (Guet et al., 2015). In each site, the
514 experimental design consisted in a randomized complete block design with 6 blocks,
515 and thus 6 repetitions per genotype. Twelve traits were evaluated on the 2 sites, as
516 previously described (Gebreselassie et al., 2017; Chateigner et al., 2020). We
517 considered traits measured in 2 sites as different traits, leading to a total of 21 traits
518 (detailed in **Table 1**). These traits can be categorized into 5 types: growth, pathogen
519 tolerance, phenology, architecture, and biochemistry. At Orleans, the trees were grown
520 through 3 successive cycles: 2008-2009, 2010-2011 and 2012-2015. During the first
521 growth cycle (2009), rust tolerance (Rust) was measured with a discrete score from 1
522 (no symptom) to 8 (generalized symptoms), as detailed in Legionnet et al. (1999**).**
523 Average branch angle (BrAngl) was evaluated with a score on proleptic shoots from 1 to
524 4 (score 1: between 0° and 30°; score 2: between 30° and 40°; score 3: between 40°
525 and 55°; score 4: and between 55°and 90°). During the second growth cycle, height (Ht)
526 and circumference at 1-meter above the ground (Circ) were measured on 2 year-old
527 trees (winter 2011). At Savigliano, trees went through two cycles: 2008-2009 and 2009-
528 2010. Only Circ was measured during the second growth cycle on 2 year-old trees
529 (winter 2010). Biochemical traits consisted in predictions of several chemical
530 compounds obtained from near-infrared spectra on wood samples collected in the same
531 years as growth traits and at both sites, as described in Gebreselassie et al. (2017).
532 Biochemical traits included: extractives content (Extractives), total lignin content
533 (Lignin), ratios between different lignin components like p-hydroxyphenyl (H), guaiacyl
534 (G) and syringyl (S) (H.G, S.G), total Glucose content (Glucose), ratio between Xylose
535 and Glucose content (XylGlu) and the ratio between 5 and 6 carbon sugars (C5.C6).
536 One phenological trait was also measured, BudFlush as discrete scores for a given day
537 of the year, measured on the apical bud (Dillen et al., 2009).

Phenotype adjustments

539 All 21 traits were independently adjusted to field micro-environmental heterogeneity
540 using the breedR package (Munoz and Sanchez, 2017). The model included blocks and
541 spatial effects (autoregressive residuals function) to account for micro-environmental
542 heterogeneity. Also a model selection was carried out using the AIC to select the effects
543 to be included in the model and to tune the autoregressive parameters. The genotypic
544 adjusted means from these models were used as phenotype for this study.

Genotype and transcriptomic data

546 RNA sequencing was carried out in 2015 on young differentiating xylem and cambium
547 tissues collected on two replicates of the 241 genotypes located into two blocks of the
548 Orleans design (Chateigner et al., 2020). We obtained sequencing reads for 459
549 samples corresponding to 218 genotypes with two replicates and 23 genotypes with 1
550 replicate. These sequencing reads were used to provide both transcriptomic and
551 genomic data.

552 For transcriptomic data, the reads were mapped on the *Populus trichocarpa* v3.0
553 primary transcripts and read counts were retrieved for 41,335 transcripts. Only
554 transcripts with at least 1 count in 10% of the individuals were kept, yielding 34,229
555 features. The raw count data were normalized by Trimmed Mean of M-values using the
556 R package edgeR v3.26.4 (Robinson and Oshlack, 2010) and we calculated the counts
557 per millions (Law et al., 2014). To make the CPM data fit a Gaussian distribution, we
558 computed a *log*2(*n*+1) instead of a *log*2(n+0.5) typically used in a voom analysis (Law et
559 al., 2014), to avoid negative values, which are problematic for the rest of the analysis.
560 For each transcript the log2(n+1) of the CPM were fitted with a mixed model including
561 experimental (batch) and genetic effects to extract their genotypic blups. Those
562 transcripts' genotypic blups were used for the rest of our analysis.

563 The genotyping data was obtained, first by mapping the RNAseq reads on the *P.*
564 *trichocarpa* genome reference (v3.0) (Goodstein et al., 2012). After the mapping, the
565 SNPs were called using 4 callers. In order to generate a high-confidence SNP set we
566 selected only the SNPs identified by at least 3 of the 4 callers and with less than 50% of
567 missing values. Remaining missing values were imputed using complementary
568 genotyping data obtained with a 12k Illumina Infinium Bead-Chip array (Faivre-Rampant
569 et al., 2016). Full details of SNP discovery, data filtering criteria and final selection are
570 given in Rogier et al. (2018). We then detected 874,923 SNPs. From these detected
571 SNPs 428,836 SNPs were retained for this study after filtering the SNPs with minimum
572 alleles frequencies lower than 0.05.

573 **eQTLs Analysis**

574 eQTLs analysis was performed using the Multi-Loci Mixed-Model (MLMM) approach
575 (Segura et al., 2012) and implemented in the R package MLMM v0.1.1. MLMM uses a
576 step-by-step forward inclusion and backward elimination approach under a mixed-model
577 framework which accounts for the confounding usually attributed to population structure
578 with a random polygenic effect. For each of the 34,229 transcripts we ran MLMM for up
579 to 10 steps and identified the optimal model according to the mBonf criterion (all
580 selected SNPs are significant at a 5% Bonferroni corrected threshold). The initial and
581 the optimal steps outputs have been saved for further analyses.

16

582 Based on the positional proximity of the genes, the eQTLS detected at each of these 2
583 steps were classified as CIS regulatory elements (non-coding DNA regulating the
584 transcription of neighboring genes), and/or as TRANS regulatory elements (regulating
585 the transcription of distant genes), according to the following rules:

586   - all eQTLs associated with the expression of a gene located in a different
587   chromosome are classified as TRANS, and the targeted gene is classified as a
588   TRANS regulated gene;
589   - all eQTLs located on the same locus as the gene it targets, according to the
590   genome annotation, are classified as CIS, and the targeted gene is also
591   classified as CIS regulated gene;
592   - the remaining eQTLs whose target gene is on the same chromosome but not on
593   the same locus, were splitted into CIS or TRANS according to their distance to
594   the middle of the gene they target. We estimated the maximum distance between
595   the CIS eQTLs identified at previous step and the middle of the gene they target
596   as 18.9 kb (eQTL being on the same position as its target gene). If the distance
597   between eQTLs and the gene they target is greater than 18.9 kb they were
598   classified as eQTLs TRANS and target Gene TRANS. Otherwise, the eQTLs and
599   the target gene were classified as CIS.

Models, prediction accuracy and cross-validation

601 Two ridge regression models were built for each trait with a single omic data as
602 predictor, genotypic data (**G** model), or transcriptomic data (**T** model), respectively with
603 p = 428,836 SNPs and q = 34,229 transcripts' expression levels variables. A third multi-
604 omic was also built with integration by concatenation of both omics data (**G+T** model).
605 These 3 models can be written as:

$$Y = \boldsymbol{X}\beta + \epsilon \quad \text{(eq 1)}$$

607 Respectively for models **G**,**T** and **G+T**, **X** represent the genotyping matrix $(n \times p)$, the
608 transcript expression level matrix for the genes $(n \times q)$ and the concatenated transcript
609 expression level and genotyping matrix $(n \times (p+q))$. With the same logic, β represent
610 the vector of effect sizes of variables of those matrices. Y is the vector of phenotype,
611 and ϵ the vector of residual errors of the model.

612 The models were computed using the R package glmnet (Friedman et al., 2010) in a 5
613 inner-fold and 10 outer-fold nested cross validation framework (Varma and Simon,
614 2006). The sampling process for the different folds was repeated 50 times. Each cross-
615 validation sample was used across all traits and for the three models. Paired t-tests in R
616 (rstatix package version 0.7.0) (Kassambara, 2021) were used for model comparisons
617 of performance.

618   The models performances was measured using R^2 between observed and predicted
619   values.

SNPs and transcript effects ranking

621   In order to study the changes operated for each feature (SNP or gene) when changing
622   prediction models from single omics to the concatenated counterpart, we compared the
623   change in ranking of the effects across models. Ranks were obtained for each
624   predicting model and trait from the ordering of squared effect sizes.

625   For each variant typologies, the estimated effects rank was compared between the
626   single omic models (**G** or **T**) and the multi-omics model with a paired wilcoxon test and a
627   Pearson correlation.

628   The difference in effect ranking between the model with concatenation and the single
629   omic models was calculated for the different typologies of each feature (SNPs and
630   genes). Then this ranking difference was averaged for each trait and regressed with the
631   concatenation advantage of each trait, which is the average accuracy difference
632   between the concatenation model and the single omics models:

$$\Delta_{predictor} = \frac{1}{\omega} \sum_{i=1}^{\omega} \left(R_i \atop (G+T) - R_i \atop (G\ or\ T) \right) \text{ (eq 2)}$$

634   Predictor represents either SNPs or transcripts. R is the ranking vector of squared effect
635   sizes of the given predictor. ω is the number of predictors (p for SNPs and q for
636   transcripts). $\Delta_{predictor}$ is the average difference in effect ranking between the multi-omics
637   model and the single omic one by trait for the given predictor.

GO analysis

639   Functional enrichment was conducted based on the gene ontology (GO) terms
640   associated with the best *Arabidopsis thaliana* homolog and based on Phytozome
641   v12.1.6 database (Goodstein et al., 2012). GO analysis was conducted using R
642   package topGO 2.44.0 (Alexa and Rahnenfuhrer, 2021) and Fisher's exact test with
643   'elim' used to correct for multiple comparisons. The significant threshold of GO terms
644   was $P \leq 0.05$.

Availability of data and code

646   Supporting data is available at: https://doi.org/10.15454/8DQXK5

647   Code for running the test and replicate the analysis are available at:

18

648   https://github.com/Tawfekh/Code-Article-Multi-omics-prediction

Acknowledgements

Author contributions

667   ARW: analyzed all the data and wrote the manuscript; HD: performed GO analyzes and
668   wrote the manuscript; LS and VS: planned and designed the project, supervised the
669   research and wrote the manuscript.

19

References

670

671 **Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L.** (2018). Genetics of trans-
672 regulatory variation in gene expression. eLife **7**: e35471.

673 **Alexa, A. and Rahnenfuhrer, J.** (2021). topGO: Enrichment Analysis for Gene Ontology
674 (Bioconductor version: Release (3.13)).

675 **Azodi, C.B., Pardo, J., VanBuren, R., Campos, G. de los, and Shiu, S.-H.** (2020).
676 Transcriptome-Based Prediction of Complex Traits in Maize. The Plant Cell **32**: 139–
677 151.

678 **Buil, A., Brown, A.A., Lappalainen, T., Viñuela, A., Davies, M.N., Zheng, H.-F., Richards,
679 J.B., Glass, D., Small, K.S., Durbin, R., Spector, T.D., and Dermitzakis, E.T.** (2015).
680 Gene-gene and gene-environment interactions detected by transcriptome sequence
681 analysis in twins. Nat Genet **47**: 88–91.

682 **de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., and Calus, M.P.L.**
683 (2013). Whole-Genome Regression and Prediction Methods Applied to Plant and Animal
684 Breeding. Genetics **193**: 327–345.

685 **Chaffey, N., Cholewa, E., Regan, S., and Sundberg, B.** (2002). Secondary xylem
686 development in Arabidopsis: a model for wood formation. Physiologia Plantarum **114**:
687 594–600.

688 **Chateigner, A., Lesage-Descauses, M.-C., Rogier, O., Jorge, V., Leplé, J.-C., Brunaud, V.,
689 Roux, C.P.-L., Soubigou-Taconnat, L., Martin-Magniette, M.-L., Sanchez, L., and
690 Segura, V.** (2020). Gene expression predictions and networks in natural populations
691 supports the omnigenic theory. BMC Genomics **21**: 416.

692 **Dillen, S.Y., Storme, V., Marron, N., Bastien, C., Neyrinck, S., Steenackers, M., Ceulemans,
693 R., and Boerjan, W.** (2009). Genomic regions involved in productivity of two interspecific
694 poplar families in Europe. 1. Stem height, circumference and volume. Tree Genetics &
695 Genomes **5**: 147–164.

696 **Ehsani, A., Sørensen, P., Pomp, D., Allan, M., and Janss, L.** (2012). Inferring genetic
697 architecture of complex traits using Bayesian integrative analysis of genome and
698 transcriptome data. BMC Genomics **13**: 456.

699 **Faivre-Rampant, P. et al.** (2016). New resources for genetic studies in Populus nigra: genome-
700 wide SNP discovery and development of a 12k Infinium array. Molecular Ecology
701 Resources **16**: 1023–1036.

702 **Farrar, D.E. and Glauber, R.R.** (1967). Multicollinearity in Regression Analysis: The Problem
703 Revisited. The Review of Economics and Statistics **49**: 92–107.

704 **Friedman, J., Hastie, T., and Tibshirani, R.** (2010). Regularization Paths for Generalized
705 Linear Models via Coordinate Descent. J Stat Softw **33**: 1–22.

706 **Gebreselassie, M.N. et al.** (2017). Near-infrared spectroscopy enables the genetic analysis of
707 chemical properties in a large set of wood samples from Populus nigra (L.) natural
708 populations. Industrial Crops and Products **107**: 159–171.

709 **Gianola, D.** (2021). Opinionated Views on Genome-Assisted Inference and Prediction During a
710 Pandemic. Frontiers in Plant Science **12**: 1533.

711 **Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks,
712 W., Hellsten, U., Putnam, N., and Rokhsar, D.S.** (2012). Phytozome: a comparative
713 platform for green plant genomics. Nucleic Acids Research **40**: D1178–D1186.

714 **Guet, J., Fabbrini, F., Fichot, R., Sabatti, M., Bastien, C., and Brignolas, F.** (2015). Genetic
715 variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in
716 European populations of black poplar (Populus nigra L.). Tree Physiology **35**: 850–863.

717 **Guo, Z., Magwire, M.M., Basten, C.J., Xu, Z., and Wang, D.** (2016). Evaluation of the utility of
718 gene expression and metabolic information for genomic prediction in maize. Theor Appl
719 Genet **129**: 2413–2427.

Idaghdour, Y. and Awadalla, P. (2013). Exploiting Gene Expression Variation to Capture Gene-Environment Interactions for Disease. Frontiers in Genetics **3**: 228.

Kassambara, A. (2021). rstatix: Pipe-Friendly Framework for Basic Statistical Tests.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology **15**: R29.

Legionnet, A., Muranty, Hé., and LefÈvre, F. (1999). Genetic variation of the riparian pioneer tree species Populus nigra. II. Variation in susceptibility to the foliar rust Melampsora larici-populina. Heredity **82**: 318–327.

Li, Z., Gao, N., Martini, J.W.R., and Simianer, H. (2019). Integrating Gene Expression Data Into Genomic Prediction. Front. Genet. **10**.

Liu, H., Luo, X., Niu, L., Xiao, Y., Chen, L., Liu, J., Wang, X., Jin, M., Li, W., Zhang, Q., and Yan, J. (2017). Distant eQTLs and Non-coding Sequences Play Critical Roles in Regulating Gene Expression and Quantitative Trait Variation in Maize. Molecular Plant **10**: 414–426.

Morgante, F., Huang, W., Sørensen, P., Maltecca, C., and Mackay, T.F.C. (2020). Leveraging Multiple Layers of Data To Predict Drosophila Complex Traits. G3 Genes|Genomes|Genetics **10**: 4599–4613.

Munoz, F. and Sanchez, L. (2017). breedR: statistical methods for forest genetic resources analysis. https://github.com/famuvie/breedR. R package version 0.12-2.

Newcombe, G., Stirling, B., and Bradshaw, H.D. (2001). Abundant Pathogenic Variation in the New Hybrid Rust Melampsora ×columbiana on Hybrid Poplar. Phytopathology **91**: 981–985.

Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., and Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics **16**: 85–97.

Robinson, M.D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology **11**: R25.

Rogier, O., Chateigner, A., Amanzougarene, S., Lesage-Descauses, M.-C., Balzergue, S., Brunaud, V., Caius, J., Soubigou-Taconnat, L., Jorge, V., and Segura, V. (2018). Accuracy of RNAseq based SNP discovery and genotyping in Populusnigra. BMC Genomics **19**: 909.

Schrag, T.A., Westhues, M., Schipprack, W., Seifert, F., Thiemann, A., Scholten, S., and Melchinger, A.E. (2018). Beyond Genomic Prediction: Combining Different Types of omics Data Can Improve Prediction of Hybrid Performance in Maize. Genetics **208**: 1373–1385.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet **44**: 825–830.

Steenackers, J., Steenackers, M., Steenackers, V., and Stevens, M. (1996). Poplar diseases, consequences on growth and wood quality. Biomass and Bioenergy **10**: 267–274.

Swanson-Wagner, R.A., DeCook, R., Jia, Y., Bancroft, T., Ji, T., Zhao, X., Nettleton, D., and Schnable, P.S. (2009). Paternal Dominance of Trans-eQTL Influences Gene Expression Patterns in Maize Hybrids. Science.

Takagi, Y., Matsuda, H., Taniguchi, Y., and Iwaisaki, H. (2014). Predicting the Phenotypic Values of Physiological Traits Using SNP Genotype and Gene Expression Data in Mice. PLOS ONE **9**: e115532.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics **7**: 91.

Varona, L., Legarra, A., Toro, M.A., and Vitezica, Z.G. (2018). Non-additive Effects in Genomic Selection. Frontiers in Genetics **9**: 78.

771   **Wang, J. and Kamp, B.J. van der** (1992). Resistance, tolerance, and yield of western black
772   cottonwood infected by Melampsora rust. Can. J. For. Res. **22**: 183–192.
773   **Ye, S., Li, J., and Zhang, Z.** (2020). Multi-omics-data-assisted genomic feature markers
774   preselection improves the accuracy of genomic prediction. J Animal Sci Biotechnol **11**:
775   109.
776   **Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C.** (2019). Machine and deep
777   learning meet genome-scale metabolic modeling. PLOS Computational Biology **15**:
778   e1007084.

Figure legends

Figure 1: Prediction accuracies
Violin plots of prediction accuracies for 21 traits in the poplar dataset according to three models, using only genotyping data (the G model colored in dark brown to the left in the panels), using only transcriptomic data (T model colored in dark blue), and concatenating both genotyping and transcriptomic data (G+T model colored in light brown to the right). Each 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 is increased (+) or decreased (-) in the multi-omics model by comparison with the single omic ones. Some traits were evaluated in two sites ("ORL" standing for Orléans in France and "SAV" for Savigliano in Italy). The white and black dots show the median and mean of the precision distributions, respectively. The dark brown and dark blue horizontal lines represent respectively the mean of precision distributions of G and T models.

Figure 2: eQTL map between SNPs and transcripts
Map of associations (dots) between SNPs and transcripts through an eQTLs analysis with multi locus detection (Step_op), with dot size reflecting the association score (-log10 of the p-value of the test). The darkened diagonal includes all CIS mediated associations, while the off-diagonal dots represent the TRANS associations.

Figure 3: Distribution of predictors' change in importance
Boxplot of the average change in importance of SNPs (panels A) and transcripts (panels B). Each dot represents the average difference per trait and per site of the predictor ranks between the multi omics model (G+T) and the single omic models (G for SNPs and T for transcripts). The red and blue boxplots show respectively the distribution of the average rank change for the TRANS-eQTLs and CIS-eQTLs. The boxplot in black shows the distribution for the predictors that have not been detected in the eQTL analysis.

Figure 4: Relationship between predictors' change in importance and muti-omics prediction advantage
Regression across traits measured at Orleans between average change in importance of predictors and advantage in performance of G+T over the single-omic counterpart. The top panel (A) shows the regression obtained with the eQTLs (TRANS-eQTLs on the left and CIS-eQTLs on the right). The bottom panel (B) shows the regression obtained with the regulated genes (TRANS on the left and CIS on the right).

Figure 5: Gene ontology terms enrichment analysis
Schematic representation of the enriched GO terms among the top targeted transcripts or eQTL gene models list for A) the circumference of the tree trunk or B) the lignin content evaluated at Orleans. Font size and color intensity are proportional to -log10(p) of the top 10 GO terms.

23

822 **Supplemental Data**

823 **Supplemental Figure S1:** eQTL map between SNPs and transcripts (Step_0)
824 **Supplemental Figure S2:** Abundance and score of CIS and TRANS eQTLs
825 **Supplemental Figure S3:** Comparison between the importance of predictors across
826 single and multi-omics models
827 **Supplemental Figure S4:** Variation of the change in importance of the eQTLs and
828 targeted transcripts
829 **Supplemental Figure S5:** Change in importance of the eQTLs and their corresponding
830 targeted transcripts
831 **Supplemental Figure S6:** Relationship between predictors change in importance and
832 muti-omics prediction advantage for traits measured at Savigliano
833 **Supplemental Figure S7:** Average number of connection for the eQTLs and targeted
834 transcripts
835 **Supplemental Table S1:** Prediction accuracies comparison between the multi-omics
836 model and the single omic ones
837 **Supplemental Table S2:** Complete gene ontology analysis of all traits
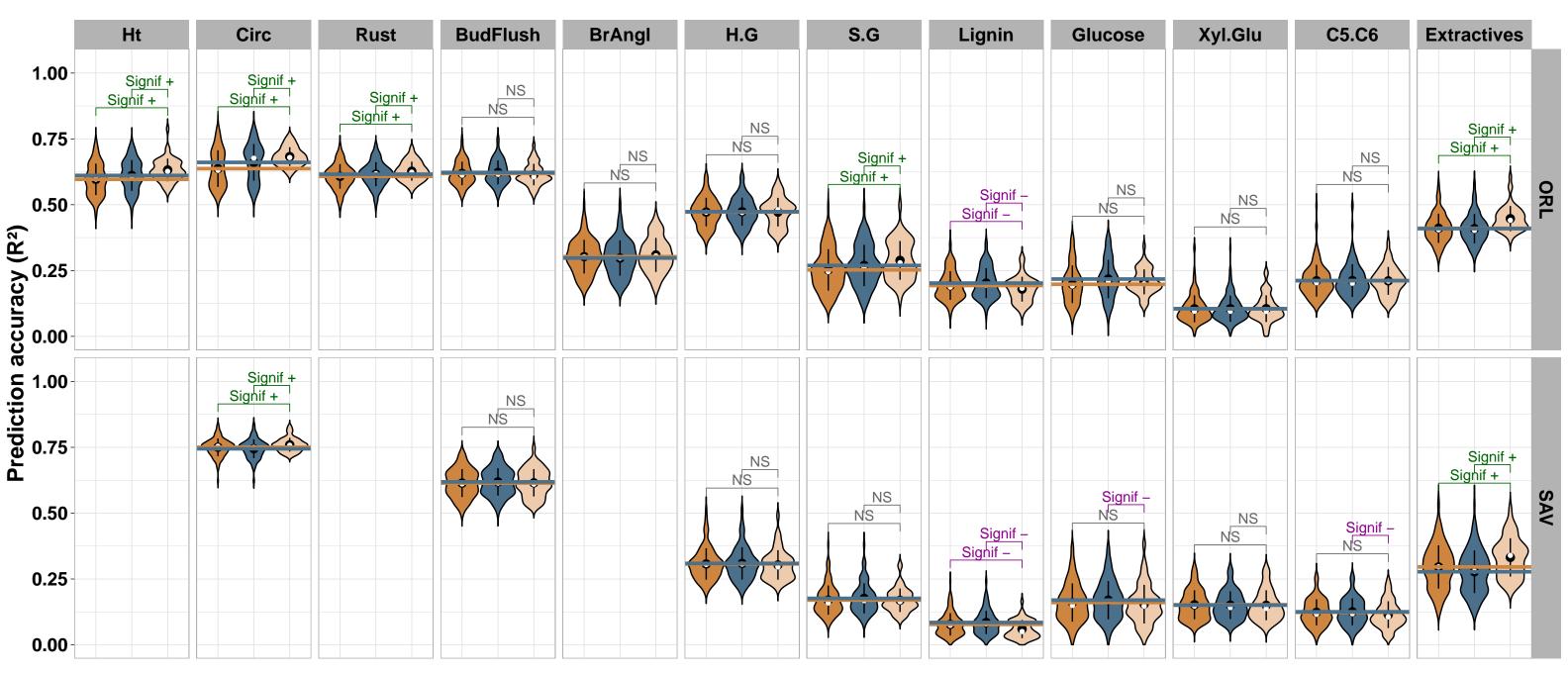
838 **Tables**

839 **Table 1: List of phenotypic traits**

840 List of phenotypic traits used in the study with their abbreviations, classified by
841 functional types, with site of measurement and year.

Functional types	Trait	Abbreviation	Site	Year
Growth	Height	Ht	ORL	2011
	Circumference	Circ	ORL	2011
			SAV	2009
Pathogen Tolerance	Tolerance to rust	Rust	ORL	2009
Phenology	Date of bud flush	BudFlush	ORL	2009
			SAV	2011
Architecture	Branching angle	BrAngl	ORL	2009
Biochemical	H/G lignin ratio	H.G	ORL	2011
			SAV	2009
	S/G lignin ratio	S.G	ORL	2011
			SAV	2009
	Lignin content	Lignin	ORL	2011
			SAV	2009
	Glucose content	Glucose	ORL	2011
			SAV	2009
	Xylose to glucose ratio	Xyl.Glu	ORL	2011
			SAV	2009
	5C/6C carbon sugar ratio	C5.C6	ORL	2011
			SAV	2009
	Extractives content	Extractives	ORL	2011
			SAV	2009

842

25

A) SNPs

B) Transcripts

A) eQTLs

TRANS: $R = -0.81$, $p = 0.0015$

CIS: $R = -0.6$, $p = 0.037$

y-axis: mean.R^2(G+T) − mean.R^2(G)

B) Targeted transcripts

TRANS: $R = 0.3$, $p = 0.34$

CIS: $R = -0.64$, $p = 0.024$

y-axis: mean.R^2(G+T) − mean.R^2(T)

x-axis: Average change in importance

A)

Top middle:
endosperm development
anatomical structure formation
formation of anatomical bounda
formation of plant organ bound
nitrogen compound transport
positive regulation of macromo
regulation of cell population
cellularization
transcription-coupled nucleoti
modification-dependent macromo

Left middle:
fatty acid homeostasis
triglyceride homeostasis
phospholipid homeostasis
regulation of nucleobase-conta
positive regulation of nucleob
nucleobase-containing compound
regulation of RNA export from
regulation of meristem structu
positive regulation of RNA exp
acylglycerol homeostasis

Right middle:
xylan catabolic process
cell wall polysaccharide metabolic proce...
xylan metabolic process
cell wall polysaccharide catabolic proce...
ergosterol metabolic process
cysteinyl-tRNA aminoacylation
cellular macromolecule catabolic process
macromolecule catabolic process
hemicellulose metabolic process
cell wall macromolecule catabolic proces...

Bottom middle:
response to organonitrogen com
porphyrin-containing compound
protein targeting to chloropla
protein import into chloroplas
establishment of protein local
protein targeting
tetrapyrrole metabolic process
protein localization to organe
protein localization to chloro
protein modification by small

99%

1%

Targeted transcripts

1%   99%

eQTL gene models

B)

Top middle:
gene expression
nucleobase-containing compound
purine nucleotide-sugar transm
nucleic acid metabolic process
spliceosomal complex assembly
RNA metabolic process
specification of symmetry
organic cyclic compound metabo
regulation of gene expression
GDP-fucose transmembrane trans

Left middle:
vesicle-mediated transport
glycerolipid biosynthetic proc
regulation of glutamine family
macromolecule localization
peptide transport
protein localization
adenylate cyclase-modulating G
amide transport
establishment of protein local
protein transport

Right middle:
retrograde vesicle-mediated tr
response to organonitrogen com
indole-containing compound cat
lipid oxidation
trehalose metabolism in respon
cellular response to organonit
fatty acid oxidation
fatty acid catabolic process
cellular lipid catabolic proce
mRNA splice site selection

Bottom middle:
Golgi vesicle transport
ribonucleoprotein complex asse
protein localization to extrac
dolichyl monophosphate biosynt
organophosphate ester transpor
ribonucleoprotein complex subu
root radial pattern formation
cellular response to ethylene
ethylene-activated signaling p
phosphorelay signal transducti

99%

1%

Targeted transcripts

1%   99%

eQTL gene models