bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The relationship between gene co-expression network connectivity and phenotypic prediction sheds light at the core of the omnigenic theory

Aurélien Chateigner[1], Marie-Claude Lesage-Descauses[1], Odile Rogier[1], Véronique Jorge[1], Jean-Charles Leplé[2], Véronique Brunaud[3,4], Christine Paysant-Le Roux[3,4], Ludivine Soubigou-Taconnat[3,4], Marie-Laure Martin-Magniette[3,4,5], Leopoldo Sanchez[*,1], and Vincent Segura[*,†,1]

[1]BioForA, INRA, ONF, Orléans, France
[2]BIOGECO, INRA, Univ. Bordeaux, Cestas, France
[3]Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA, Université Paris-Sud, Université d'Evry, Université Paris-Saclay, Gif sur Yvette, France
[4]Institute of Plant Sciences Paris-Saclay (IPS2), CNRS, INRA Université Paris-Diderot, Sorbonne Paris-Cité, Gif sur Yvette, France
[5]MIA-Paris, AgroParisTech, INRA, Paris, France

Abstract

Recent literature on the differential role of genes within networks, including the omnigenic model, distinguishes core from peripheral genes in the layout underlying phenotypes. Cores are typically few, each of them highly contributes to phenotypic variation, but because they are so few, they altogether only explain a small part of trait heritability. In contrast, peripherals, each of small influence, are so numerous that they finally lead phenotypic variation. We collected and sequenced RNA from 459 European black poplars and built co-expression networks to define core and peripheral genes as the most and least connected ones. We computed the role of each of these gene sets in the prediction of phenotypes and showed that cores contribute additively to phenotypes, consistent with a downstream position in a biological cascade, while peripherals interact to influence phenotypes, consistent with an upstream position. Quantitative and population genetics analyses further revealed that cores are more expressed than peripherals but they tend to vary less and to be more differentiated between populations suggesting that they are more constrained by natural selection. Our work is the first attempt to integrate core and peripheral terminologies from co-expression networks and omnigenic theory. In the end, we showed, that there seems to be a strong overlap between them, with core genes from co-expression networks likely being a mixture of integrative hubs with a direct effect on phenotype in agreement with the omnigenic theory, and master regulators which control the overall metabolic flow towards the phenotype.

Keywords: core, peripheral, boruta, machine learning, *Populus nigra*

[*]Equal contribution
[†]Corresponding author: `vincent.segura@inra.fr`

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Gene-to-gene interaction is a pervasive although elusive phenomenon underlying phenotype expression. Genes operate within networks with more or less mediated actions on the phenome. Systems biology approaches are required to grasp the functional topology of these networks and ultimately gain insights into how gene interactions interplay at different biological levels to produce global phenotypes (Mackay et al., 2009). New sources of information and their subsequent use in the inference of gene networks are populating the wide gap existing between phenotypes and DNA sequences and, therefore, opening the door to systems biology approaches for the development of context-dependent phenotypic predictions. RNA sequencing (RNAseq) is one of such new sources of information that can be used to infer gene networks (Han et al., 2015).

Among the many works on gene network inference based on transcriptomic data, we would like to pinpoint here two recent studies that aim at characterizing the different gene roles within co-expression networks (Josephs et al., 2017; Mähler et al., 2017). Josephs et al. (2017) studied the link between previously published concepts related to gene expression (Josephs et al., 2015), gene connectivity (Langfelder and Horvath, 2008), divergence (Williamson et al., 2005) and traces of natural selection (Josephs et al., 2015; Sicard et al., 2015) in a natural population of the plant *Capsella grandiflora*. They showed that both connectivity and local regulatory variation on the genome are important factors, while not being able to disentangle which of them is directly responsible for patterns of selection among genes. Mähler et al. (2017) recalled the importance of studying the general features of biological networks in natural populations. With a genome-wide association study (GWAS) on expression data from RNAseq, they suggested that purifying selection is the main mechanism maintaining functional connectivity of core genes in a network and that this connectivity is inversely related to eQTLs effect size. These two studies start to outline the first elements of a gene network theory based on connectivity, stating that core genes, which are highly connected, are each of high importance, and thus highly constrained by selection. In contrast to these central genes, there are peripheral, less connected genes, never far from a core hub. These peripheral genes are less constrained than core genes and consequently, they harbor larger amounts of variation at population levels.

In another recent study, Boyle et al. (2017) proposed the omnigenic theory, as an extension of the classic polygenic view for the genetic architecture of complex traits. They provide a clear but (disease) trait-centered definition of their new paradigm, explaining that numerous genes that are peripheral in a regulatory network are sufficiently connected to genes directly involved in a disease to modulate their effect and explain most of the missing heritability of the disease risk (Maher, 2008). Unlike the two precedent studies (Josephs et al., 2017; Mähler et al., 2017), which were based on co-expression networks and thus centered around connectivity for categorizing genes, this new study focuses instead on the relationship between genes and traits. Core and peripheral genes in the omnigenic theory are thus defined with respect to their proximity to the trait they affect (Liu et al., 2018). This point somehow recalls classic studies of molecular evolution in biological pathways which showed that selection pressure is correlated to the gene position within the pathway, either positively (Han et al., 2013; Lu, 2003; Rausher et al., 2008, 1999; Riley et al., 2003; Yu et al., 2011) or negatively (Han et al., 2013; Jovelin and Phillips, 2011; Song et al., 2012; Wu et al., 2010), depending on the pathway. Jovelin and Phillips (2011) showed that selective constraints are positively correlated to expression level, confirming previous studies (Drummond et al., 2005; Duret and Mouchiroud, 2000; Pál et al., 2001). Montanucci et al. (2011) showed a positive correlation between selective constraints and connectivity, although such a possibility remained contentious in previous works (Bloom and Adami, 2004; Fraser and Hirsh, 2004).

While Josephs' (Josephs et al., 2017) and Mahler's (Mähler et al., 2017) studies apparently framed the general view behind Boyle's theory (Boyle et al., 2017), based on topological features described in molecular evolution studies of biological pathways, a point remains quite unclear so far: to what extent core and peripheral genes based on connectivity within a co-expression network overlap with core and peripheral genes as defined with respect to a given trait such as in the omnigenic theory? One way to clarify this would be to study the respective roles of core and peripheral genes, as defined on the basis of their connectivity within a co-expression network, in the prediction of a phenotype. Even if predictions are still one step before validation by in vivo experiments, they already represent a landmark that may not only be correlative but also closer to causation, depending on the modeling strategy.

Our present study aims at exploring gene ability to predict traits, with datasets representing core genes and peripheral genes. By making use of two methods to predict these phenotypes, a classic additive

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

linear model, and a more complex and interactive neural network model, we further aimed at studying the mode of action of each type of genes. On the one hand, genes that are better predictors with an additive model are supposed to have overall a more additive, direct mode of action representing a gene that would be downstream in a biological pathway. We expect core genes to display such additive behavior, with a high but selectively constrained expression level (Jovelin and Phillips, 2011; Montanucci et al., 2011). On the other hand, genes being better predictors with an interactive model are supposed to be upstream in pathways. We expect peripheral genes to behave interactively, with a lower but relatively more variable expression level. With a lower variation, we also expect core genes to be worse predictors for traits than peripheral genes unless the former also bear larger effects.

To answer the questions concerning the respective roles of core and peripheral genes on phenotypic variation and the way these roles fit into the new omnigenic theory, we have sequenced the RNA of 459 samples of black poplar (*Populus nigra*), corresponding to 241 genotypes, from 11 populations representing the natural distribution of the species across Western Europe. We also have for each of these trees phenotypic records for 17 traits, covering growth, phenology, physical and chemical properties of wood. They cover two different environments where the trees were grown in common gardens, in central France and northern Italy. By predicting these traits from our gene expression data, from different gene sets, selected based on their topology in co-expression networks, we uncovered the importance of genes of varying centrality in order to characterize them and test whether this network centered definition of gene sets matches with the trait centered definition proposed in the omnigenic theory.

Results

Wood samples, phenotypes, and transcriptomes

Wood collection and phenotypic data (**Table S1**) have been previously described (Gebreselassie et al., 2017). Further details are provided in the methods section. Briefly, we are focusing on 241 genotypes planted in 2 common gardens, in Orléans (central France) and Savigliano (northern Italy), and for which phenotypic data have been collected. In Orléans, we used 2 clonal trees per genotype to sample xylem and cambium during the 2015 growing season for RNA sequencing. Because of sampling and experimental mistakes that were further revealed by the polymorphisms in the RNA sequences, we ended up with 459 samples for which the genotype identity was confirmed. These samples correspond to 218 genotypes with two biological replicates and 23 genotypes with a single biological replicate. We mapped the sequencing reads on the *Populus trichocarpa* transcriptome (v3.0) to obtain gene expression data.

Sample collection extended on a 2 weeks period, with varying weather along the days, and different operators involved. We did PCA analyses on the cofactors that were presumably involved in the experience, to look whether any confounding effect could be identified (**Figure S1**). No clear segregation was found for any of these, except for the ones associated with weather. To verify this observation, we used mixed-models to correct effects of all these cofactors, with the breedR R package (Muñoz and Sanchez, 2017), and while it properly corrected the environmental effects, it also removed information from the data, making prediction quality much poorer than without cofactor correction for most of the traits (**Figure S2**). Since phenotype is a mixture between genotype and environment, we supposed that correcting the environment also removed part of the genetic variation. Further analyses with complex neural network models, expected to account more efficiently for interactions with hidden theoretical states, did not show better results than additive models. We thus did not favor one particular type of model with uncorrected data. Moreover, we did not aim at interpreting the effect of each variable in this study but rather at inferring mechanisms from the prediction quality of the different models, which might be less prone to confounding effects.

From the 41,335 transcripts obtained from the mapping, we removed the 1,653 without reads, we normalized the read counts, stabilized their variance and transformed the counts of the 39,682 remaining transcripts to counts per million. Further details are provided in the methods section. Hereafter, we refer to this set of 39,682 transcripts as the full gene set.

Clustering and network construction

The classical approach to build a signed scale-free gene expression network is to use the weighted correlation network analysis, implemented in the WGCNA R package (Langfelder and Horvath, 2008), using a power function on correlations between gene expressions. We chose to use Spearman's rank correlation to avoid any assumption on the linearity of relationships. The scale-free topology fitting index (R^2) reached a maximum of 0.85 for a soft threshold of 15, yielding a

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

mean connectivity of 22.9 (**Figure 1A**). We detected 25 gene expression modules (**Table S2**) with automatic detection (merging threshold: 0.25, minimum module size: 30, **Figure 1B**). Spearman's correlations between 17 traits values and expression, presented in the lower panel of **Figure 1B** below the module membership of each gene, displayed a structuration when ordered following the gene expression tree. The traits themselves were line ordered according to clustering on their scaled values to represent their relationships (**Figure S3**). Interestingly, some patterns in the correlation between expression and traits did not follow what we would expect from the similarity between traits (3 traits out of 7 with data in both geographical sites). For instance, in the group composed of S.G ratios and glucose composition, the patterns were more similar between sites across traits than between traits across sites (**Figure 1B**, **Figure S3**). Complex shared regulations mediated by the environment seem to be in control of these phenotypes, suggesting site-specific genetic control. Otherwise, glucose composition in Savigliano, wood basic density, and extractives in Orléans presented similar patterns, contrarily to what would be expected from the correlations between these traits. These results suggest that the comparative analysis of correlations between gene expression and between traits allow unraveling underlying links between traits that are not obvious from factual phenotypic and genetic correlations between traits.

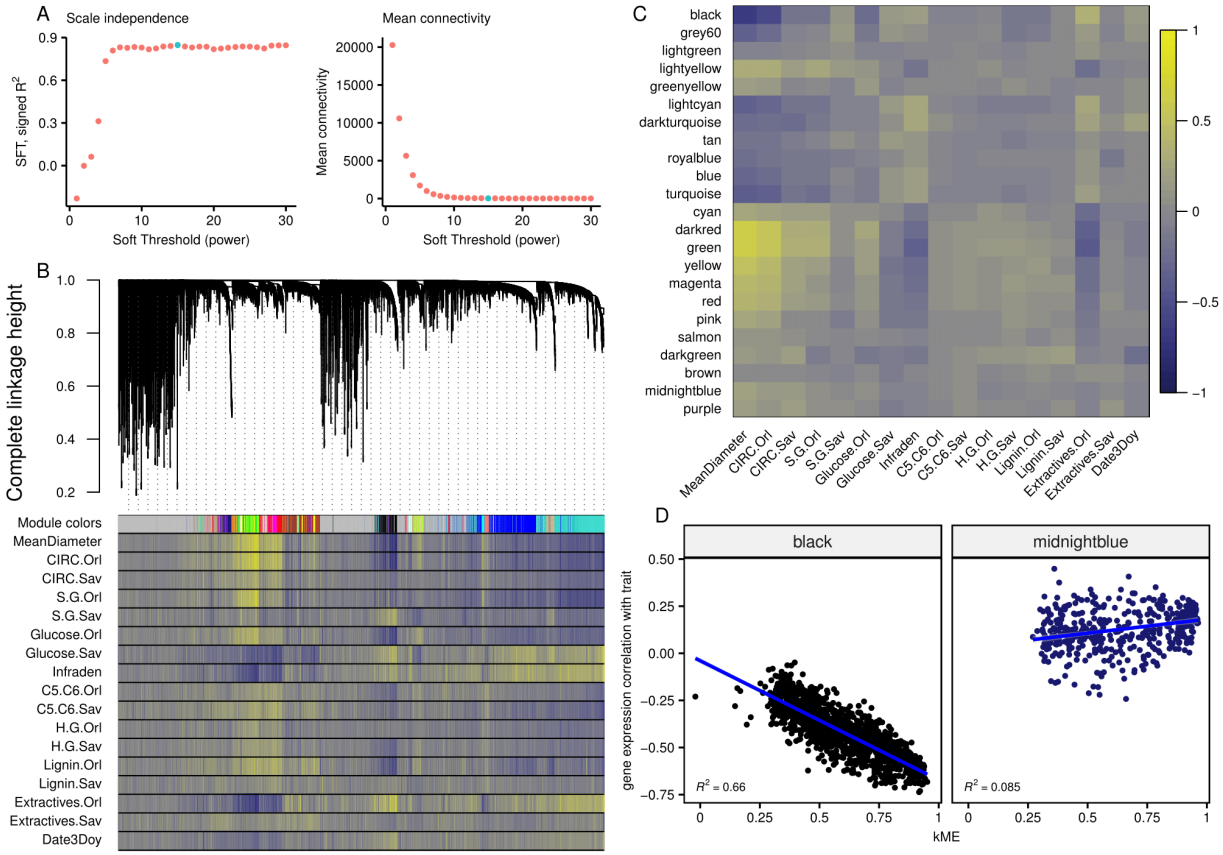


Figure 1: WGCNA analysis of gene expression data. (A) Selection of the soft threshold (green dot) based on the correlation maximization with scale-free topology (left panel) producing low mean connectivity (right panel). (B) Gene expression hierarchical clustering dendrogram, based on the Spearman correlations (top panel), resulting in clusters identified by colors (first line of the bottom panel). Spearman correlations between gene expressions and traits values are represented as color bands on the other lines of the bottom panel, from highly negative correlations (dark blue) to highly positive correlations (light yellow), according to the scale displayed in panel C. (C) Spearman correlation between eigengenes (the best theoretical representative of a gene expression module) of modules identified in the previous panel and traits, again on a dark blue (highly negative) to light yellow (highly positive) scale. (D) Focus on two modules from the previous graph, representing the correlations between gene expression correlation with mean sample diameter and centrality in the module. These two panels represent the strongest (left panel, black module, $R^2 = 0.66$) and the weakest (right panel, midnightblue module, $R^2 = 0.09$) correlations with the corresponding trait.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To get further insight into the relationships between module composition and traits, we looked at the strongest correlations between the best theoretical representative of a gene expression module (eigengene) and each trait, in order to identify genes in relevant modules with an influence on the trait (**Figure 1C**). Following a Bonferroni correction of the p-values provided by WGCNA, only 72 correlations remained significant ($p \leq 0.05$) out of the initial 425 traits by modules combinations, and 5 modules were defined as composed of genes not involved in any of the traits studied (salmon, greenyellow, brown, lightgreen and darkgrey, **Figure S4**). In significantly correlated modules, gene expression correlation with trait was also significantly correlated with centrality in the module (represented by the kME, the correlation with the module eigengene), while no correlation was found in poorly correlated modules (**Figure 1D**, **Figure S5**). In other words, there is a three-way correlation. The genes with the highest kME in a given module are the most correlated to the eigengene and, consequently, are also the most correlated to those traits with the largest correlation with the module eigengene. Although this is somehow expected, it underlines the usefulness of kME as a centrality score to further characterize the genes within each module. We thus used this centrality score to define further the topological position of our gene expressions in networks. As a gene has a score for each module, we used the gene's highest absolute score, which is the score in the module to which it was assigned. In order to avoid bias in gene selection by large groups, we selected the 10% of genes with the highest global absolute scores to define the core genes group, and the 10% with the lowest global absolute scores to define the peripheral genes group. Finally, we selected 100 samples of 3,968 (10%) random genes as control groups (**Figure S6**).

One particular module from the WGCNA clustering is the grey module. This module typically gathers genes with poor membership to any other module. In our case, it is the largest module, with 15,214 genes (38% of the full set). It gathers the vast majority of genes with very low kME (**Figure S6**) and 99% of the peripheral genes set (**Table S4**). While it is typically discarded in classic clustering studies, its eigengene displays the highest number of significant correlations with traits suggesting global non-negligible biological roles (**Figure 1C**, **Figure S4**). It could have been interesting to use it as a contrasting set for the remaining of the study in light of the omnigenic theory. However, its size is not suitable for fair comparison to the 10% core genes. We thus decided to stick to the peripheral genes set as previously defined to contrast the core genes set.

To assess the robustness of WGCNA analysis results, we compared it to a k-means clustering (R package coseq (Rau and Maugis-Rabusseau, 2017)) of the gene expressions (**Figure S7A**). The distribution of WGCNA and k-means' clusters showed a correlation of 0.42 (**Figure S7B**). K-means clustering tends to force groups of comparable size (Biernacki et al., 2006), which does not seem biologically credible. Furthermore, the correlations between the k-mean modules eigengenes and traits were lower than with WGCNA's, with a poor repartition of the different modules on the first 2 principal component analysis space (**Figure S7C**). We thus preferred WGCNA clustering to k-means clustering for this analysis and were quite confident about its robustness given its overall concordance with k-means clustering.

Boruta gene expression selection

In addition to the previously defined gene sets (full, core, random, peripheral), we wanted to have a set of genes being relevant for their predictability of the phenotype. Our hypothesis here would be that this set is the one that enables the best prediction of a given trait with a limited gene number that would be comparable to the other subsets of genes selected from their centrality within the networks. For that purpose, we performed a Boruta analysis (Boruta R package (Kursa and Rudnicki, 2010)) on the full genes set within the training sample (60% of all observations). This algorithm performs several random forests to analyze which gene expression profile is important to predict a phenotype. In the end, a pool of 637 unique gene expressions was found to be important to predict our phenotypes (**Figure S8**). Traits with the highest number of important genes were related to growth. For the other traits, we always had more genes selected when the trait was measured in Orléans compared to Savigliano (respective medians of 29 and 17.5). We hypothesize that this was due to the fact that RNA collection was performed on trees in Orléans. One exception to this pattern was the Lignin content, that we suspected to be due to a methodological difference between assessments, as previously discussed (Gebreselassie et al., 2017). On average, genes that were specific to single traits represented 62% of selected genes, genes shared across sites for a given trait were 4%, genes shared by trait category (growth, phenology, physical, chemical) were 18%, and genes shared among all traits were 16%.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Phenotype prediction with gene expression

For our 5 genes sets (full, core, random, peripheral and Boruta), we trained two classes of models to predict the phenotypes: an additive linear model (ridge regression) and an interactive neural networks model. For the former, we used ridge regression to deal with the fact that for all gene sets the number of predictors was larger than the number of observations. For the latter, we chose neural networks as a contemporary machine-learning method, which is not subjected to dimensionality problems (González-Recio et al., 2014) and is able to capture interactions without *a priori* explicit declaration between the entries, here gene expressions. In theory, both methods are able to capture the same signal but differences between their results, due to computing efficiency by design, let us capture more efficiently additivity or interactivity and are thus likely to inform us about the preferential mode of action of each gene set. **Figure 2** and **Figure S9** show that for linear modeling with ridge regression, the best genes set to predict phenotypes was the core gene set, followed by the full, Boruta, random and peripheral sets (respective median prediction R^2 over all traits of 0.33, 0.31, 0.25, 0.18 and 0.16). On the contrary, for neural network modeling, core genes constituted the worst set by far, followed by a cluster of similarly performing peripheral, random and Boruta sets (respective median prediction R^2 over all traits of 0.07, 0.21, 0.22, 0.22). We have not been able to compute neural network models with the full set as the number of predictors remains too large to be fitted within a reasonable time on computing clusters. Across phenotypes (**Figure S9**), predictions were generally less variable under neural network than under the ridge regression counterpart (interquartile range mean division by 1.48).

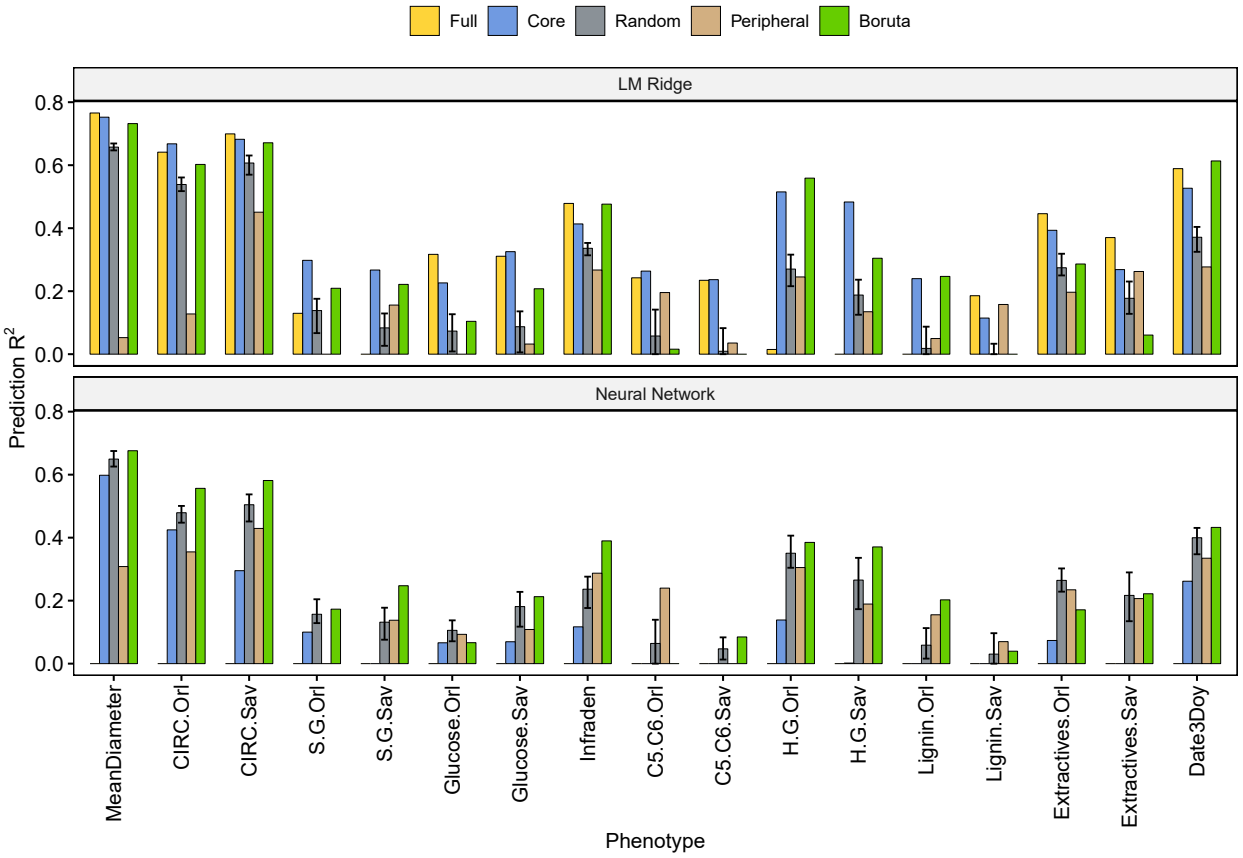


Figure 2: Predictions scores on test sets (R^2 on the y axis) for 2 algorithms (LM Ridge, top panel; neural network, bottom panel) for each phenotypic trait (on the x axis). The color of each bar represents the gene set that has been used for the prediction. Intervals for the random set represent the first and third quartiles of the distribution of the 100 different realizations, while the height of the bar corresponds to the median.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To further investigate the behavior of genes with different positions in networks with respect to the prediction model used, we computed differences between prediction scores for core and peripheral gene sets for additive (ridge regression) and interactive (neural network) algorithms (**Figure 3**). As a null reference for inference, a randomization strategy involving 100 random sets of genes was used to infer differences in prediction scores between models due to random sampling. For this, we computed a total of $4,950$ differences corresponding to all pairwise differences, excluding reciprocals and self-comparisons. A positive difference indicates an advantage of core genes sets over peripherals and, conversely, a negative difference indicates an advantage of peripheral genes. Except for 4 out of 17 cases, most traits showed a contrasting behavior of the two alternative algorithms. While most additive ridge regression models yielded positive scores across traits, the neural network counterpart showed negative scores. This hints at the fact of different gene actions in the two sets of genes. Indeed, the former ridge regression models capture mostly additive gene actions, which appeared to be prominent for core genes. Contrarily, neural network modeling is better suited for revealing gene interactions, which seem to be inherently associated with peripheral gene functioning. On average, the neural network had a mean difference of -0.08, showing that they were mainly in favor of the peripheral genes set. On the opposite, ridge regression models had mean differences of 0.24, showing that they were predicting a lot better with core genes set. It should be noted that concording behavior might come from the almost complete inability to predict the phenotype for a particular trait (a score close to 0 in **Figure 2**). In most of the cases, the contrasting pattern between core and peripherals with the two algorithms could not be drawn exclusively by chance as indicated by the distribution of randomized sets which clearly appeared to be centered on zero (mean differences of -0.002 and 0.0002 for neural network and ridge regression models respectively).

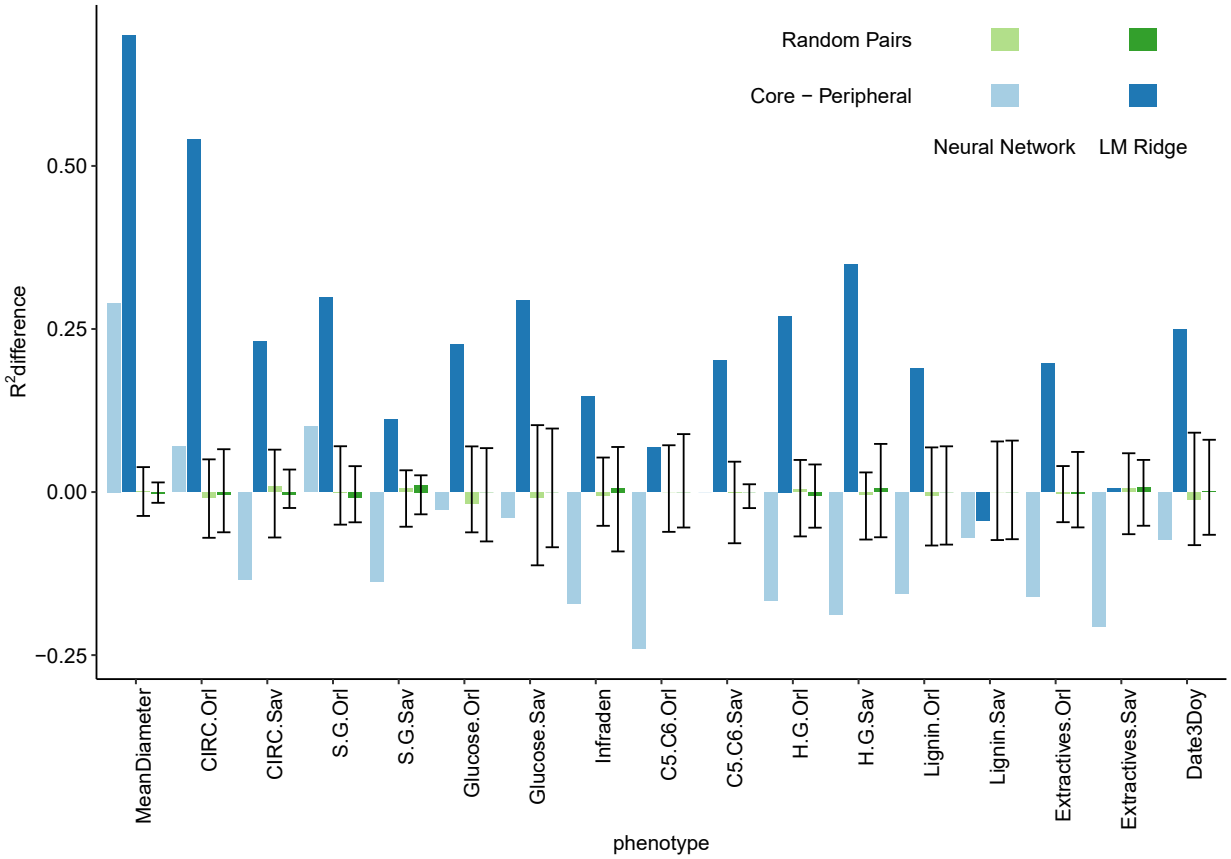


Figure 3: Difference of prediction scores (on the y-axis) between the core and peripheral gene sets (in blue) or between random sets (in green), for additive (LM Ridge in saturated colors) and interactive (neural network in faded colors) algorithms, for the different traits (on the x-axis). For the random pairs, error bars represent the first and third quartiles of the differences between pairs of randomized sets and the bar corresponds to the median.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Heritability and population differentiation of modules

To get further insights into the biological role of core and peripheral genes at population levels, we looked at the distribution of various characteristics between gene sets: gene expression level (**Figure 4B**); several classical population statistics, including heritability (h^2, **Figure 4A**), coefficient of quantitative genetic differentiation (Q_{ST}, **Figure 4C**), coefficient of genetic variation (CVg, **Figure 4E**), gene diversity (Ht, **Figure 4F**); and a contemporaneous equivalent to F_{ST} for genome scans ($ScorePCadapt$ (Luu et al., 2017), **Figure 4D**). Gene expression level, h^2, Q_{ST} and CVg were computed from gene expression data, while Ht, and $ScorePCadapt$ were computed from polymorphism data (SNP) and averaged per gene model, for more details see the methods section.

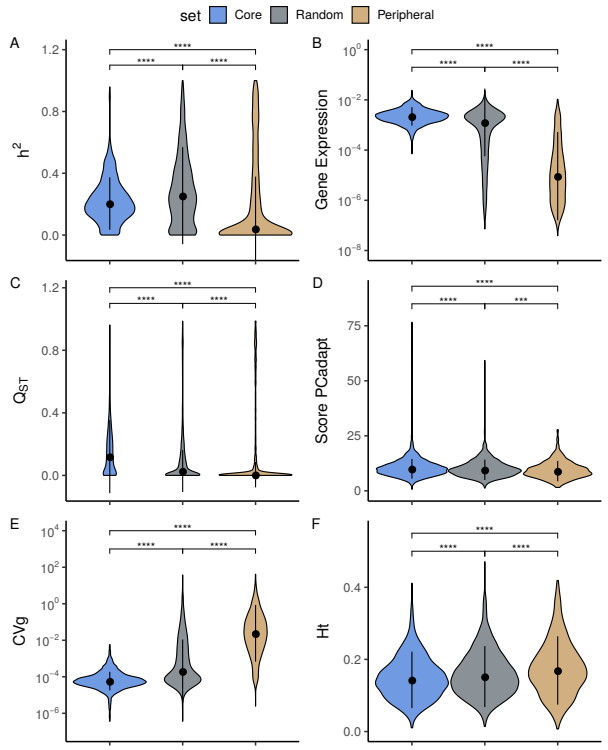


Figure 4: Distribution of various characteristics between core, peripheral and random gene sets. (A) Heritability h^2, (B) gene expression (in counts per million), (C) differentiation Q_{ST}, (D) $ScorePCadapt$, (E) genetic variation coefficient CVg and (F) overall gene diversity Ht violin plots with median (black dot) and interquartile range (black line) for each of the core (in blue), random (in grey) and peripheral (in brown) gene sets. ns: $p > 0.05$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$, ****: $p \leq 0.0001$ by Wilcoxon rank-sum test.

The extent of heritability for gene expression was higher for the random set than for core and peripheral genes sets, the latter having extremely low median heritability (0.04) (**Figure 4A**). Gene expression level (**Figure 4B**) and the extent of population differentiation from expression data (**Figure 4C**) tended to be higher in core set than in the other sets, with intermediate levels in the random set and the lowest levels in the peripheral set. According to the $ScorePCadapt$ (**Figure 4D**), core genes showed more evidence of population-specific selection patterns than the other two sets, with random genes having intermediate levels. Concerning the coefficient of genetic variation (**Figure 4E**), there was a clear difference between sets, with core genes displaying a very low variation, peripheral genes a very high, and random genes intermediate levels. Finally, there was a small difference in overall gene diversity (**Figure 4F**) that confirms the differences observed for CVg computed on gene expression level, peripheral genes being more diverse than random, and random more diverse than core genes.

Altogether, these statistics showed clear differences between core and peripheral genes: core genes are highly expressed (**Figure 4B**), highly differentiated between populations (**Figure 4C**), with generally low levels of genetic variation (**Figure 4D, 4E, 4F**); while peripheral genes are poorly expressed, poorly differentiated between populations, with generally higher genetic variation.

Discussion

Characterizing the way genes contribute to phenotypic variation could prove highly valuable to better understand the genetic architecture of complex traits. With the advent of omics data, a huge amount of information is nowadays becoming available to fill the gap between variations at the DNA and phenotype levels. In this context, two recent works used RNAseq in natural populations to build co-expression networks and study the relationship between network topology and patterns of natural selection (Josephs et al., 2017; Mähler et al., 2017). While they found differences in natural selection among genes given their connectivity within networks, they did not investigate how these differences affect phenotypic variation. With respect to the genetic architecture of complex traits, another team used the small world property of gene networks to develop a new theory, coined omnigenic, to explain the patterns observed in the GWAS results from human genetics studies (Boyle et al., 2017). This theory categorizes genes into core and peripheral genes according to the way they affect

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

a given phenotype (Liu et al., 2018). More precisely, the omnigenic theory states that core genes are typically few, they have a strong direct effect (*i.e.* not mediated through gene regulatory networks) on the phenotype but, because they are so few, they altogether do not contribute so much to the phenotypic variation. In contrast, peripheral genes are numerous, they have individually a tiny effect on the phenotype but, because they are so many, they contribute to a substantial proportion of the trait heritability. In the present study, we aimed at studying the relationship between gene connectivity in co-expression networks and phenotypic prediction. By these two features, we further aimed at testing how a network-based definition of core and peripheral genes relates to the trait-centered definition of core and peripheral genes from the omnigenic theory. We defined core and peripheral genes as the 10% most central and most peripheral genes respectively according to the outputs of WGCNA analysis. We are aware that this is somehow an oversimplification, an extreme contrast of an otherwise continuous phenomenon. Moreover, as stated in the omnigenic theory, core genes are only a modest number and peripheral genes are the remaining majority of expressed genes. While the choice of the arbitrary threshold of 10% is based on the Mahler's definition of core genes (Mähler et al., 2017), the fact of equaling both samples responded to the need for statistical comparativeness between samples of equal size. Moreover, such contrasting samples represented two conspicuous features of the distribution of kME (**Figure S6**), with a thick tail of well-connected genes and a high mass of poorly connected genes.

On average, core genes were the ones predicting the most efficiently a phenotype, for any trait category, with an additive model, even if the full set still reaches the highest global prediction R^2 (0.77 for the mean sample diameter). This might be expected from the positive and highly significant relationship observed between gene significance (correlation between gene expression and trait value) and connectivity within WGCNA modules displaying a significant correlation with traits. On the other side of networks, peripheral genes predict better with an interactive model than with an additive one and provide over both types of models the most stable predictions (interquartile ranges of 0.19 for peripheral, 0.27 for random, 0.34 for core and Boruta and 0.35 for full set). The information necessary to predict a phenotype does not seem to be particularly concentrated at any side of the network, but rather spread over it, as highlighted by the performance of random gene sets. They capitalize enough information to perform predictions more accurately than an equal number of peripheral genes. Moreover, prediction with larger peripheral sets (20% and 30% of genes) confirmed that peripheral genes need to be in a high number to reach high prediction R^2, as the median doubled between 10% and 30% sets, but not necessarily with more central genes in the network as it tended to decrease with 40% of genes (**Figure S10**, median R^2 of 0.15, 0.23, 0.33 and 0.29 respectively for 10%, 20%, 30% and 40% peripheral gene sets). In that sense, Boruta seems to be extremely useful in focussing on the information that is relevant for prediction. From the 637 genes selected by Boruta, 95 and 22 were core and peripheral genes, respectively. Although the number of core genes within the Boruta set is greater than expected by chance (Fisher's exact test $p \leq 0.0001$), a large majority of Boruta genes still do not belong to this category nor to the peripheral gene set.

Boruta selection proved to be able to select a small number of genes for all of our phenotypes, allowing for a faster and more precise prediction, with less than one-sixth of genes compared to the core or peripheral sets, and only 1.6% of the full set, with predictions being almost as accurate. This makes Boruta an advantageous alternative in genomic evaluation for breeding to more classic methods (based on the imposition of *a priori* constraints for shrinkage or variable selection (de los Campos et al., 2013) like ridge regression. All the reported predictions scores were computed on a test set, which was composed of 20% of the original individuals that were not used to train or validate the models. These results are thus representing real-life results and are not subject to over-fitting. Boruta genes were selected on the training set (60% of the original individuals) and while we could improve this set with validation data, we are fairly safe that we do not bias in favor of overfitting.

Tracking back predictabilities down to particular gene sets is an essential step before being able to understand the roles of interactivity and connectivity in a gene network underlying the phenotype. In that sense, the high levels of connectivity shown by core genes do not appear to be a prerequisite for prediction quality, while these particular genes find better fits in additive models. Contrarily, peripheral genes, while being poorly connected, display prediction quality equivalent to random or Boruta sets in interactive models. This pinpoints to an *a priori* paradoxical situation in which connectivity and interaction are not necessarily found in the same gene sets. Here, connectivity refers to the degree of membership of a given gene within a co-expression network defined independently from any phenotype. Interactivity, on the other hand, refers to the way the expression profile of a given gene is mediated before affecting the

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

phenotype. Such interactivity between gene expressions is quite different from what is usually referred to as epistasis in the genetics literature, the interaction effect between alleles from different loci on a given phenotype (Cordell, 2002), because here the input is gene expressions instead of allelic polymorphisms. Weather connectivity or interactivity relates to epistasis is beyond the scope of current work, but clearly deserves further investigation. In order to clarify this apparent paradox, one hypothetical scenario could be proposed, following the model schematized in **Figure 5**. Basically, in this model, a peripheral gene is located upstream within biological pathways and it produces an essential brick which can be further modified or complemented by the bricks of subsequent genes downstream. The peripheral genes that produce essential bricks do it with a low connection to other genes. As we progress downstream within the pathways, the bricks from peripheral genes suffer a chain of subsequent modifications due to or controlled by other genes, resulting in an impact on the final phenotype that can be highly mediated by many intermediaries, appearing as interactors, that somehow blur the brick's contribution to the ultimate phenotype. This could explain the interactive behavior of peripheral genes, as sitting far away from the final phenotype, while being poorly connected. Core genes, on our schematic model, receive upstream bricks from many peripheral genes, and their output directly impacts or influence the phenotype. This may be a reason why core genes while being highly connected hubs, behave additively, as they almost directly appear to contribute to the phenotype. We have further looked for enrichment in transcription factors (TFs) within the core and peripheral gene sets and found that TFs were overrepresented within the core gene set (Fisher's exact test $p \leq 10^{-14}$), while they were underrepresented within the peripheral gene set (Fisher's exact test $p \leq 10^{-7}$). This leads us to believe that core genes consist in fact of a mixture of highly connected regulators and genes downstream within biological pathways, which altogether contribute to the metabolic flow towards phenotypes. Consequently, they would behave additively when predicting a trait, they could contribute individually to a large proportion of phenotypic variation, and they could, therefore, suffer "first hand" the selection pressure. Core genes variation levels are low by comparison to their expression level and they might display distinct optima according to population structure, as underlined by their higher Q_{ST} and $ScorePCadapt$ in our data. As they depend much on other bricks, they have less room for variation, and are somehow "canalized". Peripheral genes, on the other hand, are highly variable with lower expression levels. They are thus the ones by which variations come to the network and propagate downstream. These observations are consistent with molecular evolution studies, as Jovelin and Phillips (2011) showed a positive correlation between selective constraints and expression level and Fraser and Hirsh (2004) showed that core genes are more expressed, but less variant compared to their expression. Finally, Montanucci et al. (2011) showed a positive correlation between selective constraints and connectivity, which also echoes in our measures and models.

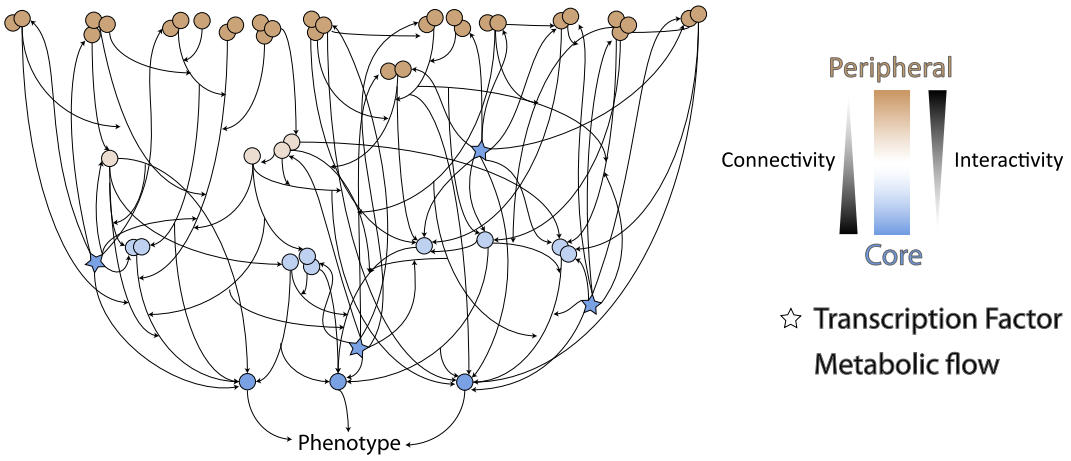


Figure 5: An expanded view of the omnigenic model which handles the observed paradox between connectivity and interactivity. The dots correspond to genes colored according to their connectivity within the network, with core genes in blue and peripheral genes in brown. Stars correspond to transcription factors and arrows represent connections within the network. Hypothetical metabolic pathways are displayed in grey in order to show the upstream-downstream positions of peripheral and core genes, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

A potential limitation of our work is that expression variation associated with a phenotype may not be entirely causal, for instance if a gene is impacted as a side effect by another gene's causality on the phenotype, or if it is reversely impacted by the phenotype. Fully revealing causation is a long path that stretches beyond the aims of this study. One perspective to gain a grasp on causation from the kind of data dealt with in this study would be to use the three-way marginal correlation results amongst DNA, gene expression and phenotype variations and check the extent to which they fit in with coherently.

Our new co-expression omnigenic network model is the first step towards a coherent integration of the terminologies used so far to define particular gene roles in the context of phenotype determinism. Our integrative approach combining predictive and explanatory functions fits well with the omnigenic theory, even when the gene network topology is not trait-centered but self-built with co-expression. It is the case within our Poplar dataset, leading us to think that this theory may be easily generalizable to contrasting biological models, further away from humans and disease-centered traits. Our study highlights the need to widen the concepts of core and peripherals in the functional topology of gene networks and also the importance of connectivity and interaction in setting the characterization of gene roles, which appeared otherwise compatible with proximity to the trait. Our results further suggest that cores' profiles might be more complex than originally proposed by the criteria of precedent studies, involving not only integrative hubs but also regulators, suggesting prospects for further analyses.

Methods

Samples collection

As described in previous works (Gebreselassie et al., 2017; Guet et al., 2015), we established in 2008 a partially replicated experiment with 1160 cloned genotypes, in two contrasting sites in central France (Orléans, ORL) and northern Italy (Savigliano, SAV). At ORL, the total number of genotypes was of 1,098 while at SAV there were 815 genotypes. In both sites, the genotypes were replicated 6 times in a randomized complete block design. At SAV, the trees were pruned at the base after one year of growth (winter 2008-2009) to remove a potential cutting effect and were subsequently evaluated for their growth and wood properties during the winter 2010-2011. At ORL, the trees had the same pruning treatment after two years of growth (winter 2009-2010) and were also subsequently evaluated for growth and wood properties after two years (winter 2011-2012). After evaluation, they were pruned again for a new growth cycle. At their fourth year of growth of this third cycle (2015), 241 genotypes present in two blocks of the French site were selected to perform sampling for RNA sequencing. In the end, we obtained transcriptomic data from 459 samples, 218 genotypes duplicated in the two blocks and 23 genotypes available from only one block. These 241 genotypes were representative of the natural west European range of *P. nigra* through 11 river catchments in 4 countries (**Table S3**).

We described 14 of the 17 phenotypic traits in previous work (Gebreselassie et al., 2017). Briefly, these traits can be divided into two categories, growth traits and biochemical traits which were all evaluated on up to 6 clonal replicates by genotype at each site after two years of growth in the second cycle. The first set is composed by the circumference of the tree at a 1-meter height measured in Savigliano at the end of 2010 (CIRC.Sav) and in Orléans at the end of 2011 (CIRC.Orl). The second set is composed, each time at both sites, of measures of ratios between the different components of the lignin, p-hydroxyphenyl (H), guaiacyl (G) and syringyl (S) (H.G.Orl, H.G.Sav, S.G.Orl and S.G.Sav), measures of the total lignin content (Lignin.Orl : measure of the lignin in Orléans, Lignin.Sav: measure of the lignin in Savigliano), measure of the total glucose (Glucose.Orl and Glucose.Sav), measure of ratio between 5 and 6 carbon sugars (C5.C6.Orl and C5.C6.Sav) and measure of the extractives (Extractives.Orl and Extractives.Sav). For each of these traits, we computed mean values per genotype previously adjusted for microenvironmental effects (block or spatial position in the field).

The 3 remaining traits were measured in 2015 on the trees harvested for the RNA sequencing experiment (2 replicates per genotype). They include the mean diameter of the stem section harvested for RNA sequencing (MeanDiameter), the date of bud flush of the tree in 2015 (Date3Doy) and the basic density of the wood (Infraden). Date of bud flush consisted in a prediction of the day of the year at which the apical bud of the tree was in stage 3 according to the scale defined in Dillen et al. (2009). Predictions were done with a lowess regression from discrete scores recorded at consecutive dates in the spring of 2015. Wood basic density was measured on a piece of wood from the stem section harvested for RNA sequencing following the Technical Association of Pulp and Paper Industry (TAPPI) standard test method T 258 "Basic density and moisture content of pulpwood".

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Transcriptome data generation

We sampled stem sections of approximately 80 cm long starting at 20 cm above ground and up to 1 meter in June 2015. The bark was detached from the trunk in order to scratch young differentiating xylem and cambium tissues using a scalpel. The tissues were immediately immersed in liquid nitrogen and crudely ground before storage at -80°C pending milling and RNA extraction. Prior to RNA extraction, the samples were finely milled with a swing mill (Retsch, Germany) and tungsten beads under cryogenic conditions with liquid nitrogen during 25 seconds (frequency 25 cps/sec). About 100 mg of milled tissue was used to isolate separately total RNA from xylem and cambium of each tree with RNeasy Plant kit (Qiagen, France), according to manufacturer's recommendations. Treatment with DNase I (Qiagen, France) to ensure elimination of genomic DNA was made during this purification step. RNA was eluted in RNAse-DNAse free water and quantified with a Nanodrop spectrophotometer. RNA from xylem and cambium of the same tree were pooled in an equimolar extract (250 ng/μL) before sending to the sequencing platform.

RNAseq experiment was carried out at the platform POPS (transcriptOmic Platform of Institute of Plant Sciences - Paris-Saclay) thanks to IG-CNS Illumina Hiseq2000. RNAseq libraries were constructed using TruSeq_Stranded_mRNA_SamplePrep_Guide_150 31047_D protocol (Illumina®, California, U.S.A.). The RNAseq samples have been sequenced in single-end reads (SR) with an insert library size of 260 bp and a read length of 100 bases. Images from the instruments were processed using the manufacturer's pipeline software to generate FASTQ sequence files. Ten samples by lane of Hiseq2000 using individually barcoded adapters gave approximately 20 million of SR per sample. We mapped the reads on the *Populus trichocarpa* v3.0 transcriptome with bowtie2 (Langmead and Salzberg, 2012) and obtained the read counts for each of the 41,335 transcripts by in-house scripts (17 million of reads were mapped per sample in median, with a minimum of 6 and a maximum of 42 million). Initially, we considered using the genotype mean to reduce our data volume. However, differences between replicates were not normally distributed, because of variation in gene expression due to plasticity. We thus could not summarize our data with their mean, as it would have removed this information and finally we chose to keep replicates as separate data samples.

Filtering the non-expressed genes, normalization and variance stabilization

As the sampling ran along 2 weeks, we expected environmental variables to blur the signal. To understand how our data were impacted, we ran the first analysis, containing a step identifying the impact of each cofactor and a step correcting confounding effects, with mixed linear models implemented in the R package breedR (Muñoz and Sanchez, 2017). However, while it was properly correcting the covariables that seemed to impact our data (environmental effects) when controlling on PCA spaces, it was also erasing useful information from the data, yielding less accurate prediction models than without any correction. We thus chose not to perform this correction, and use raw uncorrected data.

We started cleaning our raw counts data by removing the transcripts with 0 counts for all individuals. From the original 41,335 genes, 1,653 were thus removed, leaving 39,682 genes with at least 1 count in at least 1 individual. Only 1,931 genes had between 1 and 5 reads mapped across all individuals. We tried to filter more stringently, for instance by splitting the data with a 2 groups k-means clustering, but again it reduced predictions accuracy. After this first filtration, we normalized the raw counts data by Trimmed Mean of M-values (TMM, edgeR (Robinson and Oshlack, 2010)). As most features are not differentially expressed, this method takes into account the fact that the total number of reads can be strongly influenced by a low number of features. Then, we calculated the counts per million (CPM (Law et al., 2014)).

To stabilize the variance of the CPM data, we computed a $log_2(n+1)$ instead of a $log_2(n+0.5)$ typically used in a voom analysis (Law et al., 2014). The reason is that the former avoids negative values, which are problematic for the rest of the analysis. The resulting data set was called full.

Hierarchical and k-means clustering

We performed a weighted correlation network analysis with the R package WGCNA (Langfelder and Horvath, 2008) on our full RNAseq gene set. We followed the classic approach, except that we first ranked our expression data, to work subsequently with Spearman's non-parametric correlations and avoid problems due to linear modeling assumptions. We first chose the soft threshold with the highest scale-free topology fitting index ($R^2 = 0.85$), which is for a power of 15 (connectivity: mean = 22.90, median = 8.94, max = 329, **Figure 1A**). Then, we used the automatic module detection (blockwiseModules

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

function) via dynamic tree cutting with a merging threshold of 0.25, a minimum module size of 30 and bidweight midcorrelations (**Figure 1B**). All other options were left to default. This also computes module eigengenes. To sort the traits, we clustered their scaled values with the pvclust R packages (Suzuki and Shimodaira, 2015), the Ward agglomerative method ("Ward.D2") on correlations (**Figure 1B, 1C**, **Figure S3**). The clustering on euclidean distance results in the exact same hierarchical tree. Correlations between traits and gene expression or module eigengenes were computed as Spearman's rank correlations (**Figure 1B, 1C**).

We also performed a k-means clustering with the R package coseq (Rau and Maugis-Rabusseau, 2017) considering 10 initial runs, 1000 iterations, without any other data transformation, and for a number of clusters (K) between 2 and 20. At first, it identified a K without strong agreement between the two evaluation algorithms included in coseq. We thus further computed additional rounds of k-means clustering, around the previously identified K (plus or minus 5 clusters), with 100 initial runs and 10000 iterations, until both evaluation algorithms agreed.

Machine learning

Boruta gene expression selection

In addition to the inconvenience of working with a large number of features (time and power consumption), most machine learning algorithms perform better when the number of predicting variables used is kept as low as the optimal set (Kohavi and John, 1997). We thus performed an all relevant variable selection (Nilsson et al., 2007) with the Boruta function (Kursa and Rudnicki, 2010) from the eponym R package, with a 5% p-value threshold, on the training subpart of the full gene expression set, for each phenotype independently. Then, features that were not rejected by the algorithm were pooled together, so that all the important genes were in the selected gene pool.

Models

Both additive linear model (ridge regression) and interactive neural networks models were computed by the R package h2o (LeDell et al., 2019). They both used the gene expression sets as predictors and one phenotypic trait at a time as a response. Gene sets were split by the function h2o.splitFrame into 3 sets, a training set, a validation set and a test set, with the respective proportions of 60%, 20%, 20%. We checked that the split preserves the distribution of samples within populations. The training set was used to train the models, the validation set was used to validate and improve the models, while the test set was used to compute and report prediction accuracies as R^2 between observed and predicted values within this set and using the function R2 of the R package compositions (van den Boogaart et al., 2018). This set has never been used to improve the model and therefore represents a proxy of new data, avoiding the report of results from overfitted models.

For linear models, we used the function h2o.glm with default parameters, except 2-folds cross-validation and alpha set at zero to perform a ridge regression. The same splits and score reporting methods were used.

Neural networks have the reputation to be able to predict any problem, based on the Universal approximation theorem (Cybenko, 1989; Hornik et al., 1989). However, this capacity comes at the cost of a very large number of neurons in one layer, or a reasonable number of neurons per layer in a high number of layers. Both settings lead to difficult interpretation when very many gene expressions are involved. In that sense, we chose to keep our models simple, with two layers of a reasonable number of neurons. This obviously comes at the price of lower prediction power. However, we believe that these topologies give us the power to model 2 levels of interactions between genes (1 level per layer). Furthermore, since both methods yielded comparable prediction R^2 (median ridge regression $R^2 = 0.27$, mean neural network $R^2 = 0.22$), this complexity seemed appropriate. To find the best models for neural networks, we computed a random grid for each response. We tested the following four hyperparameters: (i) activation function ("Rectifier", "Tanh", "RectifierWithDropout" or "TanhWithDropout"); (ii) network structure; (iii) input layer dropout ratio (0 or 0.2) (iv) L1 and L2 regularization (comprised between 0 and 1×10^{-4}, with steps of 5×10^{-6}). Network structure corresponded to the number of neurons within each of the two hidden layers, which was based on the number of input genes (h). The first layer was composed of h, $\frac{2}{3}$h or $\frac{1}{3}$h neurons. The second layer had a number of nodes equal or lower to the first one and is also composed of h, $\frac{2}{3}$h or $\frac{1}{3}$h neurons. This represented a total of 6 different structures. We performed a random discrete strategy to find the best search criteria, computing a maximum of 100 models, with a stopping tolerance of 1×10^{-3} and 10 stopping rounds. Finally, h2o.grid parameters were the following: the algorithm was "deeplearning", with 10 epochs, 2 fold cross-validation, maximum duty cycle fraction for scoring is 0.025 constraint for a squared sum of incoming weights per unit is 10.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

All other parameters were set to default values. The best model was selected from the lowest RMSE score within the validation set.

Heritability and Qst Models

A 12k bead chip (Faivre-Rampant et al., 2016) provided 7,896 SNPs in our population. A genomic relationship matrix between genotypes was computed with these SNPs with LDAK (Speed et al., 2012), and further split into between (mean population kinship, $\mathbf{K_b}$) and within population relationship matrices (kinship kept only for the members of the same population, all the others are equal to 0, $\mathbf{K_w}$). These matrices were used in a mixed linear model to compute the additive genetic variances between (σ_b^2) and within (σ_w^2) populations for the expression of each gene as follows:

$$\mathbf{y} = \beta_0 + \mathbf{Z_b b} + \mathbf{Z_w w} + \epsilon \quad (1)$$

In this model, y is a gene expression vector across individual trees, β_0 is a vector of fixed effects (overall mean or intercept); $\mathbf{b}$ and $\mathbf{w}$ are respectively random effects of populations and individuals within populations, which follow normal distributions, centered around 0, of variance $\sigma_b^2\mathbf{K_b}$ and $\sigma_w^2\mathbf{K_w}$. $\mathbf{Z_b}$ and $\mathbf{Z_w}$ are known incidence matrices between and within populations, relating observations to random effects $\mathbf{b}$ and $\mathbf{w}$. ϵ is the residual component of gene expression, following a normal distribution centered around 0, of variance $\sigma_\epsilon^2\mathbf{I}$, where σ_ϵ^2 is the residual variance and $\mathbf{I}$ is an identity matrix. From the between and within population variance components, we computed heritability (h^2) and population differentiation estimates (Q_{ST}) for each gene as follows:

$$h^2 = \frac{\sigma_b^2 + \sigma_w^2}{\sigma_b^2 + \sigma_w^2 + \sigma_\epsilon^2} \quad (2)$$

$$Q_{ST} = \frac{\sigma_b^2}{\sigma_b^2 + 2\sigma_w^2} \quad (3)$$

To compute them, we used the function remlf90 from the R package breedR (Muñoz and Sanchez, 2017), with the Expectation-Maximization method followed by one round with Average-Information algorithm to compute the standard deviations. We computed the genetic variation coefficient (CVg) by dividing sums of σ_b^2 and σ_w^2 by expression mean, per gene.

Other population statistics

We further used a previously developed bioinformatics pipeline to call SNPs within our RNA sequences (Rogier et al., 2018). Briefly, this pipeline involves classical cleaning and quality control steps, mapping on the *Populus trichocarpa* v3.0 reference genome, and SNP calling using the combination of four different callers. We ended up with a set of 874,923 SNPs having less than 50% of missing values per genotype. The missing values were further imputed with the software FImpute (Sargolzaei et al., 2014)). We validated our genotyping by RNAseq approach by comparing the genotype calls with genotyping previously obtained with an SNP chip on the same individuals (Faivre-Rampant et al., 2016)). Genotyping accuracy based on 3,841 common positions was very high, with a mean value of 0.96 and a median value of 0.99. The imputed set of SNP was then annotated using Annovar (Wang et al., 2010) in order to group the SNPs per gene model of *P. trichocarpa* reference genome. For each SNP, we computed the overall genetic diversity statistic (Ht) with the hierfstat R package ((Goudet and Jombart, 2015) and this statistic was then averaged by gene model in order to get information on the extent of diversity. We further computed $ScorePCadapt$ with the pcadapt R package (Luu et al., 2017) with 8 retained principal components. Here again, $ScorePCadapt$ were then summarized (averaged) by gene model in order to get information about their potential involvement in adaptation. Based on the principal component analysis, PCadapt is more powerful to perform genome scans for selection in next-generation sequencing data than approaches based on F_{ST} outliers detection (Luu et al., 2017). We found a positive correlation between F_{ST} and $ScorePCadapt$ (data not shown), but PCadapt showed differences between Core, random and peripheral gene sets (**Figure 4B**) when F_{ST} did not.

Transcription factors enrichment analysis

We have tested each of the gene sets (core, peripheral, Boruta, random) for enrichment in transcription factors, with data coming from the plant TFDB (Jin et al., 2017). We selected in each set transcripts based on loci, regardless of the transcription factor families sharing different versions of the gene. Fisher's exact test was performed with the base R function fisher.test.

Data Access

This RNAseq project has been submitted to the international repository Gene Expression Omnibus (GEO) from NCBI (accession number: GSE128482). All

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

steps of the experiment, from growth conditions to bioinformatic analyses are detailed in CATdb (Gagnot et al., 2007) according to the MINSEQE "minimum information about a high-throughput sequencing experiment". Raw sequences (FASTQ) are being deposited in the Sequence Read Archive (SRA) from NCBI. Information on the studied genotypes is available in the GnpIS Information System (Steinbach et al., 2013).

Acknowledgements

The authors gratefully acknowledge the staff of the INRA GBFOR experimental unit for the establishment and management of the poplar experimental design in Orléans, the collection of wood samples in each site, and their contribution to phenotypic measurements on poplars in Orléans; Alasia Franco Vivai staff for management of the poplar experimental plantation in Savigliano, and M. Sabatti and F. Fabbrini for their contribution to phenotypic measurements on poplars in Savigliano. We acknowledge the staff of BioForA for their contribution to RNA collection in the field. We are grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrennées for providing computing and storage resources. We would also like to thank M. Nordborg for useful discussions on this work and J. Salse for useful comments on the manuscript.

Funding

Establishment and management of the experimental sites were carried out with financial support from the NOVELTREE project (EU-FP7-211868). RNA collection, extraction, and sequencing were supported by the SYBIOPOP project (ANR-13-JSV6-0001) funded by the French National Research Agency (ANR). The platform POPS benefits from the support of the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS).

Authors' Contributions

AC, LS, and VS designed the experiment, discussed the results and wrote this manuscript. AC ran the in silico experiment. MCL, VB, CPL, LT, MLMM, and VS contributed the RNAseq data production and analysis. VJ, OR and VS contributed to the SNP data production and analysis. MLMM and JCL contributed to the discussion on the methodology employed. All the authors read and approved this manuscript.

Disclosure declaration

The authors declare that they have no competing interests.

References

Biernacki, C., Celeux, G., Govaert, G., and Langrognet, F., 2006. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis*, **51**(2):587–600.

Bloom, J. D. and Adami, C., 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level: response. *BMC evolutionary biology*, **4**(1):14.

Boyle, E. A., Li, Y. I., and Pritchard, J. K., 2017. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, **169**(7):1177–1186.

Cordell, H. J., 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**(20):2463–2468.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, **2**(4):303–314.

de los Campos, G., Hickey, J. M., Pong-Wong, R., Daetwyler, H. D., and Calus, M. P. L., 2013. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. *Genetics*, **193**(2):327–345.

Dillen, S. Y., Marron, N., Sabatti, M., Ceulemans, R., and Bastien, C., 2009. Relationships among productivity determinants in two hybrid poplar families grown during three years at two contrasting sites. *Tree Physiology*, **29**(8):975–987.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H., 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, **102**(40):14338–14343.

Duret, L. and Mouchiroud, D., 2000. Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Molecular Biology and Evolution*, **17**(1):68–070.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Faivre-Rampant, P., Zaina, G., Jorge, V., Giacomello, S., Segura, V., Scalabrin, S., Guérin, V., De Paoli, E., Aluome, C., Viger, M., *et al.*, 2016. New resources for genetic studies in Populus nigra: genome-wide SNP discovery and development of a 12k Infinium array. *Molecular ecology resources*, **16**(4):1023–1036.

Fraser, H. B. and Hirsh, A. E., 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC evolutionary biology*, **4**(1):13.

Gagnot, S., Tamby, J.-P., Martin-Magniette, M.-L., Bitton, F., Taconnat, L., Balzergue, S., Aubourg, S., Renou, J.-P., Lecharny, A., and Brunaud, V., *et al.*, 2007. CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. *Nucleic Acids Research*, **36**(Database):D986–D990.

Gebreselassie, M. N., Ader, K., Boizot, N., Millier, F., Charpentier, J.-P. P., Alves, A., Simões, R., Rodrigues, J. C., Bodineau, G., Fabbrini, F., *et al.*, 2017. Near-infrared spectroscopy enables the genetic analysis of chemical properties in a large set of wood samples from Populus nigra (L.) natural populations. *Industrial Crops and Products*, **107**(January):159–171.

González-Recio, O., Rosa, G. J., and Gianola, D., 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livestock Science*, **166**:217–231.

Goudet, J. and Jombart, T., 2015. *hierfstat: Estimation and Tests of Hierarchical F-Statistics*. R package version 0.04-22.

Guet, J., Fabbrini, F., Fichot, R., Sabatti, M., Bastien, C., and Brignolas, F., 2015. Genetic variation for leaf morphology, leaf structure and leaf carbon isotope discrimination in European populations of black poplar (Populus nigra L.). *Tree Physiology*, **35**(8):850–863.

Han, M., Qin, S., Song, X., Li, Y., Jin, P., Chen, L., and Ma, F., 2013. Evolutionary rate patterns of genes involved in the Drosophila Toll and Imd signaling pathway. *BMC Evolutionary Biology*, **13**(1):245.

Han, Y., Gao, S., Muegge, K., Zhang, W., and Zhou, B., 2015. Advanced Applications of RNA Sequencing and Challenges. *Bioinformatics and Biology Insights*, **9s1**:BBI.S28991.

Hornik, K., Stinchcombe, M., and White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**(5):359–366.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., and Gao, G., 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Research*, **45**(D1):D1040–D1045.

Josephs, E., Lee, Y. W., Stinchcombe, J. R., and Wright, S. I., 2015. Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *PNAS*, **112**(50):1–6.

Josephs, E. B., Wright, S. I., Stinchcombe, J. R., and Schoen, D. J., 2017. The Relationship between Selection, Network Connectivity, and Regulatory Variation within a Population of Capsella grandiflora. *Genome Biology and Evolution*, **9**(4):1099–1109.

Jovelin, R. and Phillips, P. C., 2011. Expression Level Drives the Pattern of Selective Constraints along the Insulin/Tor Signal Transduction Pathway in Caenorhabditis. *Genome Biology and Evolution*, **3**:715–722.

Kohavi, R. and John, G. H., 1997. Wrappers for feature subset selection. *Artificial Intelligence*, **97**(1-2):273–324.

Kursa, M. B. and Rudnicki, W. R., 2010. Feature Selection with the Boruta Package. *Journal of Statistical Software*, **36**(11):1–13.

Langfelder, P. and Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, **9**(1):559.

Langmead, B. and Salzberg, S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, **9**(4):357–359.

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**(2):R29.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., *et al.*, 2019. *h2o: R Interface for 'H2O'*. R package version 3.22.1.1.

Liu, X., Li, Y. I., and Pritchard, J. K., 2018. Trans effects on gene expression can drive omnigenic inheritance. *bioRxiv*, :425108.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Lu, Y., 2003. Evolutionary Rate Variation in Anthocyanin Pathway Genes. *Molecular Biology and Evolution*, **20**(11):1844–1853.

Luu, K., Bazin, E., and Blum, M. G., 2017. pcadapt: an R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, **17**(1):67–77.

Mackay, T. F. C., Stone, E. a., and Ayroles, J. F., 2009. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, **10**(8):565–577.

Maher, B., 2008. Personal genomes: The case of the missing heritability. *Nature*, **456**(7218):18–21.

Mähler, N., Wang, J., Terebieniec, B. K., Ingvarsson, P. K., Street, N. R., and Hvidsten, T. R., 2017. Gene co-expression network connectivity is an important determinant of selective constraint. *PLOS Genetics*, **13**(4):e1006402.

Montanucci, L., Laayouni, H., Dall'Olio, G. M., and Bertranpetit, J., 2011. Molecular Evolution and Network-Level Analysis of the N-Glycosylation Metabolic Pathway Across Primates. *Molecular Biology and Evolution*, **28**(1):813–823.

Muñoz, F. and Sanchez, L., 2017. *breedR: Statistical Methods for Forest Genetic Resources Analysts*. R package version 0.12-2.

Nilsson, R., PeñaPe, J. M., Jmp, P., Björkegren JOHANBJORKEGREN, J., and Tegnér JESPERT, J., 2007. Consistent Feature Selection for Pattern Recognition in Polynomial Time. Technical report.

Pál, C., Papp, B., and Hurst, L. D., 2001. Highly expressed genes in yeast evolve slowly. *Genetics*, **158**(2):927–31.

Rau, A. and Maugis-Rabusseau, C., 2017. Transformation and model choice for RNA-seq co-expression analysis. *Briefings in Bioinformatics*, **19**(3):bbw128.

Rausher, M. D., Lu, Y., and Meyer, K., 2008. Variation in Constraint Versus Positive Selection as an Explanation for Evolutionary Rate Variation Among Anthocyanin Genes. *Journal of Molecular Evolution*, **67**(2):137–144.

Rausher, M. D., Miller, R. E., and Tiffin, P., 1999. Patterns of evolutionary rate variation among genes of the anthocyanin biosynthetic pathway. *Molecular Biology and Evolution*, **16**(2):266–274.

Riley, R. M., Jin, W., and Gibson, G., 2003. Contrasting selection pressures on components of the Ras-mediated signal transduction pathway in Drosophila. *Molecular Ecology*, **12**(5):1315–1323.

Robinson, M. D. and Oshlack, A., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, **11**(3):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., *et al.*, 2018. Accuracy of RNAseq based SNP discovery and genotyping in Populusnigra. *BMC Genomics*, **19**(1):909.

Sargolzaei, M., Chesnais, J. P., and Schenkel, F. S., 2014. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics*, **15**(1).

Sicard, A., Kappel, C., Josephs, E. B., Lee, Y. W., Marona, C., Stinchcombe, J. R., Wright, S. I., and Lenhard, M., 2015. Divergent sorting of a balanced ancestral polymorphism underlies the establishment of gene-flow barriers in Capsella. *Nature Communications*, **6**(1):7960.

Song, X., Jin, P., Qin, S., Chen, L., and Ma, F., 2012. The Evolution and Origin of Animal Toll-Like Receptor Signaling Pathway Revealed by Network-Level Molecular Evolutionary Analyses. *PLoS ONE*, **7**(12):e51657.

Speed, D., Hemani, G., Johnson, M. R., and Balding, D. J., 2012. Improved heritability estimation from genome-wide SNPs. *American Journal of Human Genetics*, **91**(6):1011–1021.

Steinbach, D., Alaux, M., Amselem, J., Choisne, N., Durand, S., Flores, R., Keliet, A.-O., Kimmel, E., Lapalu, N., Luyten, I., *et al.*, 2013. GnpIS: an information system to integrate genetic and genomic data from plants and fungi. *Database*, **2013**(0):bat058–bat058.

Suzuki, R. and Shimodaira, H., 2015. *pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling*. R package version 2.0-0.

van den Boogaart, K. G., Tolosana-Delgado, R., and Bren, M., 2018. *compositions: Compositional Data Analysis*. R package version 1.40-2.

Wang, K., Li, M., and Hakonarson, H., 2010. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, **38**(16).

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Williamson, S. H., Hernandez, R., Fledel-Alon, A., Zhu, L., Nielsen, R., and Bustamante, C. D., 2005. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences*, **102**(22):7882–7887.

Wu, X., Chi, X., Wang, P., Zheng, D., Ding, R., and Li, Y., 2010. The evolutionary rate variation among genes of HOG-signaling pathway in yeast genomes. *Biology Direct*, **5**(1):46.

Yu, H.-S., Shen, Y.-H., Yuan, G.-X., Hu, Y.-G., Xu, H.-E., Xiang, Z.-H., and Zhang, Z., 2011. Evidence of Selection at Melanin Synthesis Pathway Loci during Silkworm Domestication. *Molecular Biology and Evolution*, **28**(6):1785–1799.

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplemental Material

Supplemental tables

Table S1: Correspondence between traits, their abbreviations, and families.

Trait	Abbreviation	Family
Mean diameter of the stem section harvested for RNA sequencing	MeanDiameter	Growth
Circumference in Orléans	CIRC.Orl	Growth
Circumference in Savigliano	CIRC.Sav	Growth
Ratio between syringyl and guaiacyl lignin subunits in Orléans	S.G.Orl	Chemical
Ratio between syringyl and guaiacyl lignin subunits in Savigliano	S.G.Sav	Chemical
Total glucose in Orléans	Glucose.Orl	Chemical
Total glucose in Savigliano	Glucose.Sav	Chemical
Basic wood density of the stem section harvested for RNA sequencing	Infraden	Physical
Ratio between 5 carbon- and 6 carbon-sugars in Orléans	C5.C6.Orl	Chemical
Ratio between 5 carbon- and 6 carbon-sugars in Savigliano	C5.C6.Sav	Chemical
Ratio between p-hydroxyphenyl and guaiacyl lignin subunits in Orléans	H.G.Orl	Chemical
Ratio between p-hydroxyphenyl and guaiacyl lignin subunits in Savigliano	H.G.Sav	Chemical
Lignin content in Orléans	Lignin.Orl	Chemical
Lignin content in Savigliano	Lignin.Sav	Chemical
Extractives content in Orléans	Extractives.Orl	Chemical
Extractives content in Savigliano	Extractives.Sav	Chemical
Date of bud flush of the tree in Orléans in 2015	Date3Doy	Phenology

Table S2: Module membership of each gene (see **Supplemental file**).

Table S3: Number of genotypes sampled for each population.

Population name	Country	Number of genotypes
Adour	France	36
Basento	Italy	5
Dranse	France	16
Kuhkopf	Germany	19
Loire	France	34
NL	Netherlands	4
Paglia	Italy	13
Ramieres	France	26
Rhin	France	15
Ticino	Italy	54
ValAllier	France	19

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

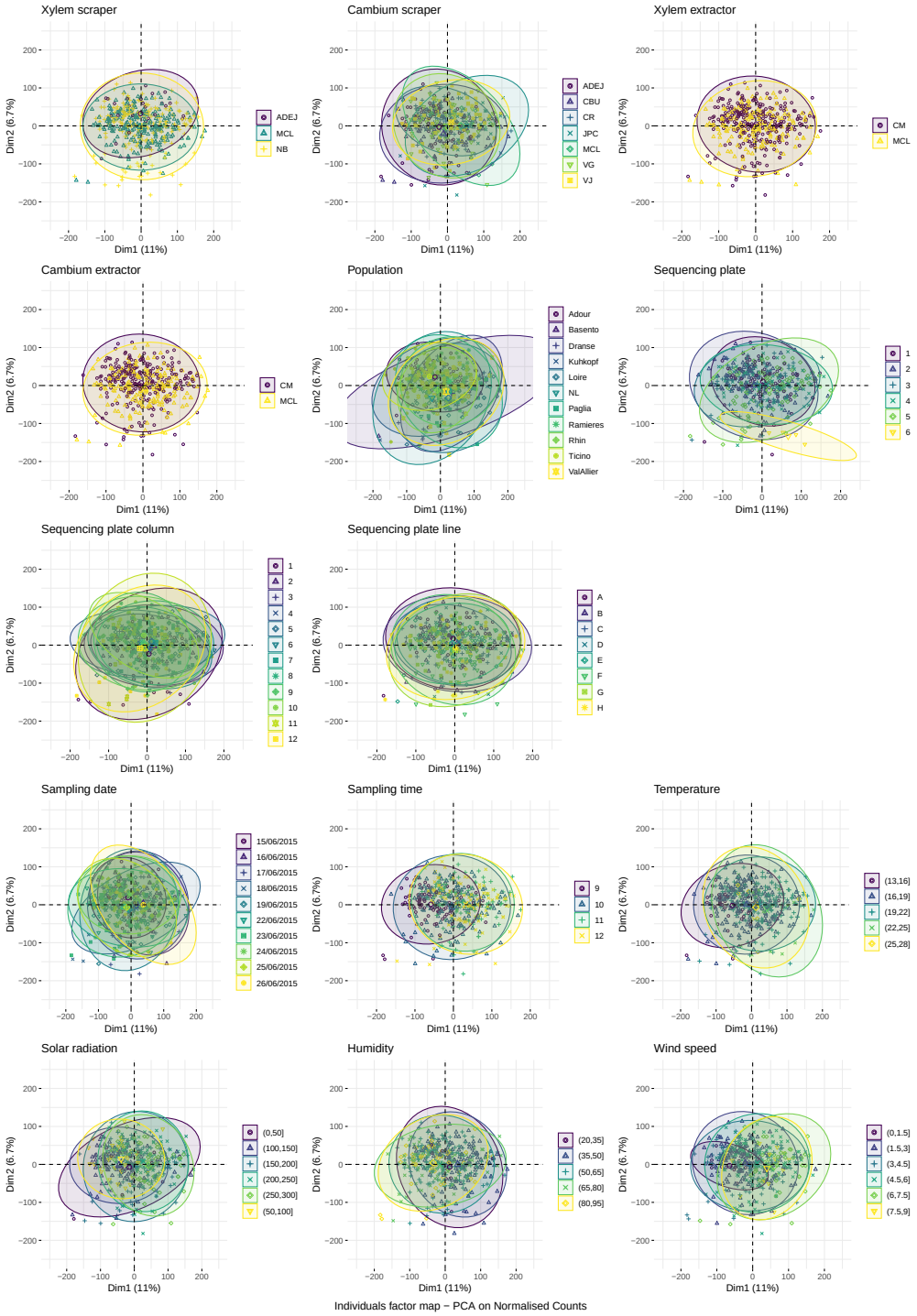
Table S4: Distribution of core and peripheral genes across modules.

Module	Number of core genes	Number of peripheral genes
black	262	1
blue	596	0
brown	191	0
cyan	51	0
darkgreen	17	0
darkgrey	0	0
darkred	20	0
darkturquoise	18	0
green	145	1
greenyellow	132	0
grey	0	3927
grey60	93	0
lightcyan	55	0
lightgreen	40	15
lightyellow	46	0
magenta	202	0
midnightblue	186	0
pink	126	0
purple	156	6
red	177	0
royalblue	29	0
salmon	140	0
tan	300	0
turquoise	807	18
yellow	179	0

bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

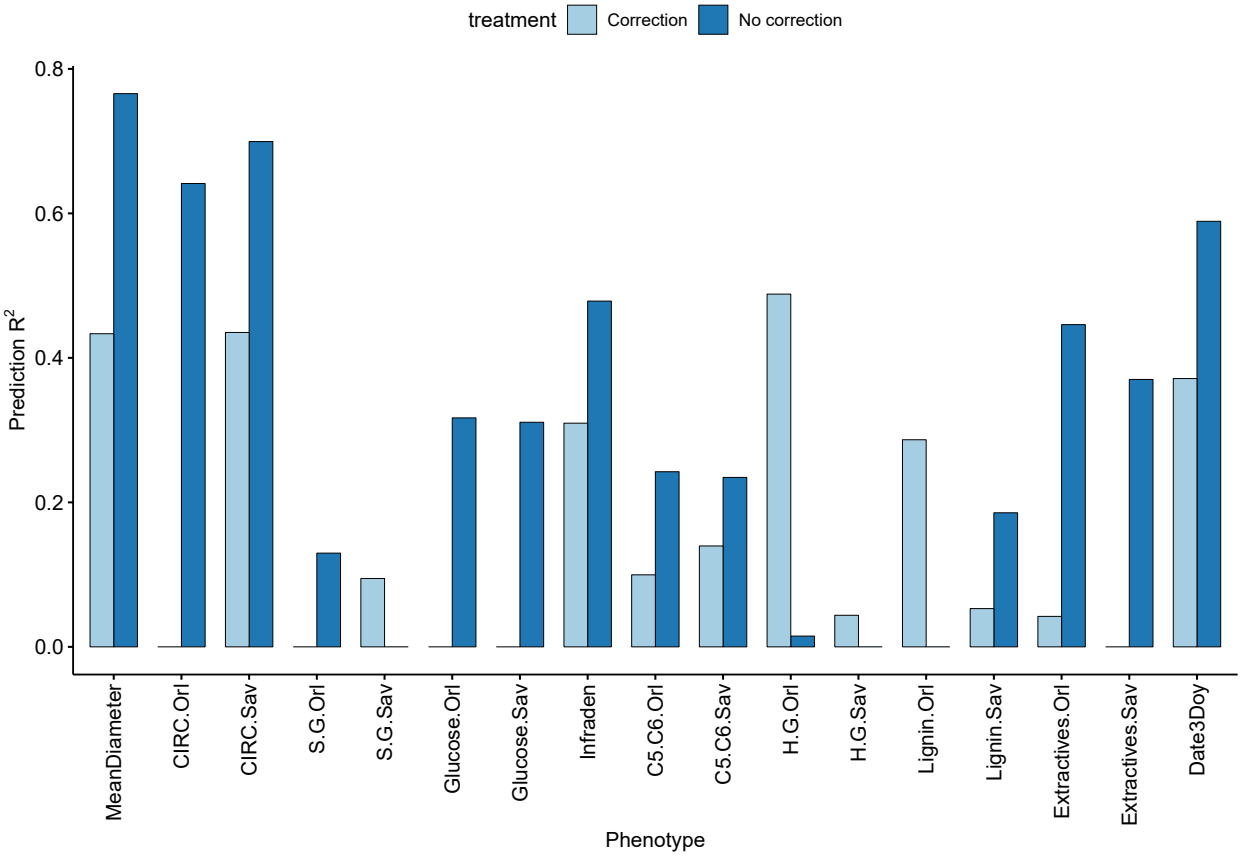
Supplemental figures

Figure S1: PCA of the different cofactors (Xylem and cambium scraper, extractor and extraction method, population, sequencing column, line and plate, the growth rate at harvest, sampling date, time, temperature, solar radiation, humidity and wind speed). Each of these represents the distribution of the individuals on the 2 first axes of the PCA (representing 17,7% of the variation), colored by class. Cofactors related to weather are presented in the 6 lower plots.



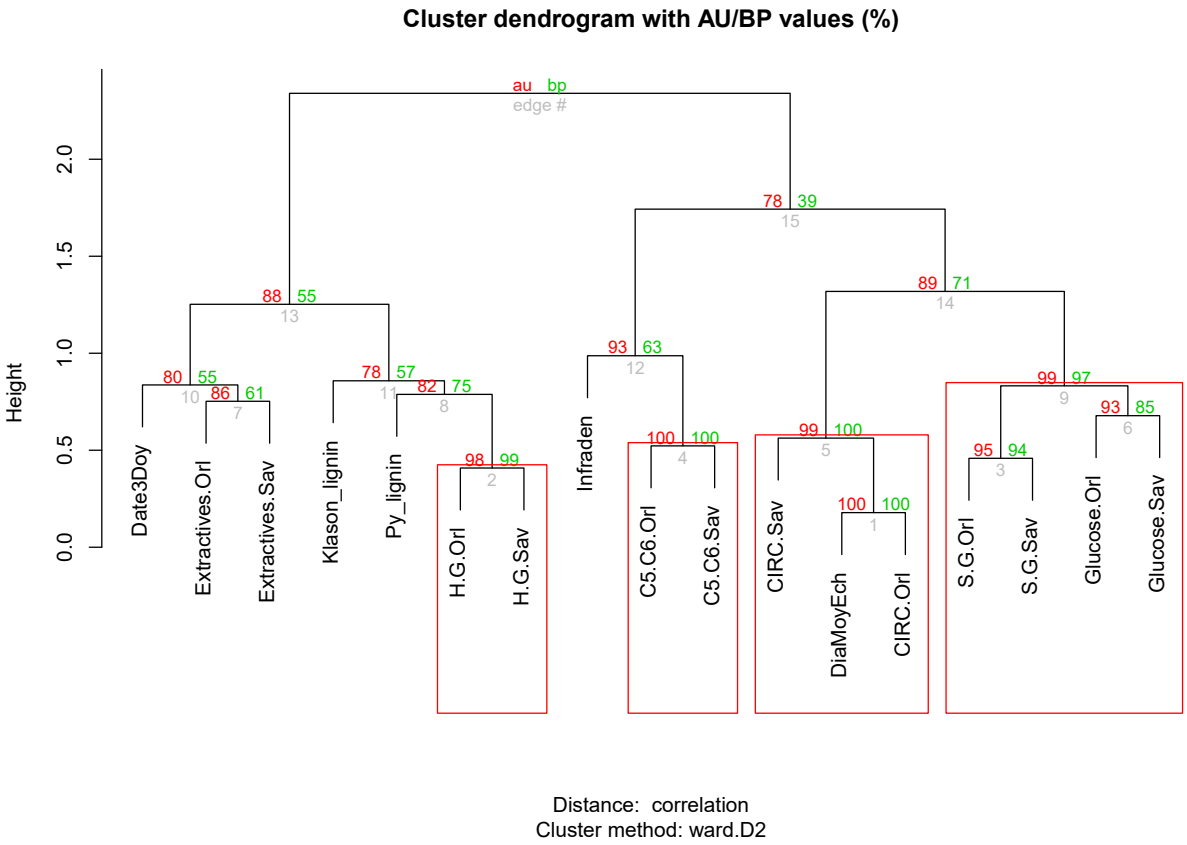
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S2: Prediction scores on test sets (R^2 on the y-axis) for the LM Ridge algorithm for each phenotypic trait (on the x-axis) with (light blue) and without (dark blue) correction of the environmental cofactors.



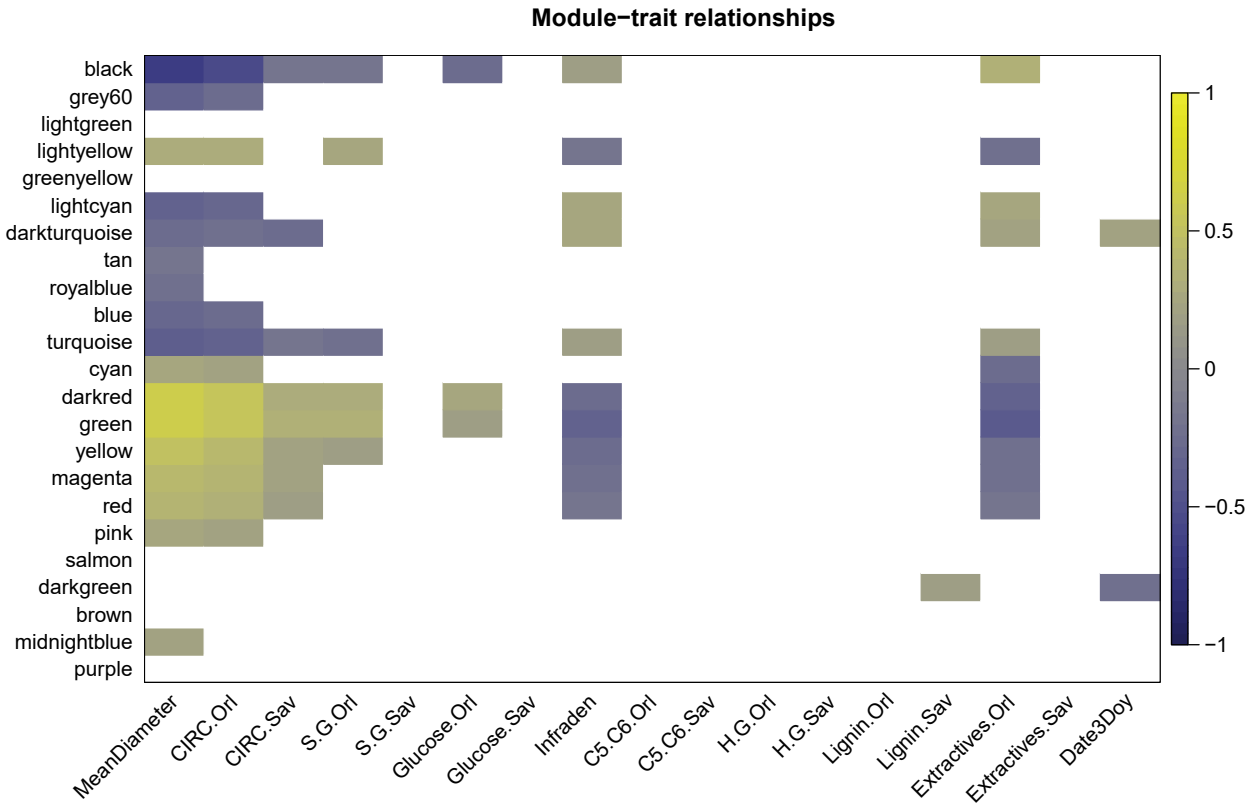
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S3: Scaled traits hierarchical clustering dendrogram computed from their correlations with Ward method ("Ward.D2") by the R package pvclust. Approximately Unbiased (au, in red) and Bootstrap Probability (bp, in green) p-values indicated the degree of belief associated with clusters. Highly supported modules are framed by a red square, grouping (a) the mean sample diameter with the two circumference traits, (b) the S/G ratios with glucose composition, (c) the two C5/C6 together, and (d) the H/G ratios.



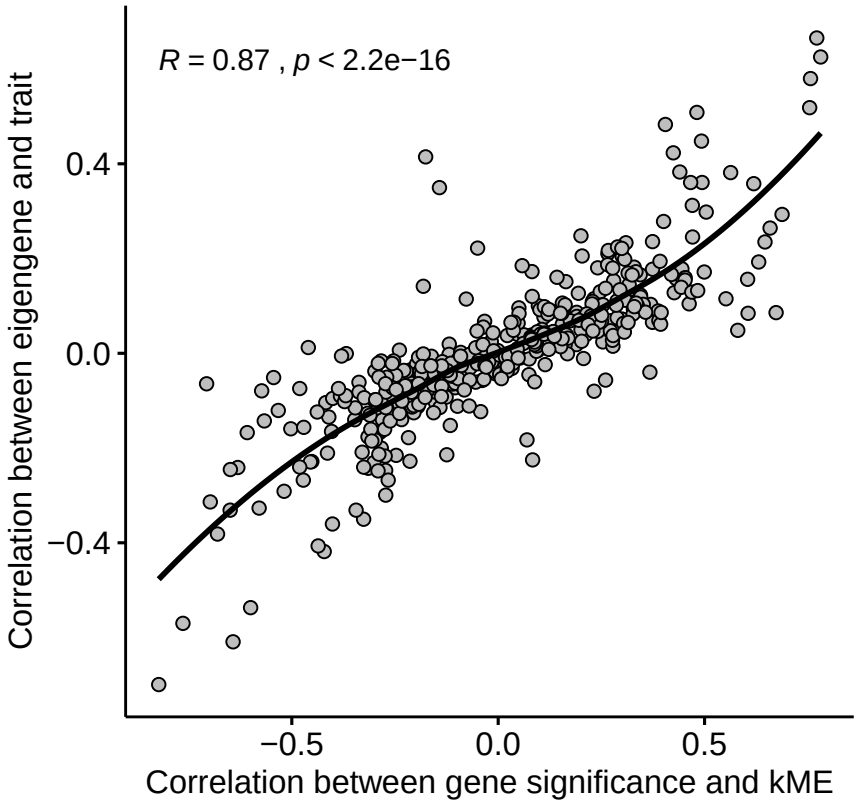
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S4: Heatmap of module-trait Spearman's correlations, on a dark blue (high negative correlation) to light yellow (high positive correlation) scale. We removed correlations with a p-value lower than 5% after Bonferroni correction. From the total of 425 correlations, 72 remained.



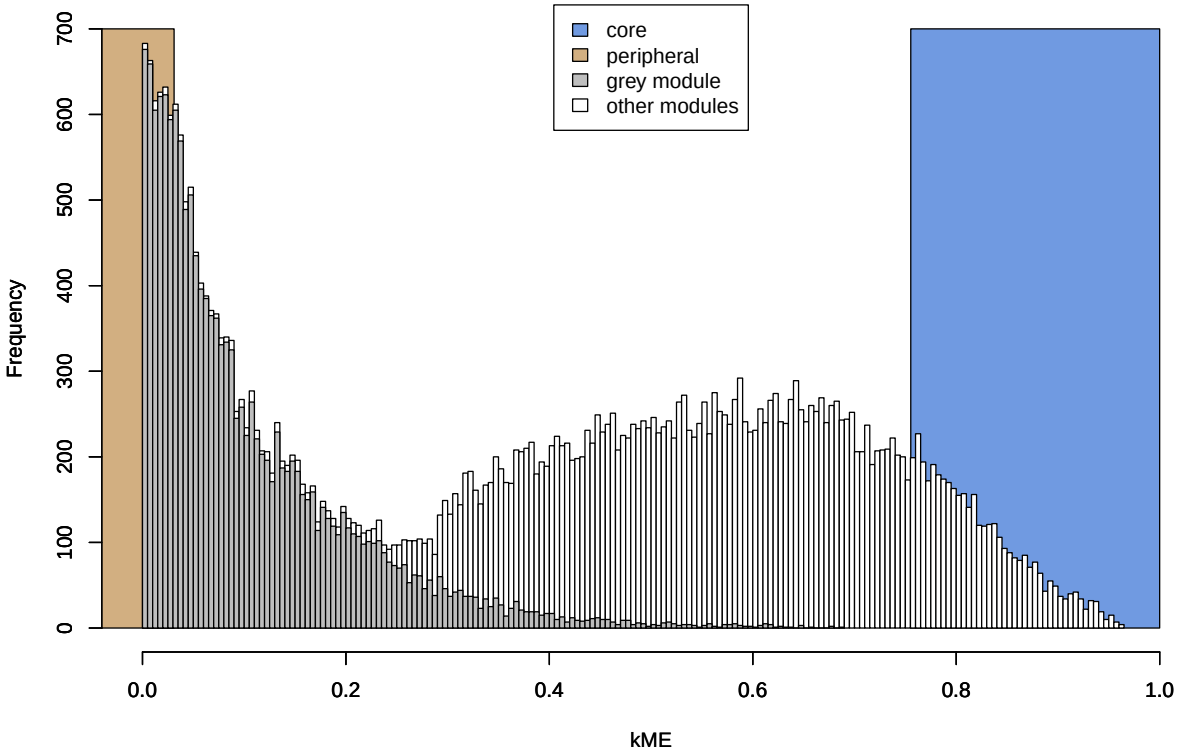
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S5: Relationship between Spearman's correlations between module-trait (y-axis) and gene significance-kME (x-axis).



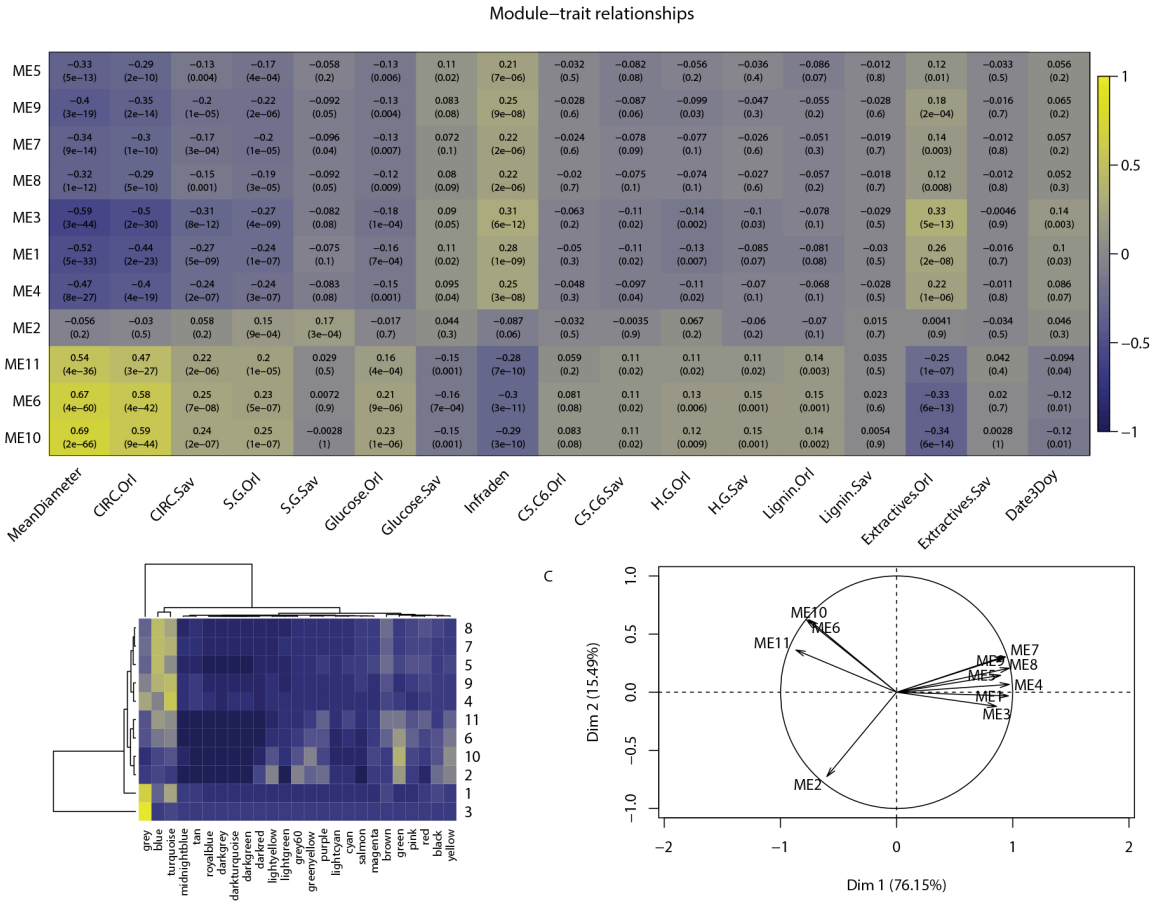
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S6: Histogram of the centrality scores. Core and peripheral sets are represented respectively by the blue and brown shading behind the bars. Random sets are distributed across the histogram and do not appear on this figure. Distribution of genes clustered in the grey module is represented by the grey bars, white bars are for other genes.



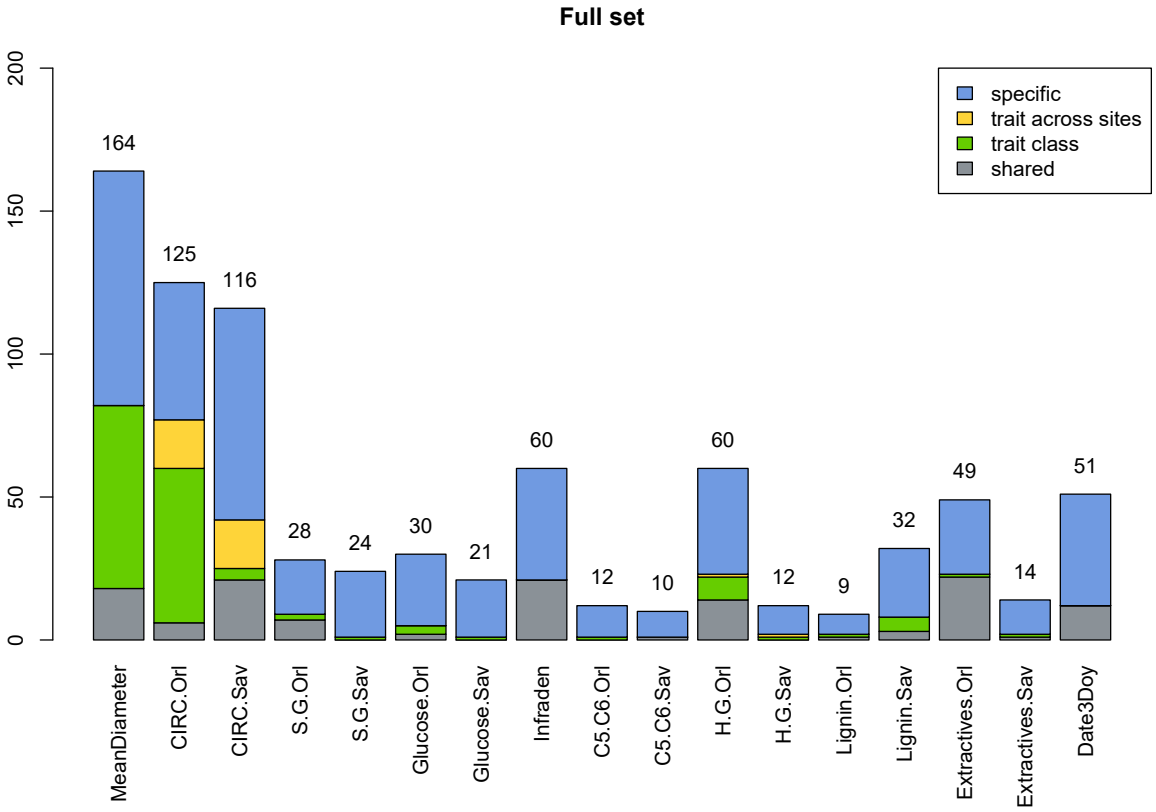
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S7: Gene expression k-means clustering (A) Correlation between eigengenes of modules identified by k-means clustering, on a light yellow (positive) to dark blue (negative) scale. P-values are indicated on the second line of each square. (B) Heatmap representing the concordance between WGCNA (abscissa) and k-means (ordinate) clusterings. (C) Principal component analysis graph of the k-means clustering.



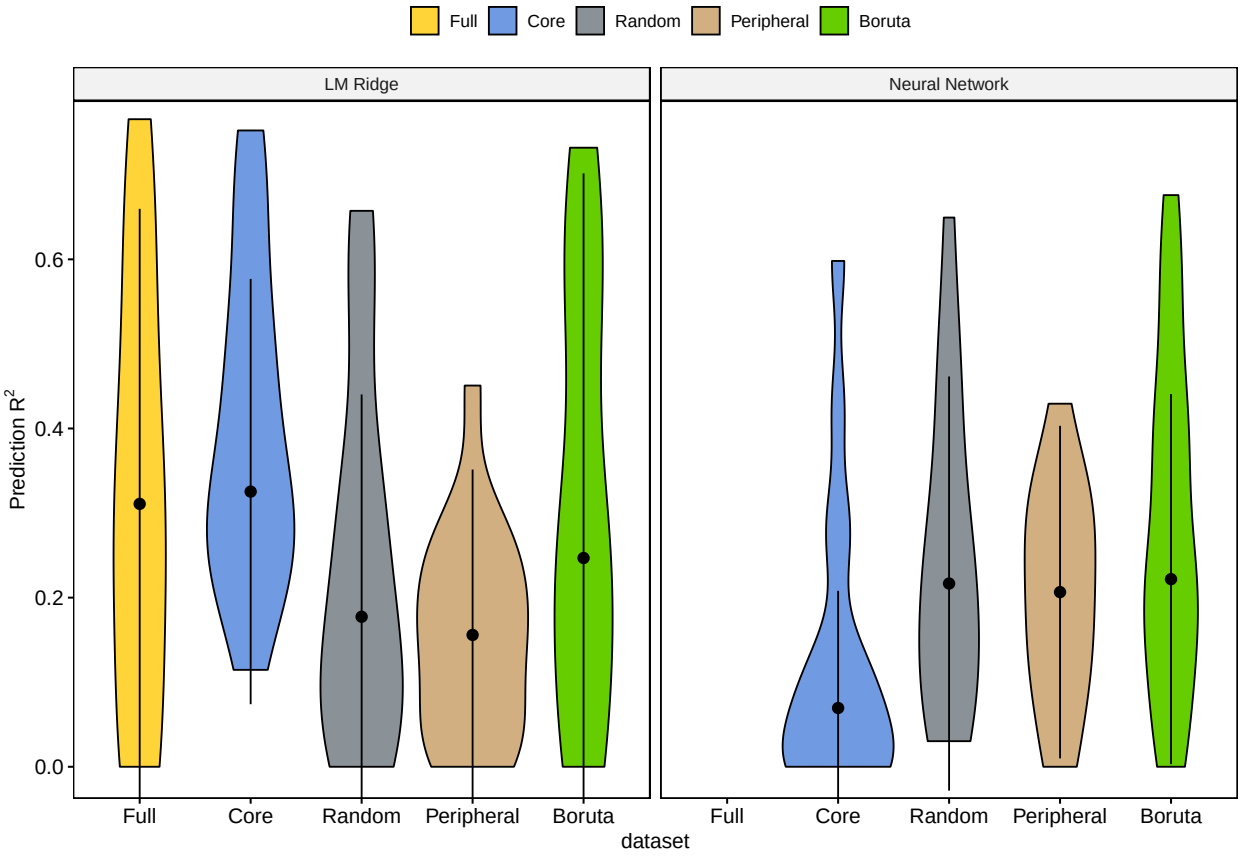
bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S8: Number and redundancy of genes whose expression is important to predict phenotypes, selected by Boruta on the training subpart of the full gene set. Genes are colored according to how shared they are: specific to a phenotype in blue, shared between sites for a given trait (Orléans and Savigliano) in yellow, shared among trait families (growth, phenology, physical, chemical) in green and among all traits in grey.



bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S9: Violin plots of prediction R^2 across all phenotypes, split by model and gene sets. The black dot represents the median, the black line above and below it represents the interquartile space.



bioRxiv preprint doi: https://doi.org/10.1101/523365; this version posted April 4, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

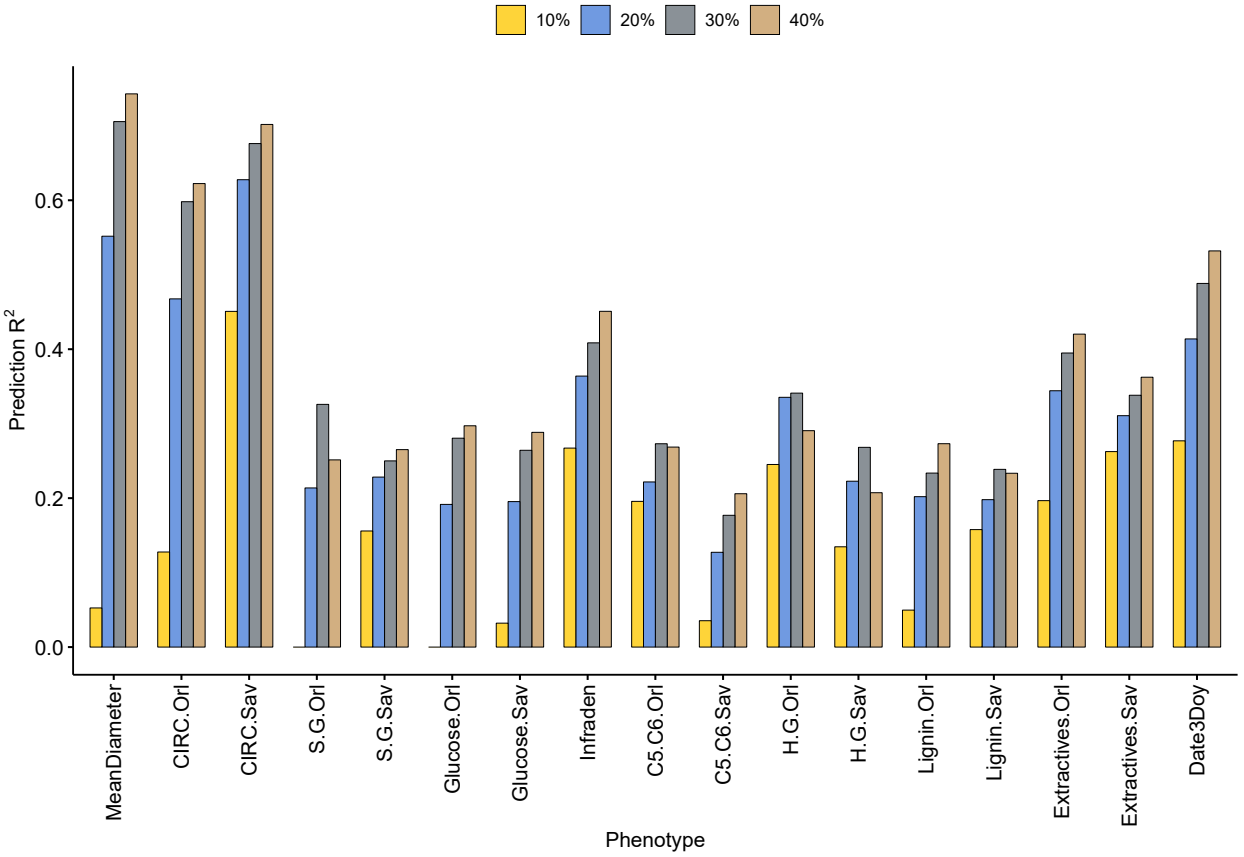



Figure S10: Predictions scores on test sets (R^2 on the y-axis) for the LM Ridge algorithm for each phenotypic trait (on the x-axis). The color of each bar represents the size of the peripheral gene set that has been used for the predictions (in percent of the full set).