Combining networks and metabolic genome-wide association studies to associate genes to metabolic pathways.

Background:

In the past 10 years, the forward genetic approach of genome-wide association studies (GWAS) has led to many breakthroughs in associating between genes and a large variety of traits. Traits, ranging from complex ones, combining multiple genes such as schizophrenia [1] to single-gene derivate intermediate phenotype interactions such as metabolites [2].

Today one of the major challenges facing GWASs is the lack of context. This lack of context is due to difficulty connecting the associated genes to a biological pathway, and segregating genes that are directly related to the phenotype i.e. 'driver' and those that are not 'passengers' [3]. A possible solution to this problem is by using a network-based approach, to cluster together known genes and metabolites from previous studies to newly annotated loci. For example, the general flavonoid/flavonol subgroup and anthocyanin subgroup were classified into different clusters by gene co-expression network analysis using all datasets in ATTED-II [4, 5]. Another example of a comparison to a known database is by associating the mGWAS results to a previous time-course stress study to construct metabolite–transcript correlation networks [6]. However, it is also possible to re-shape a GWAS analysis within an experiment to get information straight out of the GWAS analysis, as shown by the CONDOR method. In this method, the researchers combined expression quantitative trait loci (eQTL) with GWAS to create a bipartite network that represents both *cis*- and *trans*-acting SNPs and the genes with which they are associated. Then they used the modular structure of that graph to place SNPs into a functional context [7].

Although previously, GWAS and an intermediate phenotype was convert into a bipartite network, the knowledge of converting metabolomics base-GWAS into a bipartite network remains unexplored. In this work, a metabolic GWAS (mGWAS) comparison between two conditions, stress of heat and darkness and control were established. By creating a community-based bipartite network from the communities within the two bipartite networks, it was possible to unravel unique characteristics that otherwise, looking just at the two networks separately, were undetectable. One of these unique

characteristics was identified as the overall centrality of the different communities, which can highlight metabolic connections that in other ways are overlooked.

Research question:

Is it possible to associate genes to a metabolic pathway by measuring the changes in GWAS association distribution in the community structure of the bipartite network?

Methodology:

Creating the community-based bipartite network:

For GWAS from control and heat and darkness treated plants, a matrix of metabolites vs gens was created. The rows of the matrix include annotated glycerolipids metabolites that appeared in the GWAS experiment. The columns represent the genes that include SNPs with a logarithmic scale of odds (LOD) score that is higher than five (p-value < 0.05e-5). The values that were used to create the matrix were the maximum association between the designated metabolite and the associated gene. To create the networks and find the modularity within every network the package “bipartite” [8] was used. Afterwards, a new matrix was created with the module list of every bipartite network as the axis values. The values within this new matrix were the sum of co-appearing metabolite or gene between the two networks.

Measuring the bottleneck effect in the network using betweenness centrality:

To focus on the more central modules for metabolites or genes, a betweenness centrality measure was calculated on a unipartite format of the matrix using the “igraph” package [9]. To better understand if the bottleneck represents a biological press, the metabolites in the communities were classified into their annotated glycolipids classes and their amount of saturation. In addition, the genes that appeared in these communities were further analyzed in a time-course expression paradigm [10].

Gene expression of genes per community:

The genes within every community with a high betweenness score were compared using a two-way t-test, those that appear significantly different between the stress of heat and darkness and control were correlated using their regression linear curve over time. To clean out the genes that do not show a distinct trend over time, the genes were separated into three conditions: Increased excretion over time, where the slop of the liner curve was higher than 0.05; Decrees excretion over time where the

slop of the liner curve was lower than -0.05; Time and stagnant excretion where the curve was between 0.05 and -0.05. A permutation test was calculated between the sum correlations of increased excretion and decreed excretion, in comparison to 10000 random gene correlations.

Results:

Combination of mGWAS into community based bipartite network reveals bottleneck structure under stress of heat and darkness.

Genes associated by metabolic GWAS analysis are lacking context due to big gaps between genes and their endpoint metabolic phenotype. Thus, genes code affect different parts of the metabolome in a direct or indirect manner. Consequently, re-shaping the results of mGWAS unites in different genes associating patterns into functional context. To establish a bipartite network for every condition, all GWAS results for every glycerolipid were measured and the maximum value per gene related SNP was selected (fig. 1a). Then a weighted bipartite network was created using the bipartite package. Every network is based on the maximum association between the list of glycerolipids and their associated genes (fig. 1b-c). Communities were calculated per network, and a new bipartite network was established. The high level represented communities in the heat and dark condition, and the low level represented the communities of the control (fig 1d). When separating between metabolites and genes, it becomes visible that there are two metabolite-related bottlenecks in community HD34 and community HD15 (fig 1e). These results suggest that combining bipartite networks from different conditions into a community baste bipartite network unravels a bottleneck effect.

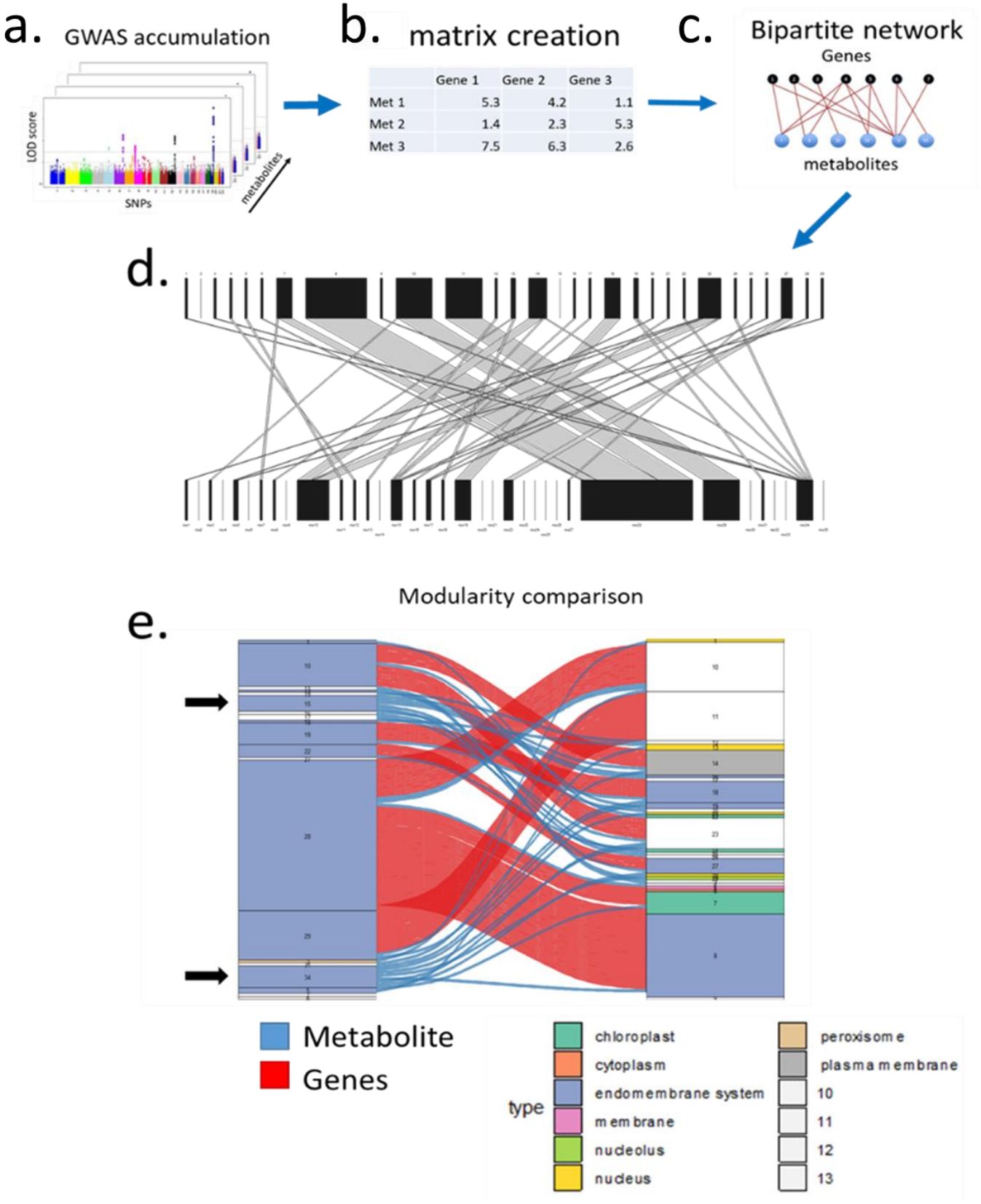
Annotation of metabolites in the bottlenecks emphasizes preference toward triacylglycerols and saturation levels.

To give an accurate measure of the bottleneck effect, betweenness centrality was measured on the unipartite community-based network (fig 2a). From the metabolic distribution of the GWAS loci created by the linkage disequilibrium plots, it appears that there is spasticity of the bottlenecks towards high saturated triacylglycerolipids for community 34 and low saturated triacylglycerolipids in community 15. (fig 2b). Therefore, the network bottleneck effect is metabolite driven strengthening the possibility of it being a biological process bottleneck.

Time course experiment adds a layer of biological context

Correlation between the genes within the communities suggests a 3-way pattern: increasing, decreasing or stagnant expression over time. The correlation between the increase and decrease group is significantly higher than a random set of genes (fig 3a). When comparing the gene expressions, heat and darkness treatment have an orderly fashion in comparison to the control (fig 3b). The slops of the linear regression of every gene in the communities have a higher absolute value than the ones in the control (fig 3c). This result might point out a biological pattern to the genes that are specific to the heat and darkness, highlighted by their communities.

Community bipartite network of multiple mGWAS.



Diagrams illustrating the community bipartite between two treatments and how it was built. For each condition, Glycerol lipid GWAS mapping where combined to create a bipartite network between genes and metabolites.

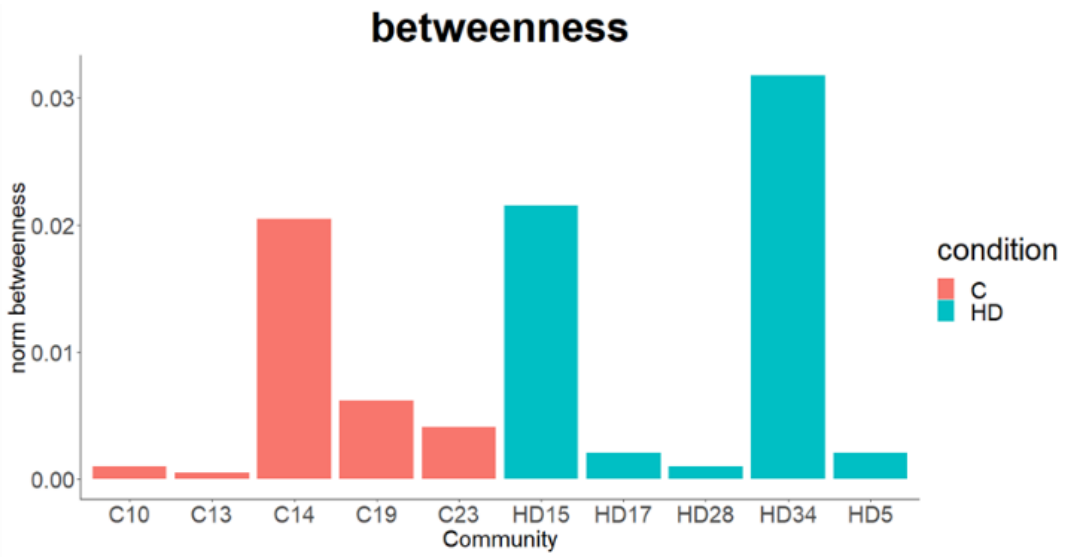
a, An illustration of the dimensions (LOD score, SNPs, and different metabolites) that are summed together for the creation of the matrixes. **b**, An example of a matrix for one condition, the column represents different genes that were significantly associated with glycerolipids in the GWAS analysis. Rows represent the metabolites that were associated with the different genes by the closest SNP to the different genes in the GWAS. **c**, Example of a bipartite network, verdicts represent the genes (upper side) and metabolites on the (lower side). edges represent associations between metabolites and genes.

d, Weighted community bipartite network of control community (upper verdicts) and Heat and darkness (lower verdicts). Edges represent genes or metabolites that appear in different conditions.

e, Weighted community bipartite network of communities that have at least one shared gene or metabolite. The red edges represent single genes that appear in both communities, and the blue edges represent a single metabolite. The colors of the verdict represent the most abundant GO term that appear in TAIR 9 database[11] The black arrow on the left side represents bottlenecks of metabolites that concentrate in specific communities under heat and darkness.

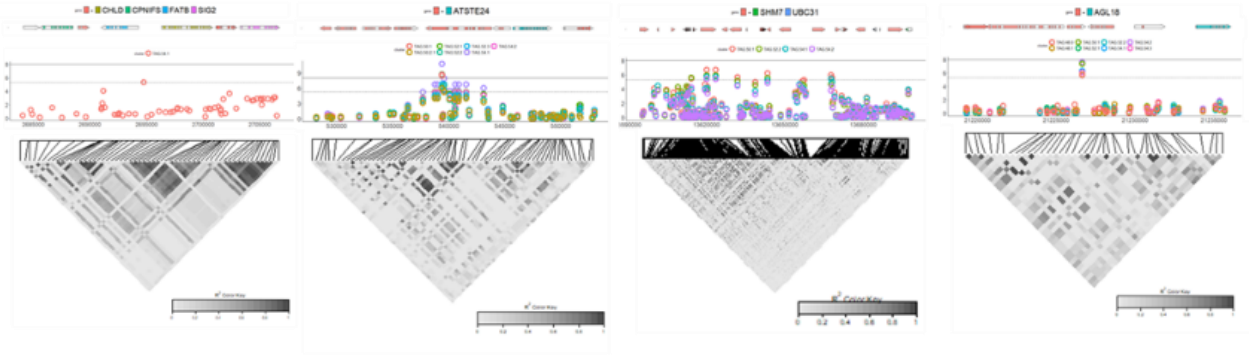
Validation and community metabolic distributions examples.

a.

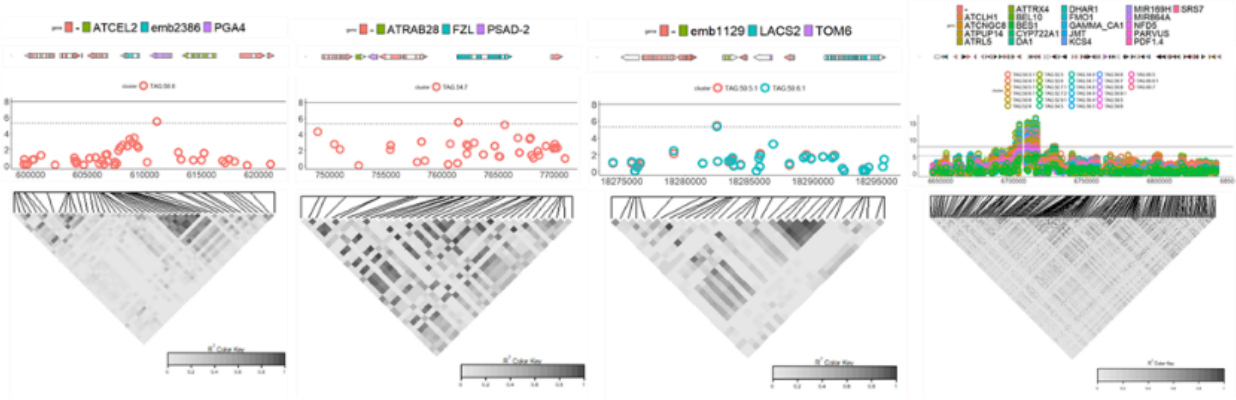


b.

15



34



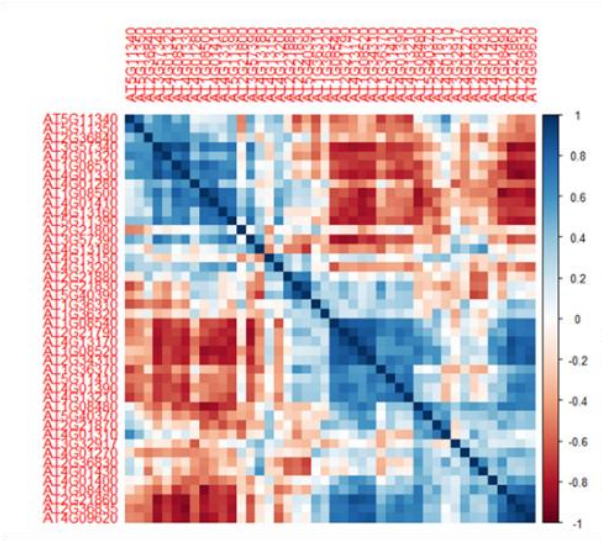
The bottleneck community shows a high specialty towards a specific metabolite type under heat and darkness. **a**, betweenness centrality measure of the metabolic distribution in the network points out that there are 2 distinct bottlenecks in the heat and darkness condition in communities 34 and 15. **b**, Different GWAS loci that construct the bottlenecks and their linkage disequilibrium plots highlight the specificity of the bottlenecks toward specific metabolic types. Most of the high unsaturated triacylglycerols are located inside community 34, whereas most saturated triacylglycerol is located inside community 15.

Time cores experiment highlights gene correlation within the communities.

15

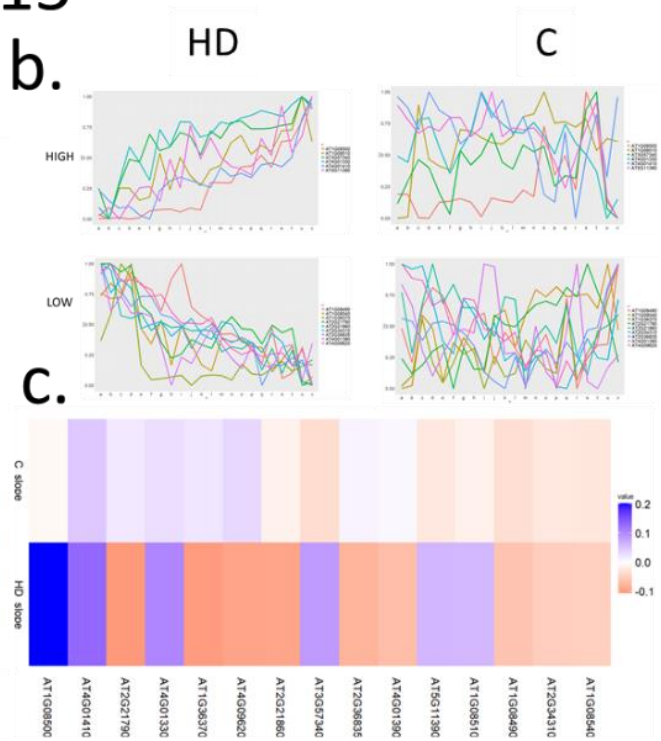
a.

44/61 change sig between C and HD



Sig level of cleaned slops(N=15,e= 10000) 0.00029

b.

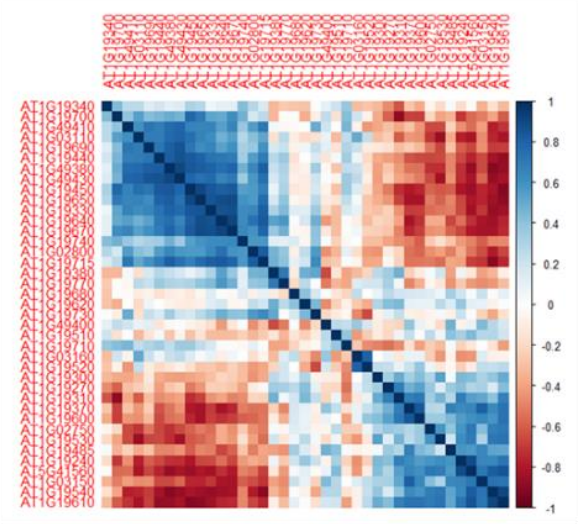


c.

34

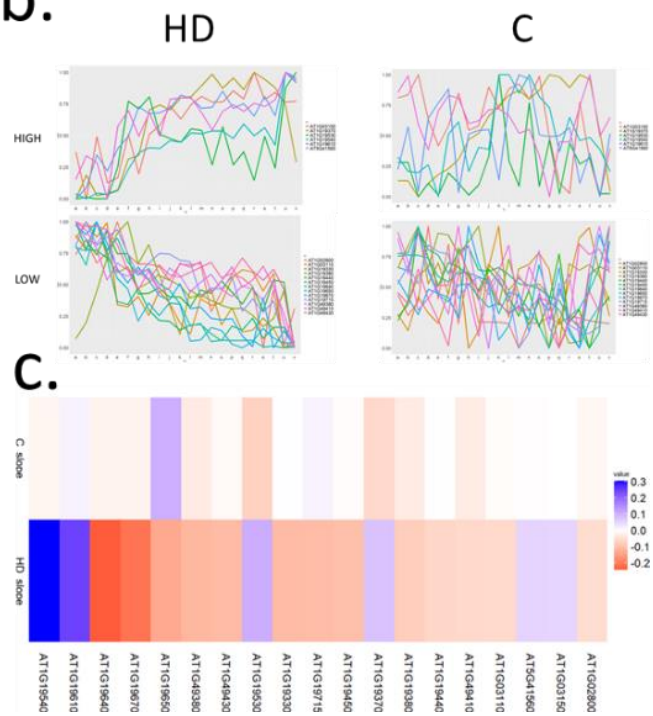
a.

39/50 change sig between C and HD



Sig level of cleaned slops(N19,e=10000) 9.999e-05

b.



c.

Time course experiment of gene excretion for the two conditions adds' a layer of validation to the genes that appear in the two bottlenecks. **a**, a correlation plot of the genes that appear significantly different between the conditions. The genes can be separated into three types in comparison to the control (increase excretion over time, decrees excretion over time, and still excretion over time). Permutation tests of the combined groups, both increased and decreased expression, have shown a significant increase in their correlation in comparison to 10000 random correlations of the same amount of genes. **b**, Normalized gene expression of the time course experiment, on the left side are the up and down expressed genes of the heat and darkness condition, and on the right side are the two conditions for the control. **c**, linear regression slope of every gene of increase and decrees excretion in the two conditions.

Discussion:

The use of GWAS to unravel genes regulating metabolomic biosynthesis was applied widely in the plant kingdom [2, 3, 12]. However, the reconstructing of a multiple metabolic GWAS into a bipartite network remains unexplored. My theoretical results show that combining the mGWAS into a bipartite community-based network unravels bottleneck effect under stress of heat and darkness, in glycerolipid degradation. In particular, the network has revealed that there are two distinct bottlenecks, one in community HD34 and the other in community HD15. Within those bottlenecks, the metabolites were high-saturated triacylglycerols (TAGs) and low acylglycerols respectively, suggesting a separate biological press of degrading high and low saturated TAGs. One of the genes that excretion decreases significantly over time in community HD34 is KCS4. This gene was identified in a previous study as associated with several polyunsaturated TAGs specifically under abiotic stress conditions. Furthermore, it was located in the ER, and acts as a branch point in the fate of fatty acids, directing the saturated ones to the very-long-chain fatty-acid pathway (VLCFA) [13]. Identifying the genes in community HD15 that their excretion decreases over time, reveal that many of them are chloroplast related by their GO terms. For example Sigma factor 2 (SIG2), a phytochrome-regulated protein important for stoichiometric control of the expression of plastid- and nuclear-encoded genes that impact plastid development and plant growth [14]. These findings suggest that bottleneck HD15 is located in a different cellular compartment than those in community

34. This result highlights the possibility of two distinct pathways of degrading TAGs by their saturation level.

In conclusion, this work suggests that by using a community-based GWAS for a comparison between stress of heat and darkness and control populations, it is possible to segregate different degradation pathways.

Word count: 1931

Bibliography:

[1] Trubetskoy, V., Pardiñas, A. F., Qi, T., Panagiotaropoulou, G., Awasthi, S., Bigdeli, T. B., ... & Lazzeroni, L. C. (2022). Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature*, *604*(7906), 502-508.

[2] Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., ... & Luo, J. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nature genetics*, *46*(7), 714-721.

[3] Califano, A., Butte, A. J., Friend, S., Ideker, T., & Schadt, E. (2012). Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature genetics*, *44*(8), 841-847.

[4] Yonekura-Sakakibara K., Tohge T., Niida R. & Saito K. (2007) Identification of a flavonol 7-O-rhamnosyltransferase gene determining flavonoid pattern in Arabidopsis by transcriptome coexpression analysis and reverse genetics. Journal of Biological Chemistry 282, 14932–14941.

[5] Yonekura-Sakakibara K., Tohge T., Matsuda F., Nakabayashi R., Takayama H., Niida R., Watanabe-Takahashi A., Inoue E. & Saito K. (2008)

[6] Wu, S., Alseekh, S., Cuadros-Inostroza, Á., Fusari, C. M., Mutwil, M., Kooke, R., ... & Brotman, Y. (2016). The combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. *PLoS genetics*, *12*(10), e1006363.

[7] Platig, J., Castaldi, P. J., DeMeo, D., & Quackenbush, J. (2016). Bipartite community structure of eQTLs. *PLoS computational biology*, *12*(9), e1005033.

[8] Dormann, C. F., Gruber, B., & Fründ, J. (2008). Introducing the bipartite package: analysing ecological networks. interaction, 1(0.2413793).

[9] Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, complex systems, 1695(5), 1-9.

[10] Wu, S., Alseekh, S., Cuadros-Inostroza, Á., Fusari, C. M., Mutwil, M., Kooke, R., ... & Brotman, Y. (2016). Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. *PLoS genetics*, *12*(10), e1006363.

[11] Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., ... & Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research, 40(D1), D1202-D1210.

[12] Medeiros, D. B., Brotman, Y., & Fernie, A. R. (2021). The utility of metabolomics as a tool to inform maize biology. *Plant Communications*, *2*(4), 100187.

[13] Luzarowska, U., Ruß, A. K., Joubès, J., Batsale, M., Szymański, J., Thirumalaikumar, V. P., ... & Brotman, Y. (2020). Hello darkness, my old friend: 3-Ketoacyl-Coenzyme A Synthase4 is a branch point in the regulation of triacylglycerol synthesis in Arabidopsis by re-channeling fatty acids under carbon starvation. BioRxiv.

[14] Oh, S., Strand, D. D., Kramer, D. M., Chen, J., & Montgomery, B. L. (2018). Transcriptome and phenotyping analyses support a role for chloroplast sigma factor 2 in red-light-dependent regulation of growth, stress, and photosynthesis. Plant direct, 2(2), e00043.