A traipse through plant synthetic biology

by

Kasey Markel

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in

Plant Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Patrick Shih, Chair
Professor Henrik Scheller
Professor James Nuñez

Summer 2024

A traipse through plant synthetic biology

By Kasey Markel

Abstract

A traipse through plant synthetic biology

by
Kasey Markel
Doctor of Philosophy in Plant Biology
University of California, Berkeley
Professor Patrick Shih, Chair

Plants have been the highest volume source of biomass and biomaterials for human use since the agricultural revolution, and even before that in most societies. Despite incredible advances in synthetic chemistry and microbial synthetic biology, plants remain the key source for calories and nutrients, and critical feedstocks for materials, fuel, and drugs. This dissertation consists of four body chapters. The first outlines the background of plant biotechnology and highlights a useful principle used in plant breeding but not well-known among plant biotechnologists. The second outlines how natural plant engineers outperform and might inspire the plant synthetic biologists of the future. The third chapter reports the first independent replication of an extremely exciting and surprising finding first reported halfway through my PhD. The fourth presents a large suite of genetic parts, specifically transcriptional repressors. Each chapter is intended to stand largely on its own, but I nonetheless believe the whole is greater than the sum of its parts, combining a conceptual advance, a look at nature's plant synthetic biologists, a replication project which lies at the heart of the scientific method, and the development of a novel suite of tools which lies at the heart of synthetic biology.

I would like to dedicate this dissertation to my parents
Karen and Curtis Markel, for giving me every opportunity

Table of Contents

Acknowledgements

It takes a village to raise a scientist, and I have had the good fortune to know many people without whom this would not have been possible. The following list will necessarily be incomplete, but I would like to thank...

My family, especially my parents Karen and Curtis, for their support of my eccentric interests ever since I was a toddler. You never tire of telling stories of a child wearing plants on his head, disappearing with no warning into tall grass, and climbing any tree that caught his fancy. Your support means everything.

My labmates in the Shih Lab, whose hands-on expertise has proved invaluable in all manner of experimental troubleshooting, and who have served as an excellent sounding board for all manner of scientific ideas.

The cohort of Plant Biology Graduate Group (PBGG) students at UC Davis with whom I began my PhD, who have taught me so much and supported each other through the difficult times that come with any significant undertaking.

The cohort of Plant and Microbial Biology (PMB) graduate students at UC Berkeley who welcomed me as one of their own when I transferred to Berkeley halfway through this degree.

The scientists and support staff at the Joint BioEnergy Institute, who have provided countless discussions and many fruitful collaborations, as well as a well-stocked and run institute in which to perform research.

Past mentors that have gotten me to this point, including Donna Widhalm, Jeffrey Prince, Barbara Whitlock, Antoine van Oijen, and Jim Haseloff. You have all shown me what it means to be a good scientist, and have gently corrected my many errors along the path.

Last but not least, my advisor Patrick Shih. You have been an excellent advisor for the last several years, and have directed me to work on an exciting set of projects. It’s been a wild ride, but I wouldn’t want it any other way.

Chapter 1: Plant synthetic biology, beginning of a great adventure

1.1 Preface

Plants have been the highest volume source of biomass and biomaterials for human use since the agricultural revolution, and even before that in most societies. Despite incredible advances in synthetic chemistry and microbial synthetic biology, plants remain the key source for calories and nutrients, and critical feedstocks for materials, fuel, and drugs.

1.2 Organization

This dissertation is primarily composed of four manuscripts, of which two are as of this writing published in peer-reviewed journals and two are available on BioRxiv and are undergoing review for peer-reviewed publication. The manuscripts each address a different topic in plant biotechnology, and as such I have opted to keep references, manuscript-level acknowledgements, and author lists to allow each manuscript to be presented in a relatively complete form. In similar fashion, the figures and supplemental figures are numbered by chapter, enabling preservation of and compatibility with the published version of the individual stories that comprise this dissertation.

The central connecting thread of my research interest has been improvement of the tools of plant synthetic biology, with the lofty goal of enabling plant biotechnology to improve human welfare around the globe. While on the surface these manuscripts address rather distinct topics, they are all linked in that grand vision.

In **Chapter 2,** we ask whether modern plant biotechnologists working on improving plants for biofuel production have things to learn from a concept widely used in classical plant breeding. As a review, this manuscript is long on ideas and short on data, though we do attempt to compile some evidence from existing literature to support the proposition.

Chapter 3 could easily be interpreted as a work of basic science focusing on a natural phenomenon remarked since the earliest naturalists. However, the motivation for that work was more biotechnology-driven than is easily seen from the final result - the moonshot proposition was to gain a sufficiently granular understanding of the remarkable plant engineering performed by wasps that human biotechnologists would be able to replicate some of their feats.

The final two manuscripts in this dissertation are not yet published in peer-reviewed publications, and primarily reflect work performed during my time at UC Berkeley rather than UC Davis where I began this PhD.

Chapter 4 began as an attempt to replicate a very surprising paper which claimed that a human RNA demethylase was capable of increasing yield in plants. This project was driven almost entirely by my own curiosity, which has always been strongest at the far reaches of science - my first single-author paper was also a replication effort. To my surprise and delight, the core claim that this human gene could increase growth and “yield” in plants replicated in my experiments. I believe this is an extremely promising line of research that deserves more attention than it is currently getting - to my knowledge there have been no further experiments published since that landmark paper, and my own manuscript has faced significant hurdles in the peer review process, caught between the Scylla of “this is a mere replication, insufficiently original to merit publication” and the Charybdis of “these results are rather unbelievable, why would a human gene boost growth in plants so much”.

Chapter 5 was the project that initially resulted in my admission to Shih lab, and has unfortunately nonetheless been the slowest to publish. It tells the story of the development of a large suite of transcriptional repressors in plants, and its inclusion here is the reason for the embargo on the dissertation for a long enough period to sort out patent applications.

Chapter 2: Defining and engineering bioenergy crop ideotypes

Including material from published work:
Markel, Kasey, Belcher, Michael, and Shih, Patrick "Defining and engineering bioenergy crop ideotypes." Current Opinion in Biotechnology 2020

2.0 Chapter preface

There is a surprisingly large disconnect between the worlds of plant breeding and plant biotechnology. As a result, useful concepts from one sometimes fail to become widespread in the other, even when they still have some relevance. In this review, we attempt to bring the classical breeding concept of an 'ideotype' to plant biotechnologists interested in improving bioenergy feedstock crops. This chapter provides much of the motivation for the more specific projects described in future chapters. In many ways, this is the core theme of the thesis, and each subsequent chapter approaches this theme at a different level of abstraction and a different point on the continuum between simple well-established plant biotechnology strategies and blue-sky cutting-edge ideas.

2.1 Abstract

Ideotypes are theoretical archetypes of crops which serve as a practical framework for plant breeders to critically evaluate what traits they should be targeting for specific applications. With advances in plant biotechnology and a growing urgency to adopt more sustainable practices across our economy, new uses for crops as bioenergy feedstocks may pivot our definition of an ideal crop that is engineered for biomass and bioenergy production, in contrast to food production. Although there is a plethora of specific applications that plant engineering efforts can contribute to, here we highlight recent advances in two broad areas of research: increasing available plant biomass and directly producing bioproducts of interest.

2.2 Introduction

Prior to our ability to transform plants, plant breeders were constrained to breeding and selecting from the morphological, physiological, and metabolic repertoire already preexisting in plant genomes. Such efforts initially were focused on breeding out deleterious traits or a narrow focus on yield. Fifty years ago, the concept of an ideotype was proposed as an alternative regime. The ideotype is an idealized form of a particular crop, which could then be a target to breed towards, rather than merely breeding away from deleterious traits [1]. This shift in mentality provided a much-needed framework to help set goals and target traits for plant breeding efforts. A useful ideotype must be "theoretically capable of greater production than the genotype it is to replace and of such design as to offer reasonable prospect that it can be bred from the material available." [1]. The discovery and development of plant transformation technologies such as Agrobacterium-mediated and biolistic transformation expanded the scope of possible ideotypes, as plant engineering efforts can now draw on a much larger effective pool of genetic material, expanding from interfertile germplasm to all sequenced and characterized genes from across the tree of life.

Feedstock crops are harvested primarily for biomass, which is then used as a substrate for downstream processes (e.g., bioconversion, fermentation, combustion, etc). Thus, it becomes useful to frame plant carbon partitioning in terms of biomass composition, and what production or deposition of small molecules or polymers would be present in feedstock ideotypes. Using new synthetic biology tools to redesign carbon flow in plants, one may alter and optimize the composition of biomass and bioproducts in a way that cannot be achieved through conventional breeding methods, ultimately improving the scalability and feasibility of renewable feedstock crops. Here, we focus on carbon allocation as a metabolic/physiological trait that may be modified to increase the utility and value of feedstock crops. Specifically, we focus on two aspects: 1) traits that may alter overall plant biomass and the usability of this biomass and 2) traits that may enhance the value of feedstock crops with the production of bioproducts, paying special attention to advances within the last two years.

2.3 Results and Discussion

2.3.1 **Engineered traits to enhance plant biomass**

The plant cell wall is a complex network of polymers and is one of the most effective carbon sequestering systems on the planet, with annual production of land plants estimated at 150 - 170 billion tons per year [2]. Cell walls represent a massive and largely untapped supply of C6 sugars in the form of cellulose, β-1,4-linked glucose. However, cell walls are naturally recalcitrant to degradation and fermentation, limiting their use [3]. Lignin is a main inhibitor of sugar release in woody crops and hemicellulose limits saccharification yields in monocot biomass crops [4]. Many engineering efforts have focused on decreasing lignin and improving fermentation characteristics.

We are only beginning to explore ways to modify the composition and deposition of plant cell wall components to improve their ability to serve as biomass feedstocks. One strategy for reducing lignin accumulation uses 3-dehydroshikimate dehydratase (QsuB) from Corynebacterium glutamicum, which converts a lignin precursor into protocatechuate. Transgenic expression of QsuB in *Arabidopsis thaliana* plastids reduced lignin accumulation and improved saccharification yield by 25-100% depending on treatment method [5]. Moreover, the C6/C5 sugar ratio of the biomass also affects saccrification yields, with higher ratios performing better. The most highly accumulated C5 sugar is xylose, but xylan synthesis mutants show dwarfism due to xylem vessel collapse. This phenotype has been rescued by returning xylan synthesis specifically to vessel tissue, leading to a 42% increase in saccharification yield compared to wild type [6]. Acetylated cell wall components are converted during fermentation to acetic acid, which inhibits fermentation. RNAi has been used to decrease expression of genes responsible for acetylation, nearly tripling saccharification yields [7]. Gene stacking has been used to generate engineered lines that contain multiple aforementioned traits [8*]. This demonstrates how modern bioengineering strategies can be used in tandem to modify the cell wall composition, a step towards engineering the optimum bioenergy crop ideotype. While ideotype specifics will vary by crop and intended application, in general an idealized biomass cell wall will have a high C6/C5 sugar ratio, low lignin concentration, and provide a favorable substrate for fermentation.

Beyond modifying the molecular composition of the cell wall, others have also focused on engineering upstream metabolic processes to increase rates of photosynthesis, carbon fixation, and biomass production. Plants often absorb more photons than they can use for photosynthesis, leading to non-photochemical quenching (NPQ) that dissipates excess energy as heat but does not contribute to biomass. Mutation of light harvesting complex components results resulted in a 25% biomass increase in Nicotiana tabacum under field conditions [9**]. It is also possible to modulate the NPQ process to shift more quickly from a heat-producing to a photosynthetic state, restoring energy capture via production of NADPH and ATP. Engineerined N. tabacum overexpressing the genes coordinating NPQ relaxation showed increases of ~15% in plant height, leaf area, and total biomass accumulation in field conditions [10]. These are promising results, as most plants use similar mechanisms making this technology applicable to bioenergy crops dependent on the maximum accumulation of lignocellulosic biomass.

Another key process that limits the theoretical maximum for biomass accumulation is photorespiration. The primary cost of photorespiration stems from the process plants use to “recycle” the unintended product formed via the oxygenase activity of RuBiSCO, leading to loss of both carbon and nitrogen. An alternative photorespiratory bypass based on the 3-hydroxypropionate bicycle was successfully engineered into cyanobacteria by expressing six heterologous genes from Chloroflexus aurantiacus. This bypass not only limits losses from photorespiration, it also fixes additional carbon and can supplement the Calvin-Benson cycle [11]. Other photorespiratory bypasses have been demonstrated to work in planta yielding more than a 25% increase in biomass in field trials [12*] . Thus, the ability to modify both the rate of carbon fixation and the fate of carbon deposition in the form of various cell wall polymers have been shown to be complementary processes for increasing the overall plant biomass of future feedstock crops.

2.3.2 **Engineered bioproducts to increase feedstock value**

Lignocellulosic bioproduction offers a much larger potential supply of biomass than food based fuels such as corn-ethanol, and also reduces the conflict between food and fuels, materials, and other products which may be produced from biomass crops. However, lignocellulosic biofuels have been slow to achieve commercial viability, in part because of low fuel prices and the chemical recalcitrance of lignocellulosic matter [13,14*]. A promising strategy to make lignocellulosic biofuels economically competitive is the co-production of higher value products directly in feedstock crops, which can be separated from the bulk carbon fuel source during processing [15*]. This can be achieved in two ways: either feedstocks for lignocellulosic biofuels can be modified so as to produce a higher value side product, or lignocellulosic biofuel can be produced from side products of other agricultural processes, such as corn stover or forestry waste. The former is amenable to feedstock bioengineering efforts to optimize for biofuel purposes, and will be discussed here. The titer required to improve feedstock value and the market size tend to correlate to each other, and inversely to the value per unit of the product, as shown in Figure 1.

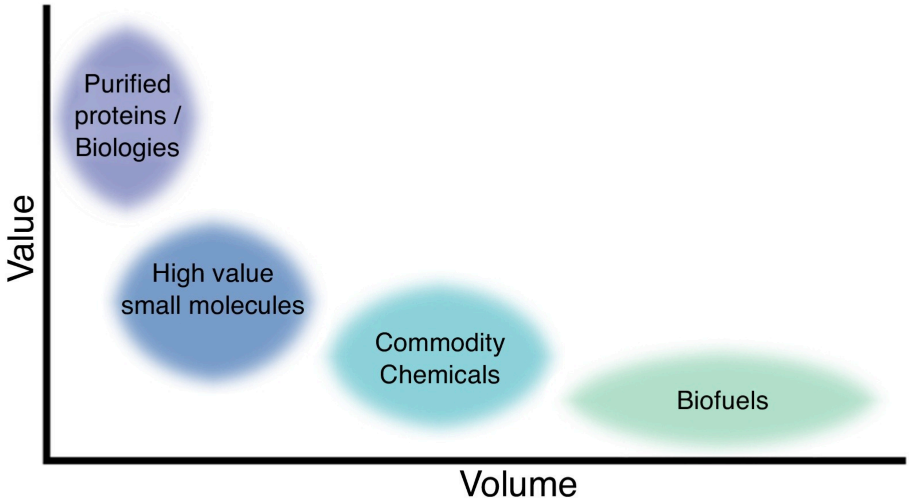


Figure 1: Tradeoff between value and volume. Cartoon diagram of inverse correlation between product volume and value per unit. Adapted from Ref. [29].

To improve the economics of biofuel production, engineered bioproducts must sell for more than the marginal cost of their purification, have a relatively large market, and not have a substantially detrimental effect on plant growth when produced in commercially relevant titers. The highest volume, lowest price added value product consists of biofuels of higher value than ethanol, which for most biofuel crops remains the base product. Higher value fuel products include lipids for biodiesel and jet fuel. Biodiesel-grade lipids have recently been produced in engineered sorghum that accumulates 8% dry weight oil in leaves in the form of lipid droplets [16**]. These droplets can be extracted using simple, cheap techniques during the standard processing pipeline for lignocellulosic biofuels, minimizing additional purification costs. Jet fuel is also a high volume product with an annual market size of 290 billion liters in 2015, with prices usually ranging around $1 per liter. There is no practical alternative available for liquid aviation fuels [17*], which account for a small but rapidly growing fraction of total anthropogenic greenhouse gas emissions - currently 2.3%, and growing at approximately 6% per year [18]. Jet fuels have been produced from the oilseed crop camelina, and efforts are underway to increase jet fuel yield [19]. Another promising high-volume side product is 1,5-pentanediol, a commodity chemical used in polyester and polyurethane production. The present market value is around $6000/ton, with a market size of 18 million USD [15*].

Using plants as a production chassis for high value low volume products has received substantial attention in recent years, with several analyses suggesting plants may allow for cheaper production for edible vaccines, bulk enzymes, and monoclonal antibodies than alternative systems [20]. These high value products split into two major classes: high value small molecules and proteins. The anti-malarial drug Artemisinin is a high value small molecule that has received particular interest in recent years ever since its first biosynthesis in a yeast system [21]. However, despite high humanitarian value, Artemisinin has not been economically valuable enough to allow for sustained production in yeast systems, leading efforts to use plants as a lower cost production method [22]. Plants have been the primary production platform for high value small molecules for millennia - medicines, spices, and drugs have mostly been sourced from plant hosts. In modern times, the large scale production of cannabis and opium poppies attest to the scalability and cost-efficacy of in planta small molecule production, even though opiates [23] and cannabinoids [24*] have been produced through engineered microbes. Despite advances in microbial engineering and synthetic chemistry, plants remain the production platforms of choice for these high value small molecules, demonstrating the low cost and high scale that can be achieved with plant systems.

A variety of lignocellulosic biomass crops are currently being researched, and each has different strengths in terms of growing conditions and potential output. For example, switchgrass grows with very few inputs on marginal land, whereas sorghum has the potential to serve both as animal feed and lignocellulosic biofuel feedstock. Rather than a single ideotype for all biomass crops, different crops may be more amenable hosts to particular applications. Complex metabolic pathways to produce high value small molecules have been successfully implemented in model plant species, and some biomass crops seem particularly amenable to metabolic engineering for high value small molecule production, as shown in Figure 2. Ultimately, engineered feedstock crops that produce co-products may help offset costs associated with a future plant-based bioeconomy that will have to compete with petrochemicals.

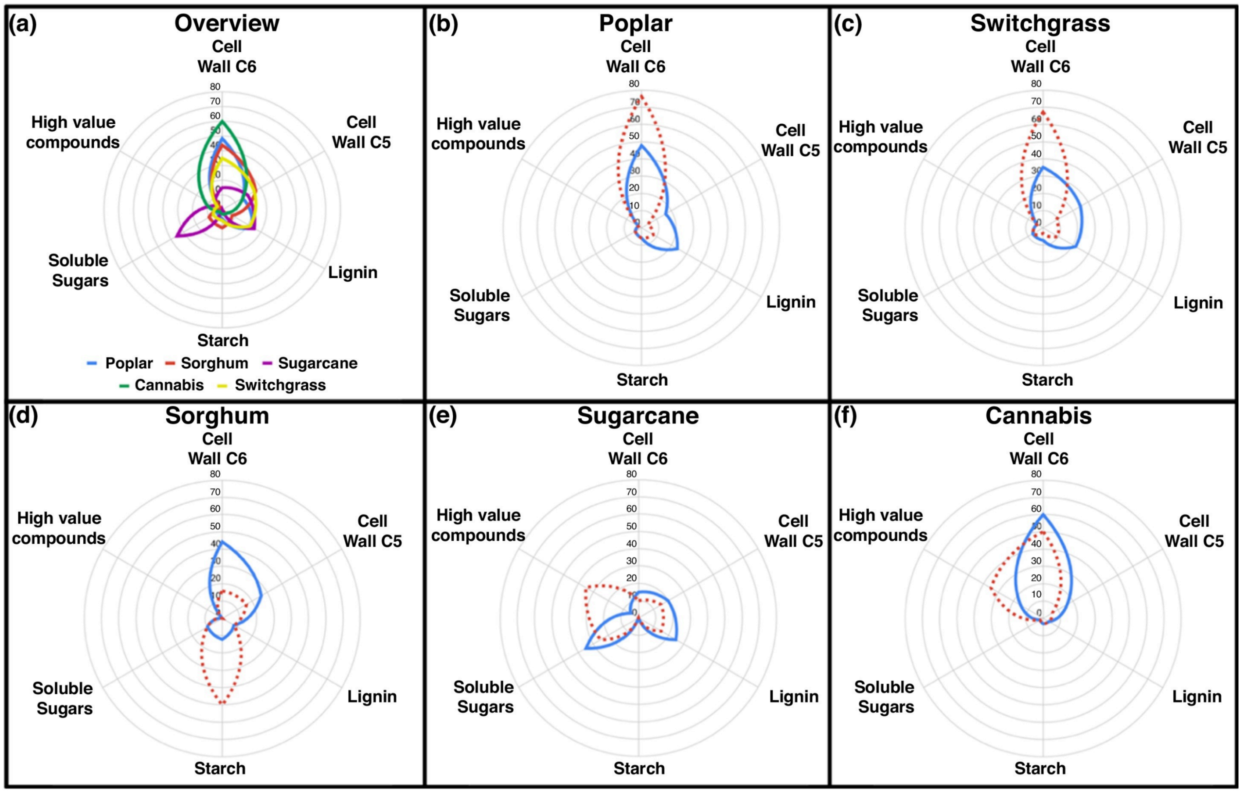
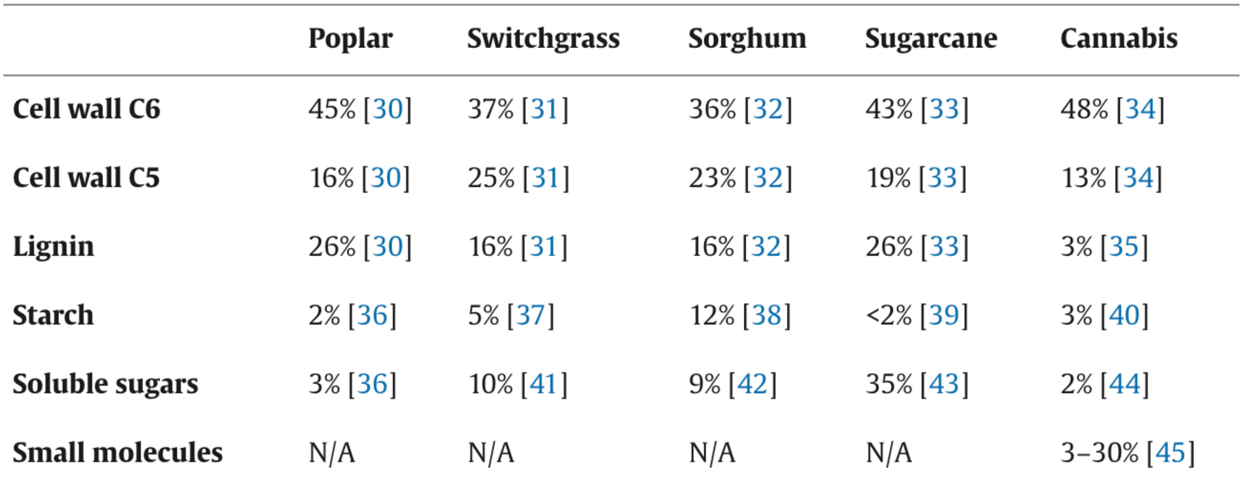


Figure 2. Carbon allocation in five bioenergy and bioproduct crops. Spider-plots demonstrating carbon allocation between six major categories: cell wall C6, cell wall C5, lignin, starch, soluble sugars, and high value small molecules. Panel A shows best estimates of current carbon allocation between these six categories. For this panel, blue indicates poplar, red indicates sorghum, purple indicates sugarcane, green indicates cannabis, and orange indicates switchgrass. Panels B–F show current allocation (solid blue line) and a hypothetical ideotype (dashed red line) for the crops poplar, switchgrass, sorghum, sugarcane, and cannabis, respectively. Concentric rings are percent dry weight, the outermost circle for all charts is 80% to allow for comparison between panels. Data are averages from literature sources, and should be taken as approximates. See Table 1 for data sources.

Table 1. Percentage of carbon allocation in potential feedstock crops listed in Figure 2. All data are presented as percentage dry weight and rounded to the nearest percent. When data are tissue-specific, averages or data from the most abundant tissue were used. All data are from untreated, wild-type, or control plants



	Poplar	Switchgrass	Sorghum	Sugarcane	Cannabis
Cell wall C6	45% [30]	37% [31]	36% [32]	43% [33]	48% [34]
Cell wall C5	16% [30]	25% [31]	23% [32]	19% [33]	13% [34]
Lignin	26% [30]	16% [31]	16% [32]	26% [33]	3% [35]
Starch	2% [36]	5% [37]	12% [38]	<2% [39]	3% [40]
Soluble sugars	3% [36]	10% [41]	9% [42]	35% [43]	2% [44]
Small molecules	N/A	N/A	N/A	N/A	3–30% [45]

2.4 Conclusion

Dedicated crops have been used in first-generation food-to-ethanol production for over 100 years [25], and in the United States annual production has increased 10-fold since 1990 [26]. Ethanol accounts for over 90% of all biofuel produced in the United States, nearly all of which is derived from dedicated fields of corn, consuming 38% of corn production [26]. The production of biofuel products from food crops causes competition between food and fuel, raising the price of staple foodstuffs [27,28*]. Lignocellulosic "second generation" biofuels substantially reduce this problem by either growing on marginal land where food crops are not viable, or by production from agricultural residues rather than diverting a food crop into the biofuel pathway.

Biofuels have sometimes been presented as an environmentally friendly and low-carbon alternative to fossil fuels, but current implementations have failed to deliver on this promise. Biofuels grown from established agricultural fields generally achieve GHG emission reductions of 20-80% compared to fossil fuels [29]. However, land use change associated with the conversion of natural land to biofuel production leads to a "carbon debt" that takes decades to centuries to pay back, negating any GHG savings [30]. Furthermore, conversion of natural land to biofuel production is a major driver of rainforest loss [31**]. Growing biofuel feedstock on marginal lands or producing biofuel as one of multiple products are the two main strategies to reduce this tradeoff. Here we consider re-designing biofuel feedstock crops to reduce cell wall recalcitrance, increase biomass, and generate additional products to add value and improve resource use efficiency.

Modern biotechnology has expanded the possibilities of crop ideotypes by allowing for plant phenotypes not attainable through classical breeding. Petrochemical fuels have been instrumental for global industrialization, and their use remains indispensable at the present. However, climate considerations as well the practical limitations inherent in using a finite resource call for the development of alternative sources of liquid fuel and materials. Plant biomass is the most viable means of production sufficiently scalable to take the place of petrochemicals in the economy of the future, and ideotype breeding serves as a useful paradigm for the design and improvement of biomass feedstock crops.

2.5 Acknowledgements

K.M. was funded by start-up funding provided by the University of California, Davis. M.S.B. was supported by a NSF Graduate Research Fellowship. P.M.S. was supported by the Department of Energy (DE-AC02-05CH11231) and start-up funding provided by the University of California, Davis.

2.6 References

1. Donald CM: **The breeding of crop ideotypes**. *Euphytica* 1968, **17**:385–403.

2. Pauly M, Keegstra K: **Cell-wall carbohydrates and their modification as a resource for biofuels**. *Plant J* 2008, **54**:559–568.

3. Himmel ME, Ding S-Y, Johnson DK, Adney WS, Nimlos MR, Brady JW, Foust TD: **Biomass recalcitrance: engineering plants and enzymes for biofuels production**. *Science* 2007, **315**:804–807.

4. DeMartini JD, Pattathil S, Miller JS, Li H, Hahn MG, Wyman CE: **Investigating plant cell wall components that affect biomass recalcitrance in poplar and switchgrass**. *Energy & Environmental Science* 2013, **6**:898.

5. Eudes A, Sathitsuksanoh N, Baidoo EEK, George A, Liang Y, Yang F, Singh S, Keasling JD, Simmons BA, Loqué D: **Expression of a bacterial 3-dehydroshikimate dehydratase reduces lignin content and improves biomass saccharification efficiency**. *Plant Biotechnol J* 2015, **13**:1241–1250.

6. Petersen PD, Lau J, Ebert B, Yang F, Verhertbruggen Y, Kim JS, Varanasi P, Suttangkakul A, Auer M, Loqué D, et al.: **Engineering of plants with improved properties as biofuels feedstocks by vessel-specific complementation of xylan biosynthesis mutants**. *Biotechnol Biofuels* 2012, **5**:84.

7. Scheller HV: **Method of Reducing Acetylation in Plants to Improve Biofuel Production**. *US Patent* 2012,

8*. Aznar A, Chalvin C, Shih PM, Maimann M, Ebert B, Birdseye DS, Loqué D, Scheller HV: **Gene stacking of multiple traits for high yield of fermentable sugars in plant biomass**. *Biotechnol Biofuels* 2018, **11**:2. Multiple technologies were used to stack and engineer three novel traits into a single plant system that increase saccharification yields and the fermentability of the resulting feedstock. This technique can be further expanded as new technology is developed and tested, eventually producing biofuel crops with a multitude of engineered traits.

9**. Kirst H, Gabilly ST, Niyogi KK, Lemaux PG, Melis A: **Photosynthetic antenna engineering to improve crop yields**. *Planta* 2017, **245**:1009–1020. Biomass increase in dense canopy conditions was achieved by diminishing the capacity of crops in a monoculture to capture photon energy. This technology could be applied to any crop utilizing light-harvesting antenna complexes via gene knockout, and due to current changes in GMO classification would circumvent costly regulatory measures.

10. Kromdijk J, Głowacka K, Leonelli L, Gabilly ST, Iwai M, Niyogi KK, Long SP: **Improving photosynthesis and crop productivity by accelerating recovery from photoprotection**. *Science* 2016, **354**:857–861.

11. Shih PM, Zarzycki J, Niyogi KK, Kerfeld CA: **Introduction of a Synthetic CO2-fixing Photorespiratory Bypass into a Cyanobacterium**. *J Biol Chem* 2014, **289**:9493–9500.

12*. South PF, Cavanagh AP, Liu HW, Ort DR: **Synthetic glycolate metabolism pathways stimulate crop growth and productivity in the field**. *Science* 2019, **363**. Many variations on an engineered photorespiratory bypass were generated and tested extensively under greenhouse and field conditions. Transgenic plants were scored by increases in biomass and photosynthetic light-use efficiency.

13. Holwerda EK, Worthen RS, Kothari N, Lasky RC, Davison BH, Fu C, Wang Z-Y, Dixon RA, Biswal AK, Mohnen D, et al.: **Multiple levers for overcoming the recalcitrance of lignocellulosic biomass**. *Biotechnol Biofuels* 2019, **12**:15.

14*. Hassan SS, Williams GA, Jaiswal AK: **Moving towards the second generation of lignocellulosic biorefineries in the EU: Drivers, challenges, and opportunities**. *Renewable Sustainable Energy Rev* 2019, **101**:590–599. An excellent and up-to-date review of the current state of lignocellulosic biofuel production. They also consider non-crop feedstocks such as forestry residues, which helps contextualize biofuel crops in the larger bio-economy.

15*. Huang K, Won W, Barnett KJ, Brentzel ZJ, Alonso DM, Huber GW, Dumesic JA, Maravelias CT: **Improving economics of lignocellulosic biofuels: An integrated strategy for coproducing 1,5-pentanediol and ethanol**. *Appl Energy* 2018, **213**:585–594. Detailed technoeconomic analysis of a co-production strategy of 1,5-pentanediol and lignocellulosic biofuels. The analysis considers non-engineered feedstocks, but their modeling shows high economic sensitivity to changes in feedstock sugar composition, which have recently been engineered in feedstock crops.

16**.Vanhercke T, Belide S, Taylor MC, El Tahchy A, Okada S, Rolland V, Liu Q, Mitchell M, Shrestha P, Venables I, et al.: **Up-regulation of lipid biosynthesis increases the oil content in leaves of Sorghum bicolor**. *Plant Biotechnol J* 2019, **17**:220–232.

17*. Kalghatgi G: **Is it really the end of internal combustion engines and petroleum in transport?** *Appl Energy* 2018, **225**:965–974. This analysis considers alternatives to liquid fuels and discusses the advantages which have allowed liquid fuels to be the primary source of energy for transportation since the transition from animal power. Particular attention is paid to the high cost and difficulty scaling alternative ways of storing energy to power transportation.

18. Nair S, Paulose H: **Emergence of green business models: The case of algae biofuel for aviation**. *Energy Policy* 2014, **65**:175–184.

19. Shonnard DR, Williams L, Kalnes TN: **Camelina-derived jet fuel and diesel: Sustainable advanced biofuels**. *Environ Prog Sustain Energy* 2010, **29**:382–392.

20. Nandi S, Kwong AT, Holtz BR, Erwin RL, Marcel S, McDonald KA: **Techno-economic analysis of a transient plant-based platform for monoclonal antibody production**. *MAbs* 2016, **8**:1456–1466.

21. Ro D-K, Paradise EM, Ouellet M, Fisher KJ, Newman KL, Ndungu JM, Ho KA, Eachus RA, Ham TS, Kirby J, et al.: **Production of the antimalarial drug precursor artemisinic acid in engineered yeast**. *Nature* 2006, **440**:940–943.

22. Fuentes P, Zhou F, Erban A, Karcher D, Kopka J, Bock R: **A new synthetic biology approach allows transfer of an entire metabolic pathway from a medicinal plant to a biomass crop**. *Elife* 2016, **5**.

23. Galanie S, Thodey K, Trenchard IJ, Filsinger Interrante M, Smolke CD: **Complete biosynthesis of opioids in yeast**. *Science* 2015, **349**:1095–1100.

24*. Luo X, Reiter MA, d'Espaux L, Wong J, Denby CM, Lechner A, Zhang Y, Grzybowski AT, Harth S, Lin W, et al.: **Complete biosynthesis of cannabinoids and their unnatural analogues in yeast**. *Nature* 2019, **567**:123–126. A recent *tour de force* of metabolic engineering. This paper demonstrates the implementation and optimization of a complex pathway in a yeast host for bioproduction purposes. Of particular interest to this review is the comparison between state-of-the-art microbial biosynthesis and *in planta* biosynthesis.

25. Hunt VD: **Gasohol handbook**. 1981,

26. **USDA ERS - U.S. Bioenergy Statistics**. *USDA ERA* 2018,

27. Rosegrant MW: *Biofuels and grain prices: impacts and policy responses*. International Food Policy Research Institute Washington, DC; 2008.

28*. German L, Goetz A, Searchinger T, Oliveira G de LT, Tomei J, Hunsberger C, Weigelt J: **Sine Qua Nons of sustainable biofuels: Distilling implications of under-performance for national biofuel programs**. *Energy Policy* 2017, **108**:806–817. Summary of an Energy Policy special edition about biofuels. Focuses in particular on the negative impacts of current biofuel policies, and suggests research directions and required characteristics of environmentally sustainable future biofuels.

29. Rathore D, Nizami A-S, Singh A, Pant D: **Key issues in estimating energy and greenhouse gas savings of biofuels: challenges and perspectives**. *Biofuel Research Journal* 2016, **3**:380–393.

30. Fargione J, Hill J, Tilman D, Polasky S, Hawthorne P: **Land clearing and the biofuel carbon debt**. *Science* 2008, **319**:1235–1238.

31**.Rulli MC, Casirati S, Dell'Angelo J, Davis KF, Passera C, D'Odorico P: **Interdependencies and telecoupling of oil palm expansion at the expense of Indonesian rainforest**. *Renewable Sustainable Energy Rev* 2019, **105**:499–512. A comprehensive analysis of the negative externalities associated with a major current source of biodiesel: oil palm plantations. The authors determine the extent of rainforest loss due to land conversion for monoculture bioenergy crops, as well as the driving policies, which largely stem from distant and wealthy nations.

32. Budzianowski WM: **High-value low-volume bioproducts coupled to bioenergies with potential to enhance business development of sustainable biorefineries**. *Renewable Sustainable Energy Rev* 2017, **70**:793–804.

33. **Hybrid Poplar Composition - Idaho National Laboratory**. *Bioenergy Feedstock Library - Idaho National Laboratory* 2018,

34. **Switchgrass Composition - Idaho National Laboratory**. *Bioenergy Library* 2011,

35. **Sorghum Composition - Idaho National Laboratory**. *Bioenergy Library* 2016,

36. **Sugarcane Bagasse composition - Idaho National Laboratory**. *Bioenergy Library* 2016,

37. Godin B, Agneessens R, Gerin PA, Delcarte J: **Composition of structural carbohydrates in biomass: Precision of a liquid chromatography method using a neutral detergent extraction and a charged aerosol detector**. *Talanta* 2011, **85**:2014–2026.

38. Neutelings G: **Lignin variability in plant cell walls: contribution of new models**. *Plant Sci* 2011, **181**:379–386.

39. He J, Ma C, Ma Y, Li H, Kang J, Liu T, Polle A, Peng C, Luo Z-B: **Cadmium tolerance in six poplar species**. *Environ Sci Pollut Res Int* 2013, **20**:163–174.

40. Decker SR, Carlile M, Selig MJ, Doeppke C, Davis M, Sykes R, Turner G, Ziebell A: **Reducing the effect of variable starch levels in biomass recalcitrance screening**. *Methods Mol Biol* 2012, **908**:181–195.

41. Zhao YL, Steinberger Y, Shi M, Han LP, Xie GH: **Changes in stem composition and harvested produce of sweet sorghum during the period from maturity to a sequence of delayed harvest dates**. *Biomass Bioenergy* 2012, **39**:261–273.

42. Canilha L, Chandel AK, Suzane dos Santos Milessi T, Antunes FAF, Luiz da Costa Freitas W, das Graças Almeida Felipe M, da Silva SS: **Bioconversion of sugarcane biomass into ethanol: an overview about composition, pretreatment methods, detoxification of hydrolysates, enzymatic saccharification, and ethanol fermentation**. *Biomed Res Int* 2012, **2012**.

43. Bagheri M, Mansouri H: **Effect of induced polyploidy on some biochemical parameters in Cannabis sativa L**. *Appl Biochem Biotechnol* 2015, **175**:2366–2375.

44. Wyman CE, Balan V, Dale BE, Elander RT, Falls M, Hames B, Holtzapple MT, Ladisch MR, Lee YY, Mosier N, et al.: **WITHDRAWN: Comparative Data on Effects of Leading Pretreatments and Enzyme Loadings and Formulations on Sugar Yields from Different Switchgrass Sources**. *Bioresource Technology* 2011, doi:10.1016/j.biortech.2011.04.030.

45. Xu F, Zhou L, Zhang K, Yu J, Wang D: **Rapid Determination of Both Structural Polysaccharides and Soluble Sugars in Sorghum Biomass Using Near-Infrared Spectroscopy**. *Bioenergy Res* 2015, **8**:130–136.

46. Lingle SE, Thomson JL: **Sugarcane Internode Composition During Crop Development**. *Bioenergy Res* 2012, **5**:168–178.

47. Godin B, Lamaudière S, Agneessens R, Schmit T, Goffart J-P, Stilmant D, Gerin PA, Delcarte J: **Chemical Composition and Biofuel Potentials of a Wide Diversity of Plant Biomasses**. *Energy & Fuels* 2013, **27**:2588–2598.

48. Brenneisen R: **Chemistry and Analysis of Phytocannabinoids and Other Cannabis Constituents**. In *Marijuana and the Cannabinoids*. Edited by ElSohly MA. Humana Press; 2007:17–49.

Chapter 3: Gall wasps, plant bioengineers

Including material from published work:
Kasey Markel, Vlastimil Novak, Benjamin P. Bowen, Yang Tian, Yi-Chun Chen, Sasilada Sirirungruang, Andy Zhou, Katherine B. Louie, Trent R. Northen, Aymerick Eudes, Henrik V. Scheller, Patrick M. Shih "Cynipid wasps systematically reprogram host metabolism and restructure cell walls in developing galls." *Plant Physiology* 195.1 (2024): 698-712.

3.0 Chapter preface

Plant biotechnology is still in its infancy. Despite all the progress of the last ~40 years, we remain incredibly limited in the sorts of changes we can cause to plant metabolism, phenotype, and developmental habit. Following a time-honored tradition, we look to nature for inspiration, in this case to gall-inducing wasps that have been bending plants to their will in incredibly specific fashion since before *Homo sapiens* evolved. While we did not fully unravel the secrets of how exactly the wasps achieve their impressive engineering feats, we gain a better understanding of what is possible and learn some glimmers of what might be the mechanisms behind the phenomenon.

While the previous chapter discussed the importance of the development of ideotypes for modern crops to serve as phenotypic targets, this chapter investigates a natural system that demonstrates that plant phenotypes are much more engineerable and plastic than is generally imagined. While full utilization of the level of engineering mastery demonstrated by gall inducers remains beyond the grasp of current plant biotechnologists, it can be used as inspiration regarding how ambitious it may be prudent to be for ideotype development over the coming decades as plant biotechnology matures as a field.

3.1 Abstract

Many insects have evolved the ability to manipulate plant growth to generate extraordinary structures called galls in which insect larva can develop while being sheltered within and feeding on the plant. In particular, Cynipid (Hymenoptera: Cynipidae) wasps have evolved to form some of the most morphologically complex galls known and generate an astonishing array of gall shapes, colors, and sizes. However, the biochemical basis underlying these remarkable cellular and developmental transformations remains poorly understood. A key determinant in plant cellular development is the deposition of the cell wall to dictate the physical form and physiological function of newly developing cells, tissues, and organs. However, it is unclear to what degree cell walls are restructured to initiate and support the formation of new gall tissue. Here, we characterize the molecular alterations underlying gall development using a combination of metabolomic, histological, and biochemical techniques to elucidate how leaf cells are reprogrammed to form galls. Strikingly, gall development involves an exceptionally coordinated spatial deposition of lignin and xylan to form *de novo* gall vasculature. Our results highlight how Cynipid wasps can radically

change the metabolite profile and restructure the cell wall to enable the formation of galls, providing new insights into the mechanism of gall induction and the extent to which plants can be entirely reprogrammed to form novel structures and organs.

3.2 Importance

Galls are abnormal plant growths induced by another organism such as insects. Some of the most ornate galls found in nature are formed by cynipid wasps on the leaves of oak trees and used to house and feed wasp larvae. These galls display an astounding array of colors and structures. The molecular changes underlying this remarkable morphological transformation and development of an induced organ are largely unknown. We utilized a wide range of analytical, histological, and biochemical techniques to reveal a previously undescribed level of cell and tissue organization in many aspects of gall growth.

3.3 Introduction

Diverse organisms from fungi and bacteria to plants and insects have independently evolved the ability to manipulate the growth of plants to their advantage, forming abnormal structures referred to as galls. Galls induced by bacteria and fungi are generally morphologically simple and often referred to simply as ‘tumors’, whereas galls induced by insects are sometimes intricately and precisely structured and have fascinated naturalists since the time of ancient Greece (2, 3). While the exact mechanisms of gall induction remain largely unknown, chemical signals from the insect causing plant growth reprogramming has been the primary working hypothesis since the time of Charles Darwin, who presents “the poison secreted by the gall-fly produces monstrous growths on the wild rose or oak-tree” as one of the final arguments suggesting all plants (and in fact all life) share common ancestry. Interestingly, some gall inducers create galls on several species of host plants, and in these cases the gall morphology is remarkably similar (5). This demonstrates that the gall is an extended phenotype of the gall inducer (6, 7), which exerts greater control over gall morphology than the plant host. The genes underlying this extended phenotype in insect gall inducers remain almost entirely unknown, but the phenotype itself is striking: precise control over host growth, metabolism, and structure.

Deciphering of the mechanisms of gall induction has been a longstanding goal of gall research (1, 8). One major theory is that gall inducers synthesize plant hormones or hormone analogs, the local concentration gradients of which play a key role in gall development (9, 10). Synthesis of plant hormones – principally auxin and cytokinin – is known to be a key component in the generation of the simple galls induced by *Agrobacterium* (11). Nonetheless, hormones likely play some role in insect gall induction, a hypothesis supported by the detection of high concentrations of plant hormone analogues in gall tissue (12), though studies of other gall types have found galls to be auxin-depleted compared to normal tissue (13). Even more conspicuous evidence comes from the ability of some gall-inducing insects to synthesize plant hormones such as auxin and most likely cytokinin (13–15). Taken together, the balance of evidence suggests that phytohormone synthesis plays a role in the induction of at least some galls, but the exact mechanism is unclear and there are almost certainly other unknown elements to the induction of galls. In particular, simple concentration gradients of hormones are insufficient to explain the morphological diversity and complexity of insect-induced galls; thus there is a need to discover, study, and expand our understanding of the many non-phytohormone compounds that may contribute to the development and morphology of complex galls.

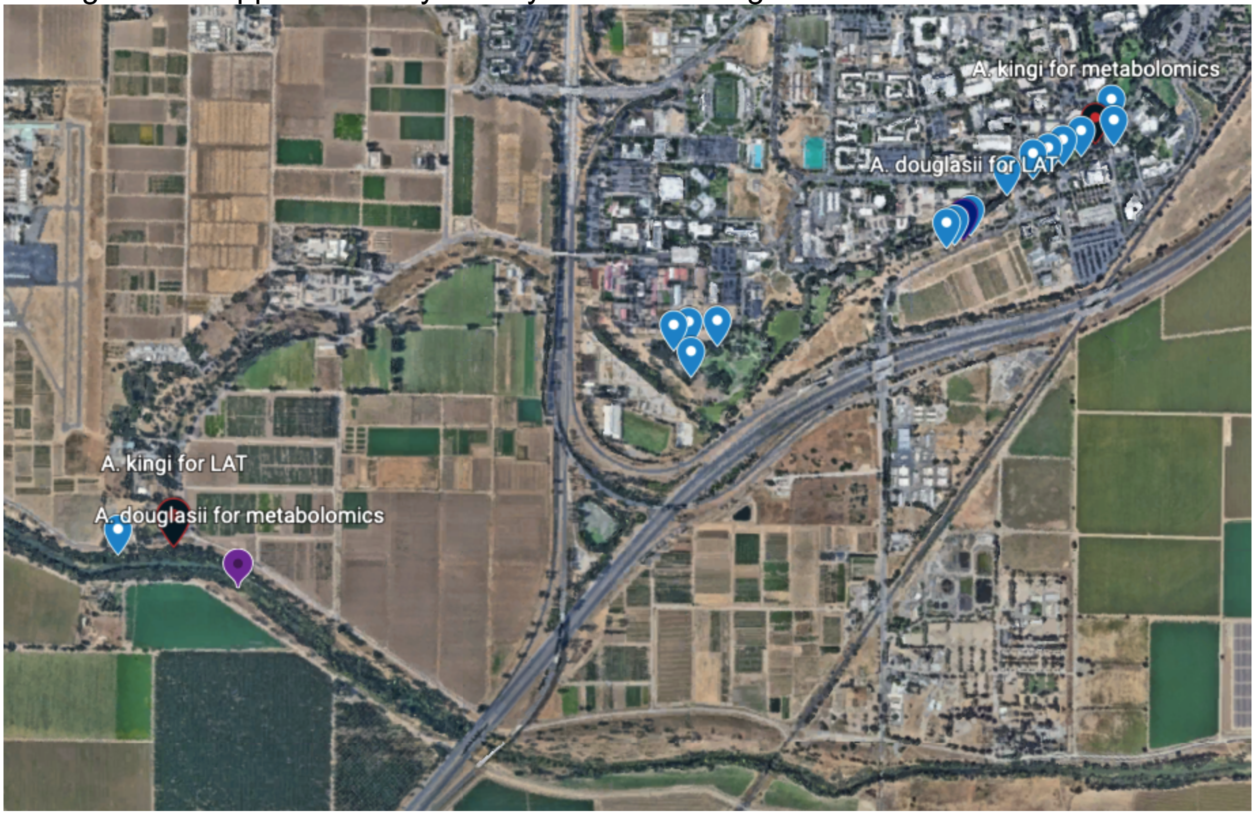
Because cell walls physically surround and constrain plant cells, any new growth such as the development of a gall requires breakdown, remodeling, and/or new deposition of cell wall material. As such, cell wall remodeling is key to organogenesis, such as the generation of new leaves (16). Despite the central role cell walls play in determining the structure and function of plant tissues, little is known about how plant cell walls are modified during the development of insect galls. Previous qualitative studies have shown changes in lignin (17), tannins (18)), and several polysaccharides (19, 20); however, there is a need for a more global understanding of all the metabolic changes underlying the transformation of the plant cell metabolites and cell walls during the formation of galls. Similarly, a more detailed spatial understanding of the molecular alteration in plant cell walls associated with gall induction may reveal unique insights into the relatively unexplored interplay between host cell reprogramming and cell wall remodeling.

Among the most morphologically complex and charismatic galls are those induced by *Cynipid* wasps on oak trees (21). Over 1300 species of cynipid wasps have been described (22), and many species alternate between a sexual and parthenogenic generation, each of which produces a distinct gall type (23). The diversity and morphological complexity of cynipid galls make them an excellent system to study the morphological, metabolic, and cell wall changes associated with gall induction. Recent molecular biology research on cynipid galls has been largely limited to transcriptomics studies (24, 25). These analyses have shed some light on the question of how cynipid wasps hijack the gene expression machinery of plants, but the metabolic consequences of these changes in gene expression remain largely unexplored. However, because insect gall induction is believed to be dependent on the generation of gradients of signaling molecules such as phytohormones and requires changes to cell wall structure and composition without obvious mRNA correlates, transcriptomics alone cannot tell the whole story. To provide a more comprehensive understanding of oak gall development, we perform a detailed analysis of the biochemical changes associated with gall induction and the concurrent alterations to cell wall structure and composition.

3.4 Results and Discussion

3.4.1 Morphological characterization of two distinct gall types

We collected cone galls induced by *Andricus kingi* and urchin galls induced by *Antron douglasii* from the leaves of the valley oak *Quercus lobata* in and near the UC Davis arboretum (trees sampled shown in Supplemental Figure 1, sampling dates and other galls identified shown in Dataset S1). Cone galls (Figure 1A) are cone shaped, usually red but often white along one side, and approximately 5 mm across at maturity. Urchin galls (Figure 1B) are rarer and larger, light purple in color, and urchin-shaped with 5-15 spikes. Both are attached to the leaf by a thin (<200 μm) section of tissue that projects orthogonal to the plane of the leaf blade and defines the axis of rotational symmetry for cone galls and approximate symmetry for the urchin galls.



Supplemental Figure 1: Map of gall collection sites. Round tags indicate individual oak trees from which galls were harvested, red tags indicate harvests of galls induced by *Andricus kingi* that were used in further experiments, purple tags indicate harvests of galls induced by *Antrol douglasii* in further experiments. For each experiment, all galls of a particular type were harvested from one tree on one day.

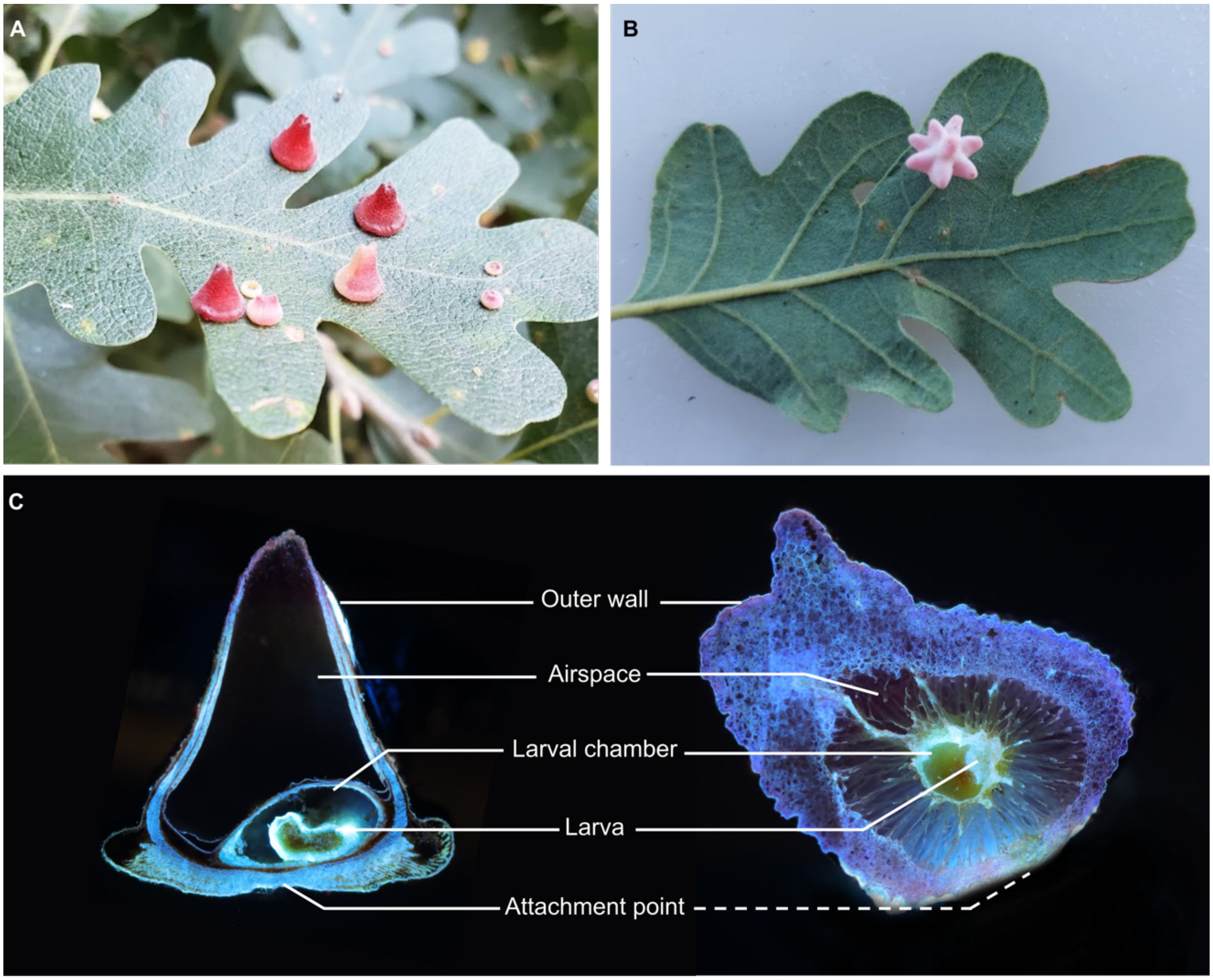
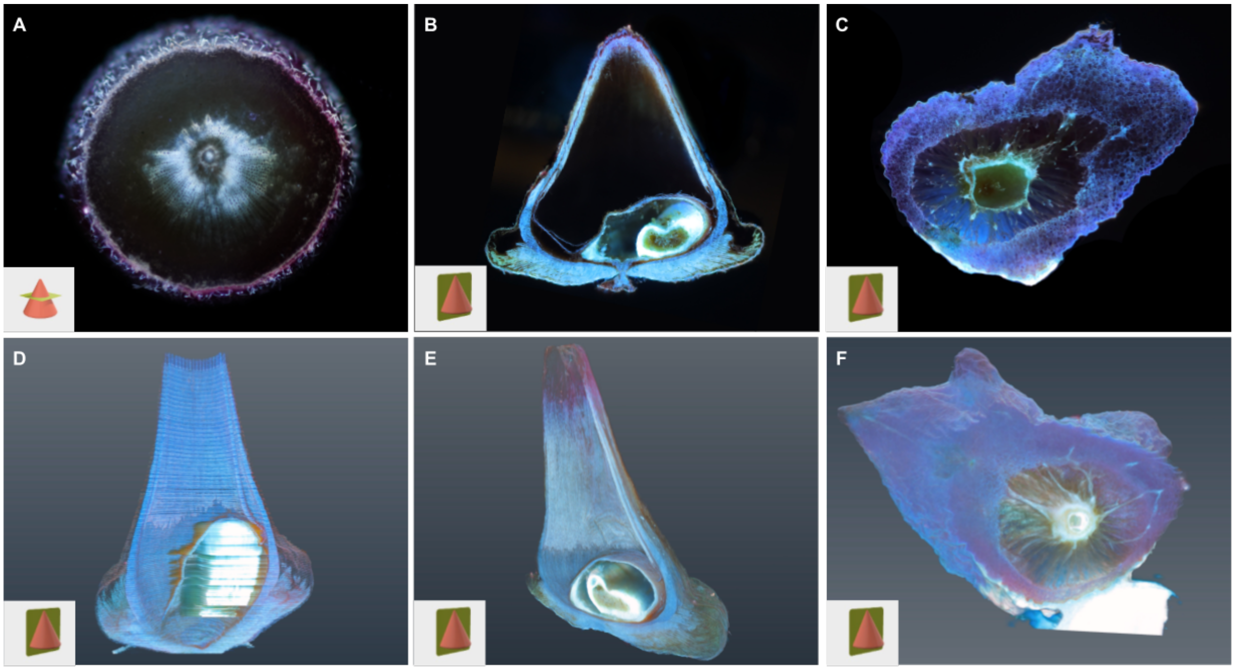


Figure 1: Comparison of anatomical features shared across morphologically disparate galls. A: Cone galls induced by *Andricus kingi*. B: Urchin gall induced by *Antron douglasii*. C: Longitudinal sections of cone (left) and urchin (right) galls, imaged with laser ablation tomography. Dashed line to urchin gall attachment point indicates attachment point is out of plane and location is approximate.

Galls induced by cynipid wasps are complex three-dimensional structures; however, the vast majority of studies have been constrained to two-dimensional sections, which has been insufficient to comprehensively characterize the relationship between plant and insect tissue. To fill this gap, we used laser ablation tomography (LAT) to generate high-resolution three-dimensional reconstructions of galls consisting of thousands of two-dimensional slices with ~8 μm resolution (Figure 1C, Supplemental Figure 2). Three-dimensional models reveal the internal structure of cone (Supplemental Videos 1 and 2) and urchin (Supplemental Video 3) galls. The *Andricus kingi* larva within the cone galls is highly autofluorescent, facilitating easy discrimination between larval and plant tissue. Surprisingly, we found the larva in different orientations in the two cone galls imaged, with the long axis aligned to the longitudinal axis of the cone gall in one case and orthogonal in the other. This variation in the orientation of the larval chamber in conjunction with the tight conservation of the overall gall structure suggests something other than the larva provides the “orientation lodestar”, most likely the attachment point to the leaf.



Supplemental Figure 2: Laser ablation tomography sections and three-dimensional models. A: Transverse section of cone gall. B: Longitudinal sectional of cone gall. C: longitudinal section of urchin gall. D: three-dimensional model of cone gall imaged in transverse sections (apparent horizontal lines are an artifact of aliasing in assembling individual slices and stochastic variance in image brightness), corresponds to supplemental movie 1. E: three-dimensional model of cone gall imaged in longitudinal sections, corresponds to supplemental movie 2. F: three-dimensional model of urchin gall, corresponds to supplemental movie 3. Inset diagrams in lower left corner indicate orientation as viewed, which corresponds to view as imaged with the exception of D, where view shown here is orthogonal to the imaging plane.

While the morphology of both types of gall is very different, they both consist of a relatively thick outer wall of plant cells, an airspace, and an inner layer of plant cells surrounding a fluid-filled cavity which houses the developing larva. The urchin gall larval chamber is suspended by thin strands of plant tissue in the center of the airspace, providing thermal insulation. The thick exterior wall and airspace have been demonstrated to be important for protection of the larva from the elements (26) and hypothesized to be important for protection from predators and parasitoids (21). An evolutionary arms race between gall inducers and these natural enemies is likely a contributing source of the tremendous variation in cynipid gall morphology.

The three-dimensional models show that plant epidermal cells surrounding the larval chamber are patterned in a smooth ovate structure. While the galls imaged with LAT were relatively mature, the insect larva remained fairly undeveloped, and likely incapable of chewing through the plant cells surrounding the chamber. However, they were within an order of magnitude of the size of the adult wasps, which means they had almost certainly grown quite substantially to reach their current size. These facts together support the hypothesis that up to and including this gall developmental stage, insect larvae are absorbing nutrients through the fluid within the larval chamber, much like other animals feed on energy reserves within an egg or plant seedlings feed on endosperm, and in contrast to the mechanical chewing found in galling thrips (27) and during the final stages of cynipid development (28). The fluid of the larval chamber is most likely to be translocated photosynthate and nutrients, highlighting the importance of vasculature to support proper gall development.

3.4.2 Metabolomic profiles of different gall types are distinct and unique

We examined the metabolomic profiles of the two gall types, looking for common patterns that may suggest the homologously shared mechanism of gall induction as well as differences that may explain the differences in gall morphology. Previous research has focused either on a small number of pre-selected metabolites (29–31) or on the transcriptional profile of galls (24, 25). We utilized untargeted metabolomics to identify metabolite differences between gall and normal leaf tissues, enabling detection of thousands of mass features between samples. The most recent common ancestor of the two species of gall wasp studied most likely also induced galls (32), and therefore shared changes in the metabolomic profile of the two galls may suggest key elements of the basic mechanism of gall induction, whereas differences between the two galls may be either a cause or result of more idiosyncratic elements of gall structure or random changes due to drift.

Metabolic changes associated with initial induction of galls are expected to be especially pronounced during the early stages of gall development. Therefore, cone and urchin galls were subdivided into 5 and 4 developmental stages respectively using mass as a proxy for developmental stage. To our knowledge, the resulting dataset is the first metabolomic analysis of cynipid gall development incorporating a developmental time-series design. We obtained 8690 mass features; the full datasets for positive and negative mass spectrometry modes are available as Dataset S2 and S3, respectively, heat map of mass feature peak height available in Supplemental Figure 3. Principal component analysis demonstrated mass feature composition was distinct for leaf, urchin gall, and cone gall samples (Figure 2A). The majority (63%) of these mass features were shared between at least two sample types, and 29% were shared among all three (Figure 2B).

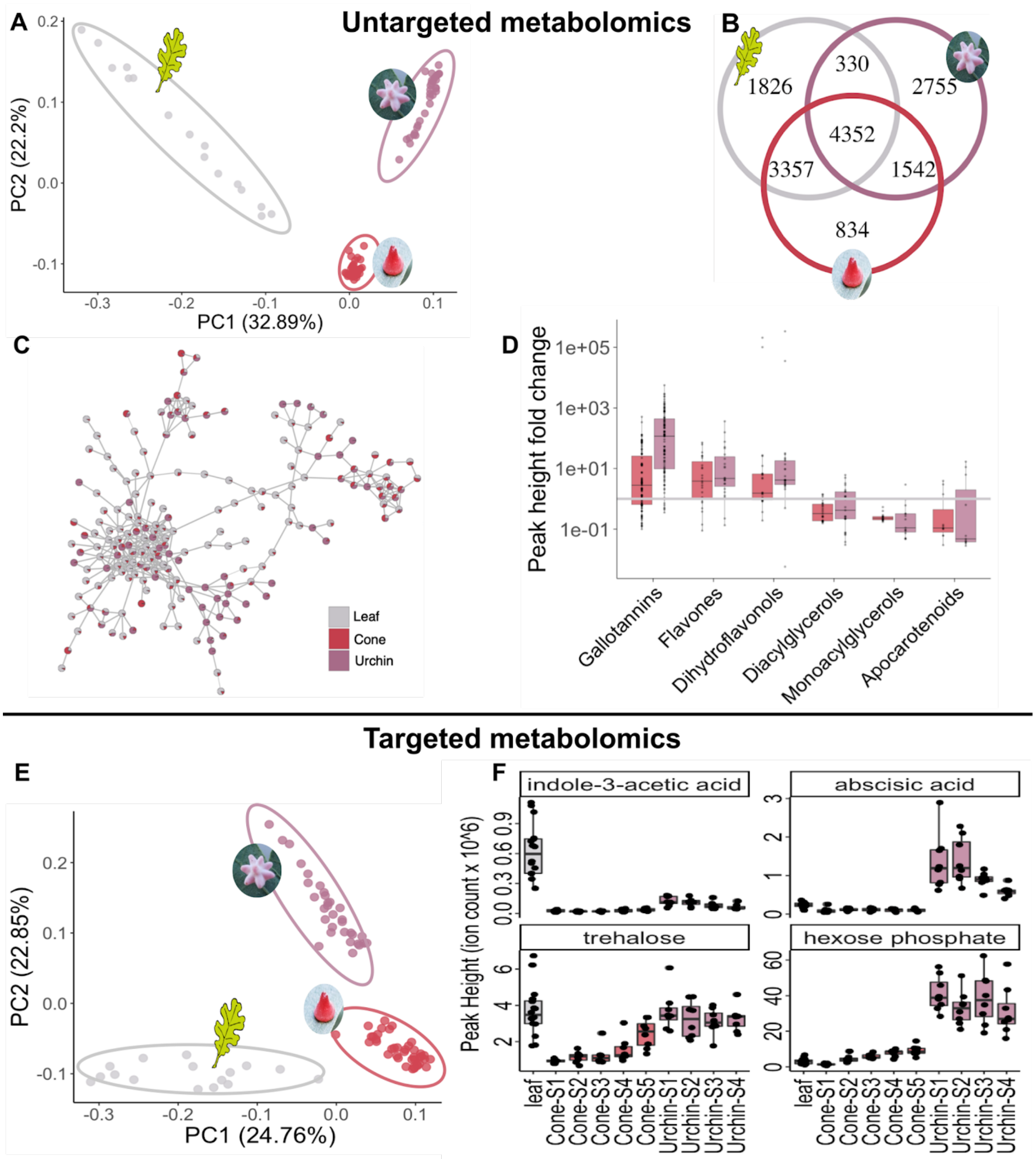
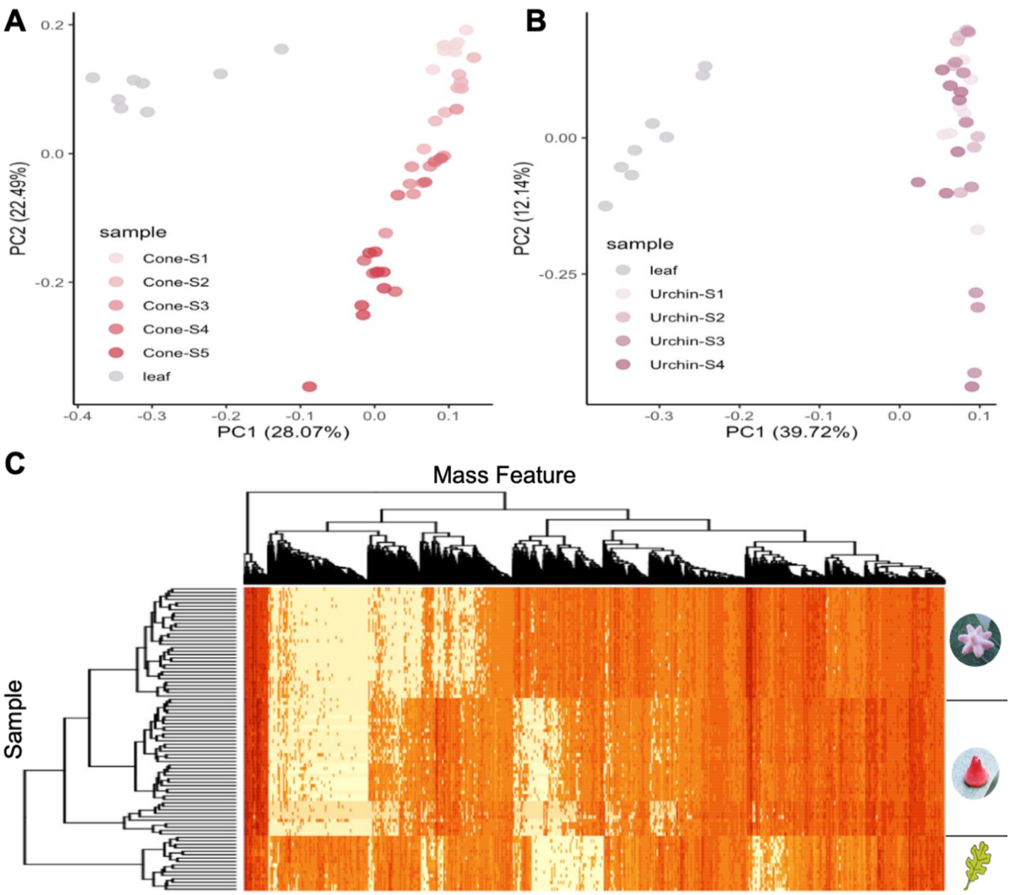


Figure 2: **Galls are metabolically distinct from each other and leaf tissue.** A: Principal component analysis of all 8690 mass features recorded in positive mode in untargeted metabolomics B: Venn Diagram of the mass features present in each sample type in untargeted metabolomics. C: Molecular networking. Each node is a mass feature, each edge indicates a cosine score of fragmentation pattern of at least 0.7, nodes are pie charts indicating relative peak heights between the three sample types. D: Natural products classes of putative identifications of mass features in untargeted metabolomics. E: Principal component analysis of all 209 metabolites

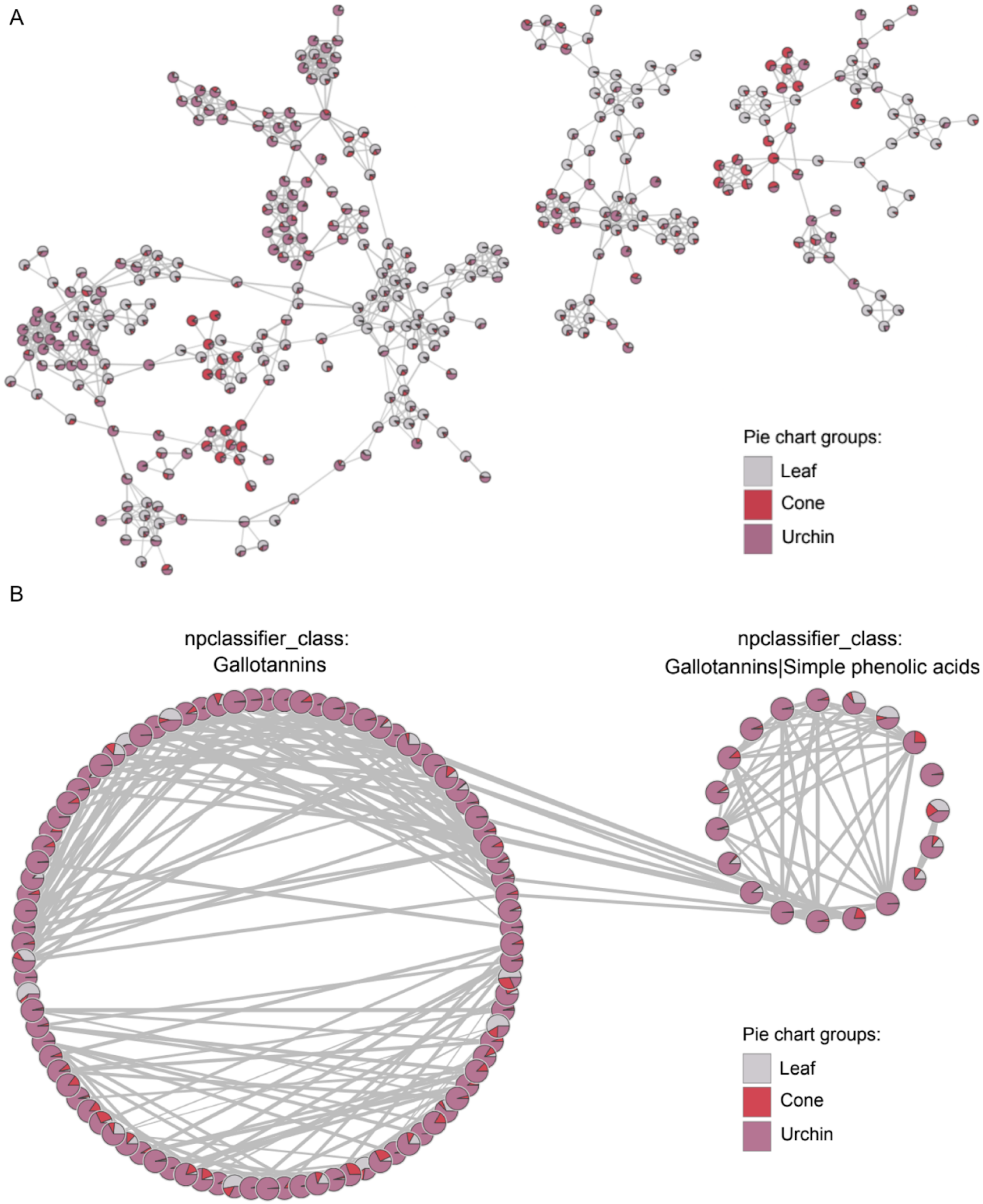
positively identified with mass-charge ratio, secondary fragmentation pattern, and retention time confirmed against a library for the same instrument in positive and negative mode, removing whichever was lower to generate a nonredundant dataset. F: Metabolite data for leaf and several growth stages of each type of gall for two hormones and two sugars. MS-MS mirror plots with more precise identification information abscisic acid, trehalose, and hexose phosphate are available in Supplemental Figures 6, 7, and 8 respectively. Indole-3-acetic acid peak height was too low to trigger MS-MS, identification was based on retention time, m/z ratio, and other mass feature characteristics shown in Dataset S4.

We performed network analysis using GNPS, which revealed that mass features overrepresented in particular sample types often clustered, demonstrating similar classes of compounds were enriched in specific galls (Figure 2C, Supplemental Figure 3). Several interactive networks are available online at NDExbio – further described in methods. We used m/z ratio and networking to generate putative identifications for each mass feature and used NPClassifier to categorize them (33), revealing increases in several expected compound classes such as gallotannins (whose name derives from 'gall') in gall tissue compared to leaf. We also observed an increase in two flavonoid categories and a decrease in two acylglycerol categories as well as apocarotenoids (Figure 2D, Supplemental Figure 5). Our finding of increased flavonoid accumulation corroborates a recent report of upregulation of flavonoid biosynthetic genes in cynipid-oak galls (25), which also may be the underlying basis of the pigmentation of the galls themselves. The decrease in acylglycerols is consistent with the same study observing that 2 of the top 50 upregulated genes were annotated as "hydrolysis of fatty acids." It has been proposed that fatty acids are converted into sugars to feed the growing larva (25). Finally, cynipid galls have previously been shown to contain lower concentrations of chlorophyll and carotenoids (34, 35), suggesting reduced photosynthesis as an explanation for the reduction in apocarotenoids observed here.



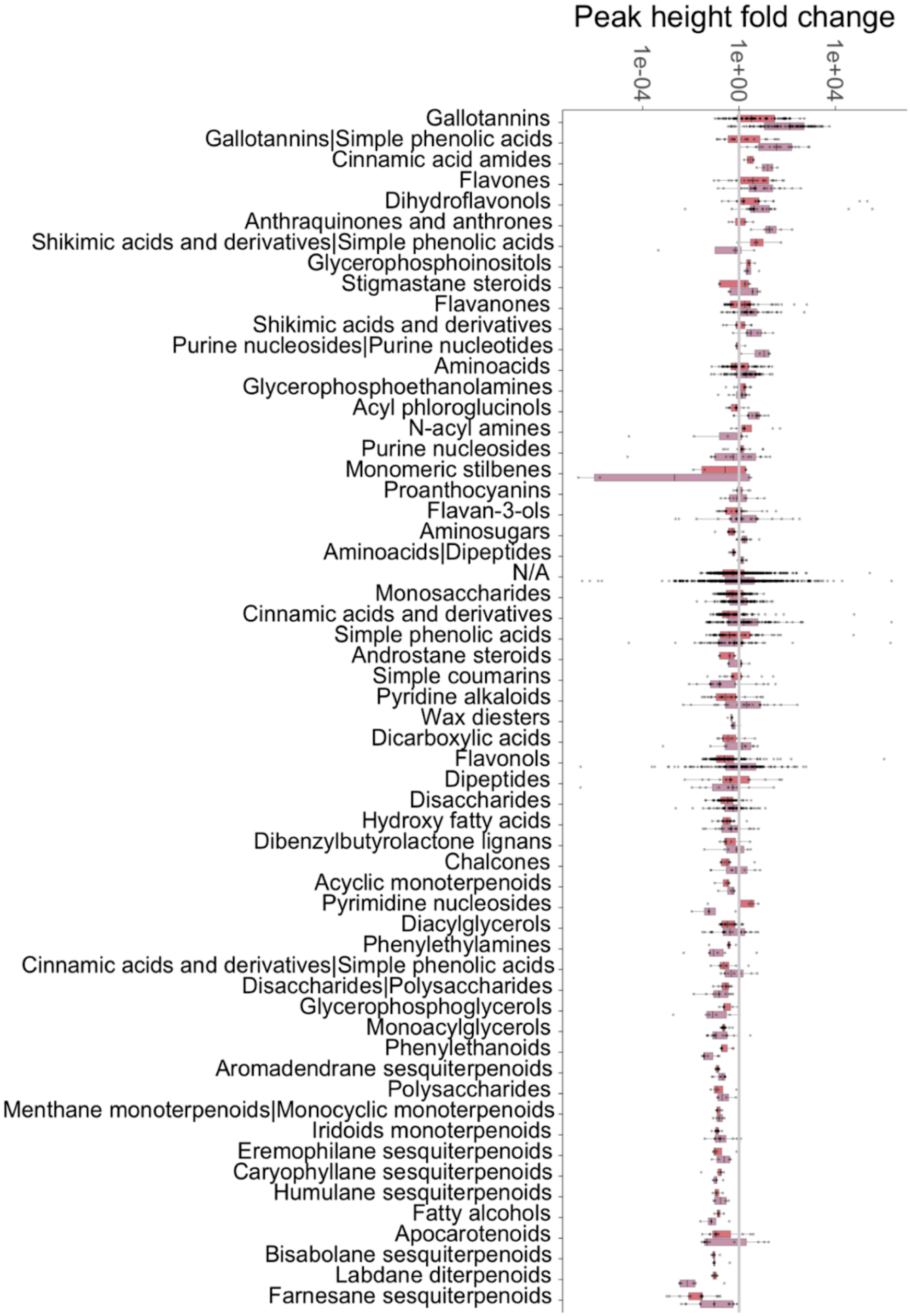
Supplemental Figure 3: Gall metabolomes cluster by developmental stage. A: PCA of the developmental stages of cone galls. B: PCA of the developmental stages of urchin galls. C: Heatmap of untargeted metabolomics with dendrogram. Rows are samples, columns are metabolites.

To provide a more detailed and quantitative understanding of metabolite changes, we next performed targeted metabolomics based on a library of standards with known retention time and fragmentation data to identify specific metabolites that broadly cover a wide sampling of primary metabolism and many core plant metabolites. Targeted metabolomic analyses combined the positive and negative mode datasets by choosing whichever had higher peak height (following methodology from (36)), resulting in a non-redundant dataset of 209 metabolites with confidence score of at least “Level 1”, meaning at least two independent and orthogonal data are used to confirm metabolite identity (37). Identification evidence including MS1, MS2, and chromatographic peak comparisons are available as Dataset S4 and S5 for positive and negative modes respectively. Principal component analysis of this stringently curated dataset revealed a distinct separation of sample types (Figure 2E), the full dataset is available as Dataset S6. Additional principal component analyses of the growth stages of each type of gall reinforce clear distinction between gall and leaf metabolites and show some clustering by gall growth stage (Supplemental Figure 3).

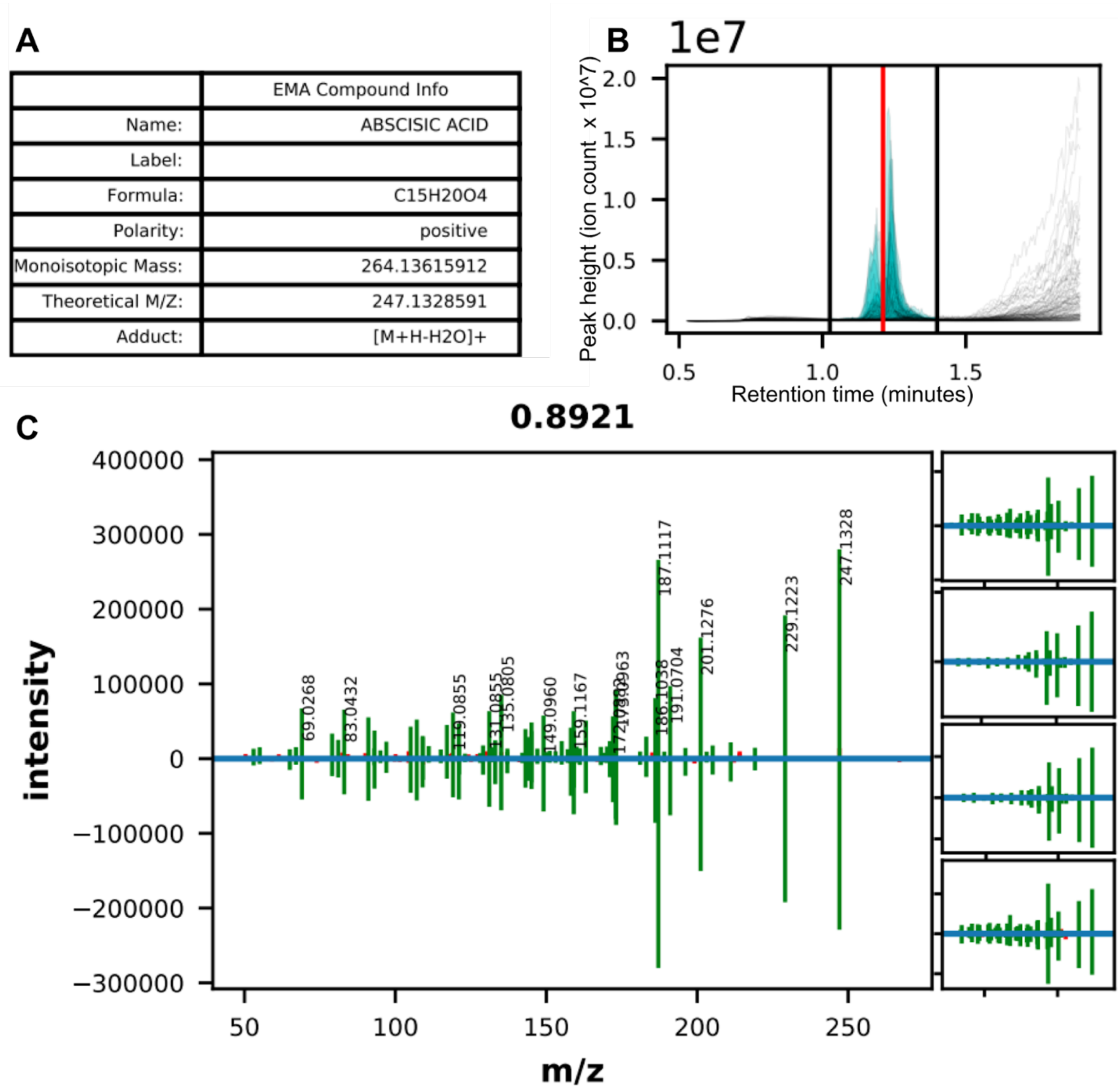


Supplemental Figure 4: Gall metabolomes are distinct. A: Network from untargeted metabolomics. Interactive networks with putative identifications of mass features available at NDExBio. B: Network of mass features classified as gallotannins by NPClassifier.

Of these 209 identified metabolites from targeted metabolomics, 43 had peak heights in urchin galls greater than four times higher than the leaf average, and 22 had peak heights in cone galls greater than four times higher than the leaf average. Of these highly gall-abundant metabolites, 11 were enriched in both gall types, much more than would be predicted if peak height were independent in both gall types (p = 0.001, hypergeometric test). Peak height data for the 54 metabolites >4x higher peak height in galls compared to leaf tissue is available in Dataset S7, these metabolites are candidates for either causes or conserved metabolic effects of the gall induction process, and may be useful leads for future efforts to determine the mechanism of gall induction. We used NPClassifier to classify all 209 metabolites by pathway and evaluated whether any pathways were overrepresented among the metabolites enriched in galls. For both gall types, there were fewer fatty acids than chance (p=0.052 for cone galls, p = 0.024 for urchin galls, hypergeometric test), supporting the results from the untargeted metabolomics.

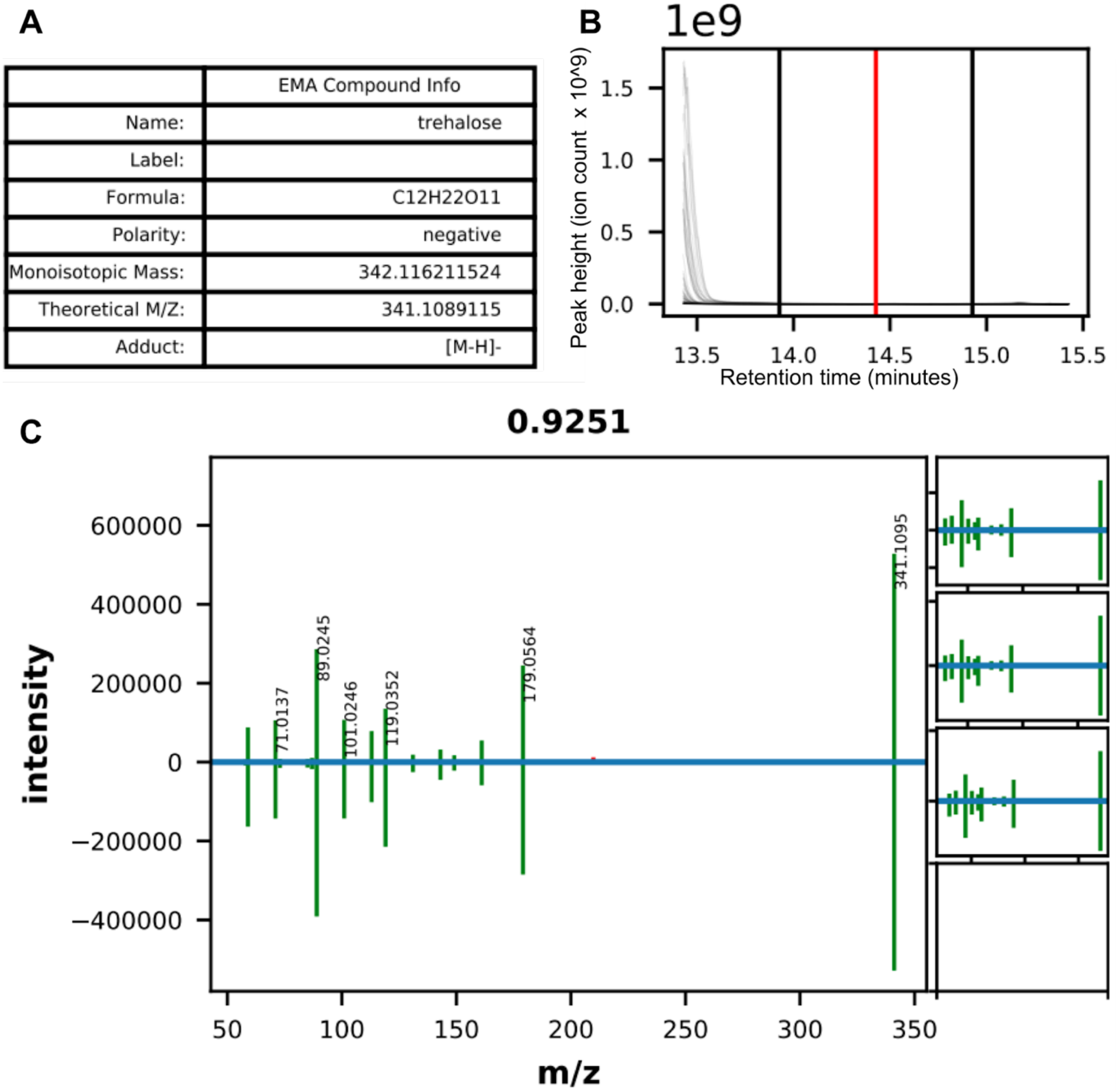


Supplemental Figure 5: **Galls differ in chemical class concentration** Peak height fold change for all NPClassifier categories for both types of gall compared to leaf tissue.



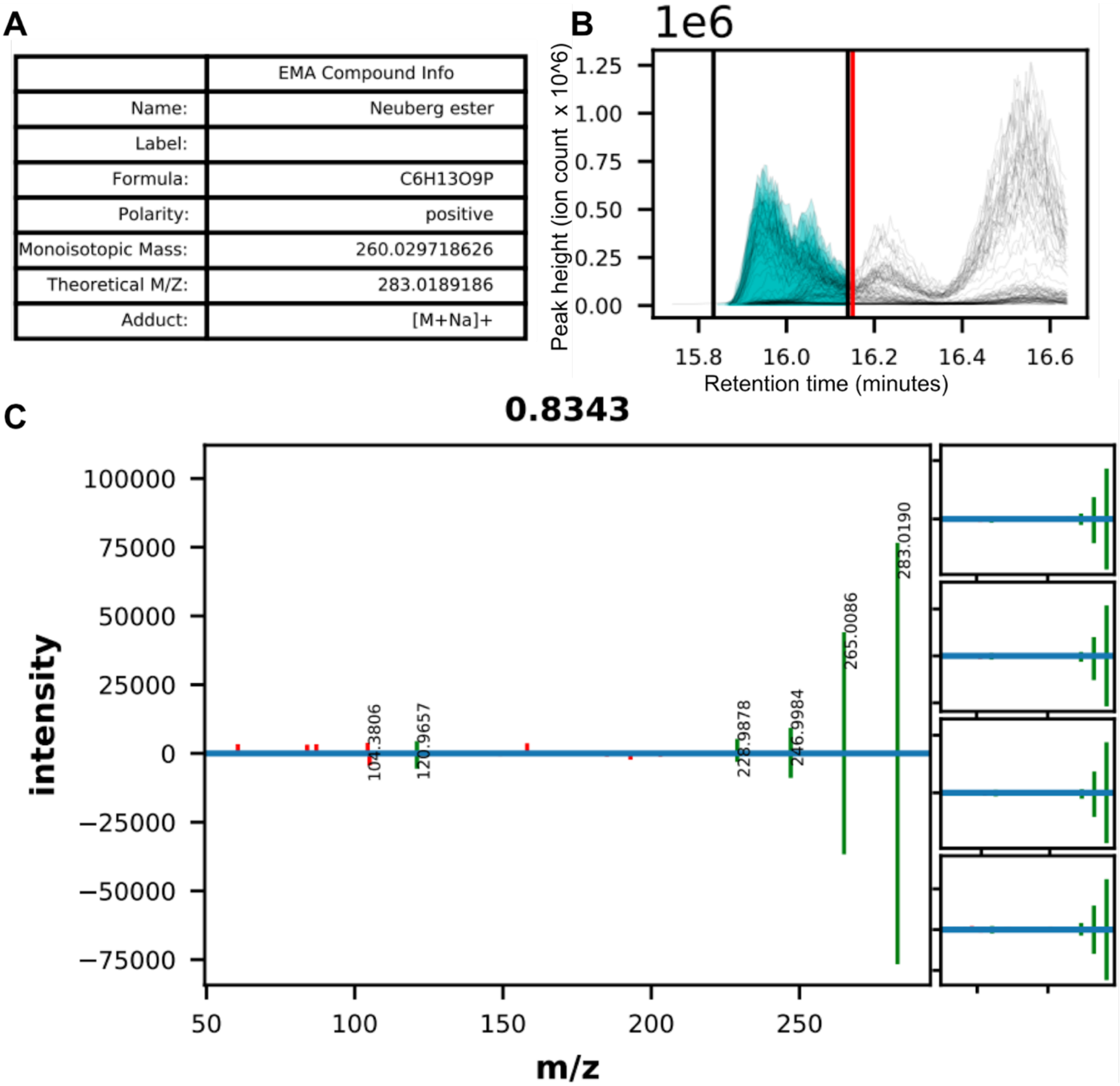
	EMA Compound Info
Name:	ABSCISIC ACID
Label:	
Formula:	C15H20O4
Polarity:	positive
Monoisotopic Mass:	264.13615912
Theoretical M/Z:	247.1328591
Adduct:	[M+H-H2O]+

Supplemental Figure 6: MS-MS details for abscisic acid. A: Information table. B: Extracted ion chromatogram across all files. The predicted retention time is the red vertical line, the black vertical lines are the integration bounds. C: MS-MS mirror plot, our data on top and standard MS-MS fragmentation pattern on bottom. Cosine score = 0.8921. Panels to the right are the next four highest scoring comparisons.



	EMA Compound Info
Name:	trehalose
Label:	
Formula:	C12H22O11
Polarity:	negative
Monoisotopic Mass:	342.116211524
Theoretical M/Z:	341.1089115
Adduct:	[M-H]-

Supplemental Figure 7: MS-MS details for trehalose. A: Information table. B: Extracted ion chromatogram across all files. The predicted retention time is the red vertical line, the black vertical lines are the integration bounds. Due to software constraints, this graph is too zoomed out on the Y axis because of another peak around 13.2 minutes. The maximum peak intensity for trehalose is 7.65 x 10^6, which is not visible due to the Y axis scaling. The peak is at 14.42 minutes (very close to the predicted retention time), and the integration bounds are set well, as can be seen in Dataset S5. C: MS-MS mirror plot, our data on top and standard MS-MS fragmentation pattern on bottom. Cosine score = 0.9251. Panels to the right are the next four highest scoring comparisons. In this case, only alternative three matches were found by the software.



A

	EMA Compound Info
Name:	Neuberg ester
Label:	
Formula:	C6H13O9P
Polarity:	positive
Monoisotopic Mass:	260.029718626
Theoretical M/Z:	283.0189186
Adduct:	[M+Na]+

Supplemental Figure 8: MS-MS details for hexose phosphate. A: Information table. While the software identified this feature as Neuberg ester meaning Fructose-6-phosphate, we are not confident of this specific isomeric identification, only that this is a hexose phosphate. B: Extracted ion chromatogram across all files. The predicted retention time is the red vertical line, the black vertical lines are the integration bounds. C: MS-MS mirror plot, our data on top and standard MS-MS fragmentation pattern on bottom. Cosine score = 0.8343. Panels to the right are the next four highest scoring comparisons.

3.4.3 Conserved metabolite changes across different galls reveals drastic changes in plant hormone and sugar concentrations

We next examined the concentration of plant hormones detected in the metabolomic analyses, which have been hypothesized to play important roles in the gall induction process. Structurally complex galls can be thought of as a novel organ functioning for the benefit of the gall inducer, and hormone concentration gradients are known to be central to the growth of organs such as leaves (38), flowers (39, 40), and fruits (39). Interestingly, the transcriptomic profile of galls induced by phylloxera on grape leaves shares many similarities to the transcriptome of fruits (41).

We found major differences in the concentration of auxin (indole-3-acetic acid) and abscisic acid between galls and normal leaf tissue (Figure 2F). Existing literature shows that auxin and cytokinin are sometimes increased and sometimes decreased in gall tissue compared to normal plant tissue, suggesting there may be multiple separate mechanisms of plant growth manipulation used by different groups of gall inducers (reviewed in (42)). This is perhaps not surprising given that the gall-inducing habit has evolved independently many times in separate lineages (43). In both cone and urchin galls, we see a massive decrease in the concentration of auxin (Figure 2F). This is somewhat surprising given the relatively low baseline levels of auxin in the middle of a leaf lamina (44), and even more surprising in light of the fact that RNAseq of a closely related cynipid-induced oak gall showed upregulation of auxin-response genes (24). While it is possible that these discordant results reflect different ground truths in these closely-related cynipid galls, it is also possible that upregulation of auxin biosynthetic genes does not result in increased auxin accumulation, highlighting a potential pitfall of interpretations of small molecule concentration solely made by transcript levels without direct biochemical measurement.

A recent gall tissue-specific RNAseq study found upregulation of auxin biosynthetic genes only in the larval chamber tissue, which comprises a relatively small portion of the total gall biomass, with low expression of auxin responsive genes throughout the remainder of the gall (25). This finding may help reconcile the seemingly contradictory results: while auxin is involved somehow in the gall induction process, if only the gall larval chamber contains high concentrations of auxin, then depending on the mass ratio of the larval chamber compared to the exterior of the gall we would expect to see some reports of higher auxin concentration and some reports of lower concentration within galls, which is indeed what has been reported (42).

Abscisic acid concentration is increased in urchin galls, but not cone galls (Figure 2F). Abscisic acid is often associated with stress, and has been shown to increase in response to attempted gall induction on resistant plants, while remaining constant between gall and normal tissue in susceptible plants (45). In another gall system, abscisic acid was reported to be decreased in gall tissue compared to normal plant tissue (46). In light of these diverging results among very phylogenetically distant gall systems, it is interesting to see different behavior in abscisic acid response even among two closely related galls on the same plant host. It is also worth noting that the only mass features identified as apocarotenoids increased in gall tissue in Figure 2D were putatively identified as abscisic acid. Since this was an independent mass spectrometry run, that both strengthens the results from this targeted analysis and their removal from the apocarotenoid class (of which abscisic acid is clearly a non-central example) strengthens the finding that apocarotenoids are depleted in gall tissue.

We also observed a striking pattern in the concentration of trehalose, a disaccharide known to play important signaling and regulatory roles. Trehalose plays a stress signaling role in plants, and exogenous application of trehalose induces resistance against pathogens (47). The massive reduction of trehalose concentration in cone galls may suggest the wasps are silencing this defense response. Trehalose also plays important roles in insects; it is a major circulating carbohydrate in the hemolymph (48), as well as a regulator of long-term hibernation-like states (49). Therefore, further research is necessary to fully understand the significance of the trehalose reduction in gall tissue.

We next examined hexose phosphates, central metabolic intermediates which are a primary output of photosynthesis and primary input into cell wall assembly. Hexose phosphates are substantially enriched in all surveyed developmental stages of urchin gall tissue compared to leaf, but remain constant at leaf-like levels in cone galls (Figure 2F). On average, hexose phosphate levels in urchin galls are over ten times higher than the leaf baseline. In general, the majority of hexose phosphates are destined for generation of starch or cell wall polysaccharides, suggesting the rerouting of metabolism to support gall development and larval feeding.

3.4.4 Gall cell layers are chemically distinct and highly lignified suggestive of *de novo* vascularization

Though the three-dimensional models generated by laser ablation tomography offer unique structural insights, they lack the chemical information. Metabolomic analysis offers chemical information, but without spatial data. To address the intersection of these interests, we turned to histochemical staining. Histochemical staining is a standard approach to identifying plant tissue types, yet there are no published micrographs of either of the galls studied here. Therefore, we next used a series of classic plant histology stains on cone galls (chosen for microscopy as they were more abundant) to examine the chemical composition and distribution to better understand the chemical changes associated with gall development. Safranin O, Congo red, Mäule stain, cellulose azure, orange G, FastGreen FCF, and aniline blue failed to show any interesting spatial patterns within the gall material (Supplemental Figure 9). Toluidine blue was useful for generating contrast to determine cell wall morphology and differentiate cell layers (Supplemental Figure 10). Weisner reagent (phloroglucinol + HCl) revealed the most striking spatial pattern, demonstrating tight spatial regulation of lignin deposition in gall tissue (Figure 3A, B). Two sclerenchyma cell layers are strongly stained (Supplemental Figure 10A), and the central sponge layer between them contains bundles of 4-9 cells in cross section with moderate lignification, which is suggestive of vasculature.

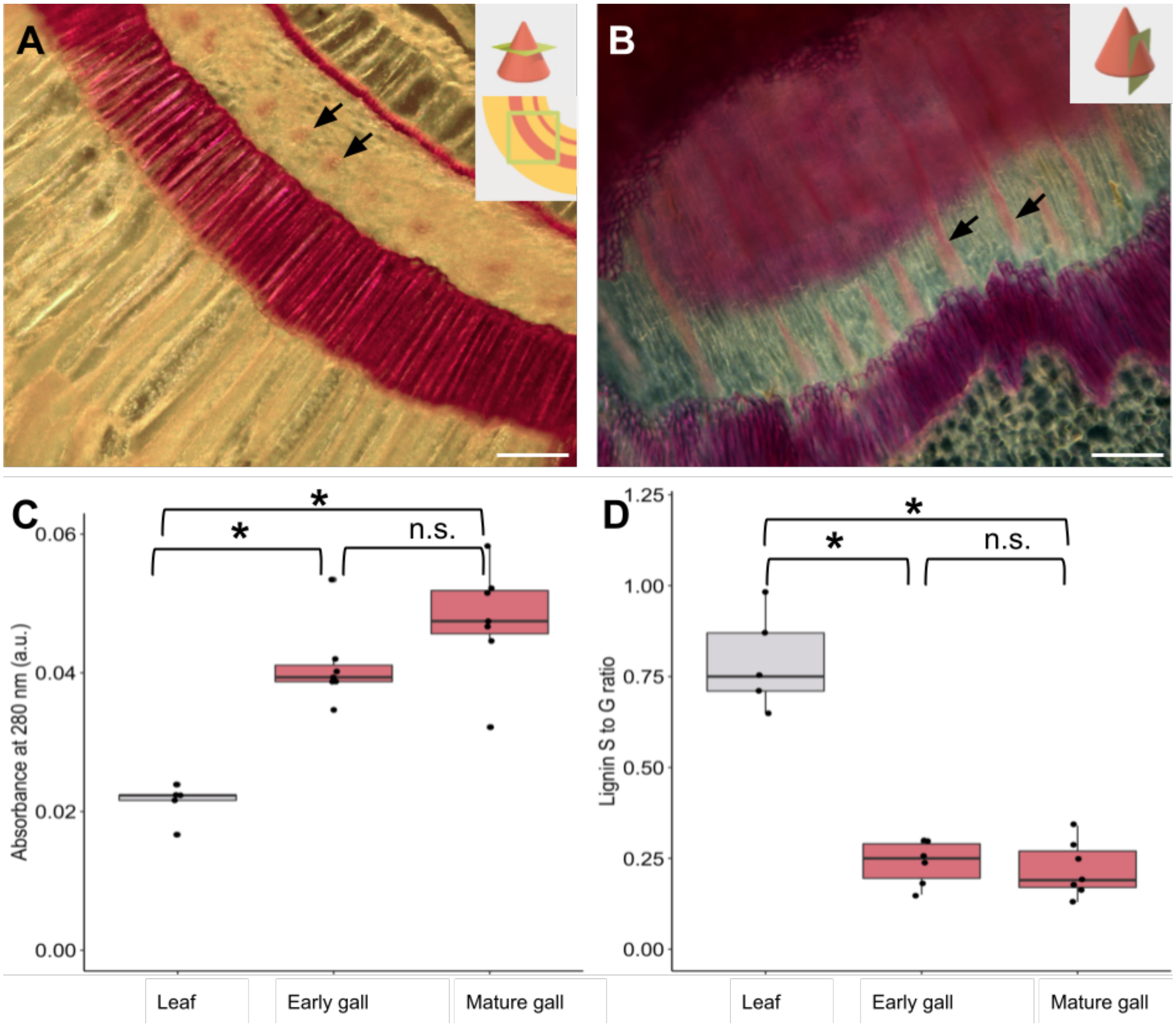
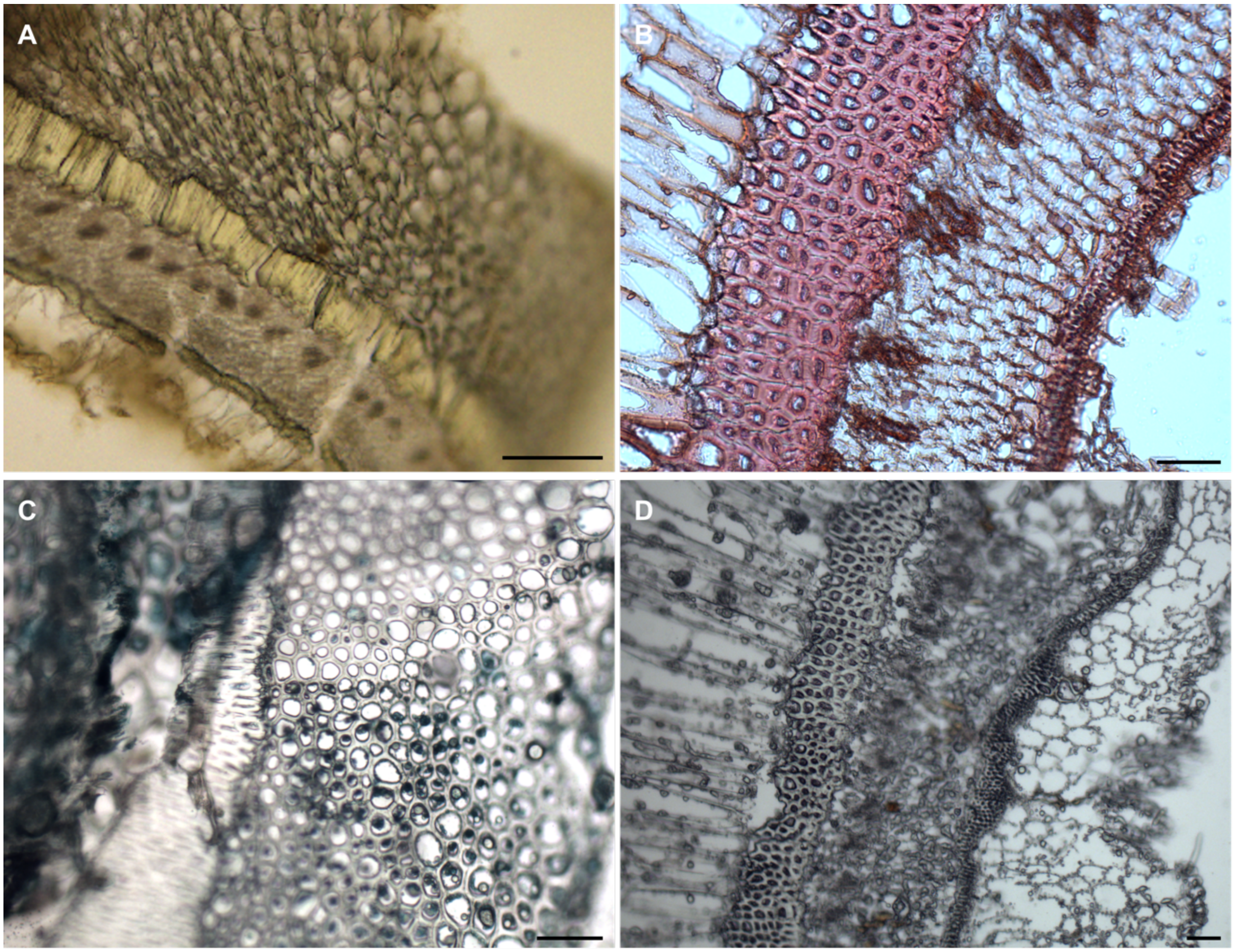
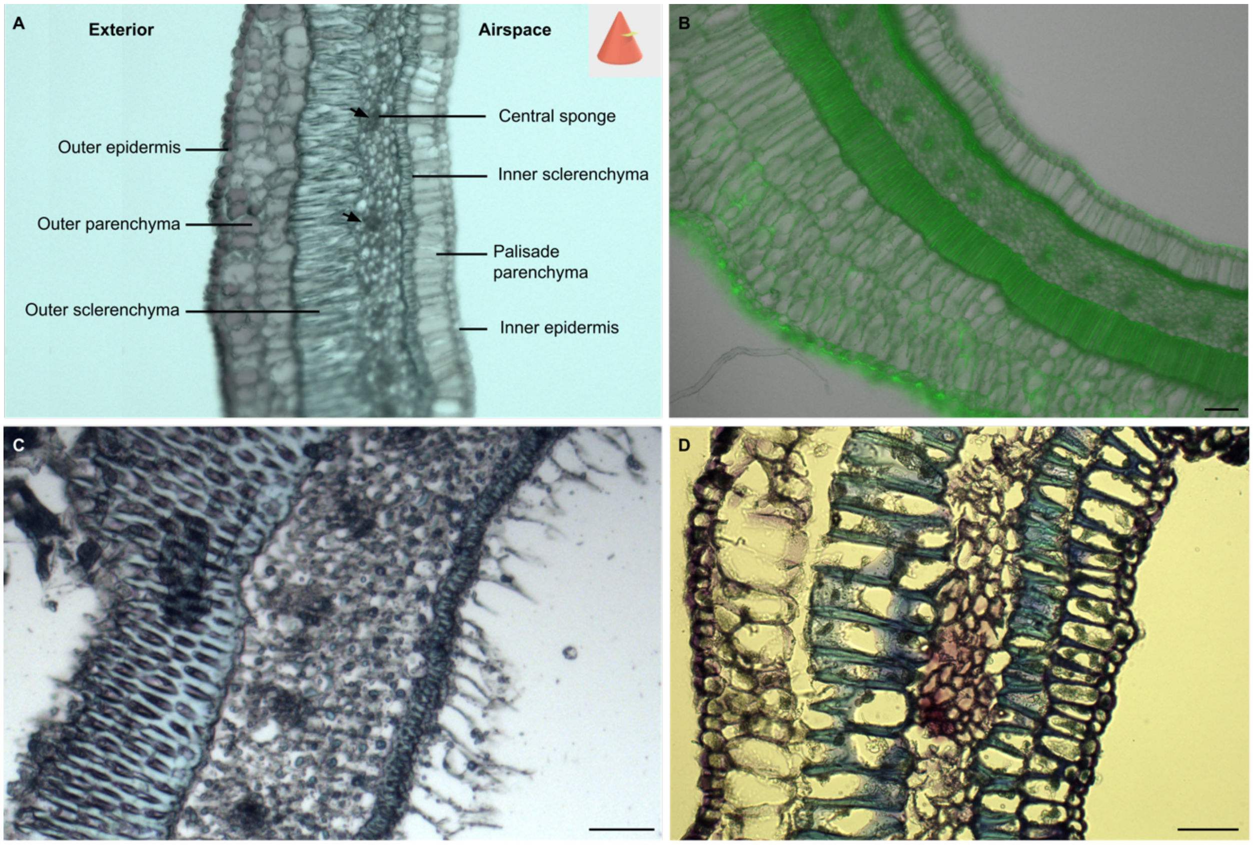


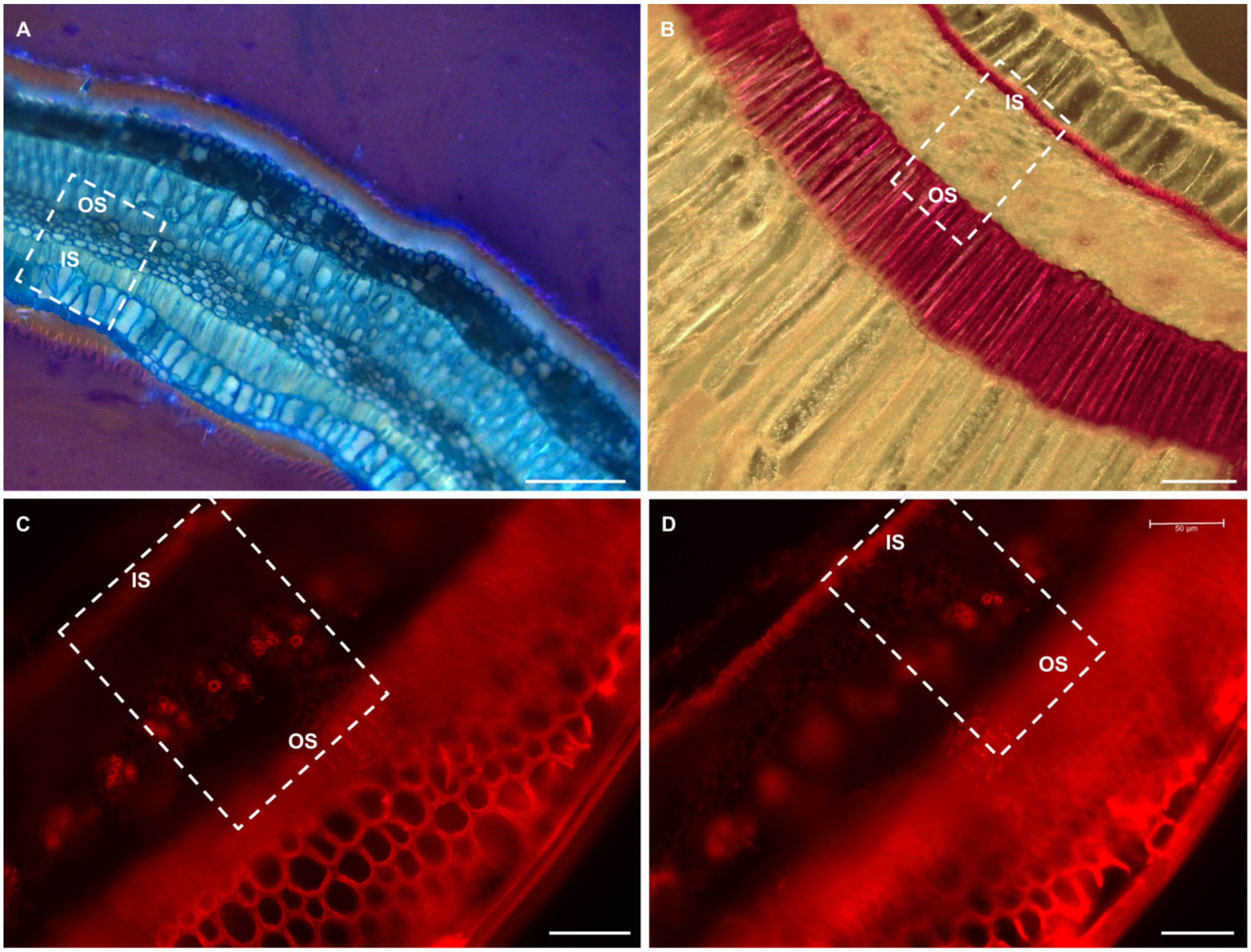
Figure 3: Lignin deposition in cone galls is spatially coordinated in a gall-specific pattern. A: Transverse section of cone gall stained with Weisner stain, showing two heavily lignified cell layers and one cell layer containing bundles of 4-9 highly lignified cells (arrows). Scale bar = 100 μm. B: Darkfield image of tangential longitudinal section of gall stained with Weisner stain, which stains heavily lignified tissue pink. The same two heavily lignified cell layers are visible, as well as the small moderately lignified bundles (arrows), now in longitudinal section. Scale bar = 100 μm. C: Lignin concentration in leaf tissue, early-development cone galls, and mature cone galls, as determined by thioglycolic acid assay. Asterisks indicate $p < 0.05$ by Kruskal-Wallis test with Benjamini-Hochberg correction for multiple comparison, n.s. Indicates $p > 0.05$. D: Lignin subunit S to G ratio as determined by pyro-GC MS. Asterisks indicate $p < 0.05$ by Kruskal-Wallis test with Benjamini-Hochberg correction for multiple comparison, n.s. Indicates $p > 0.05$



Supplemental Figure 9: Sample of additional histological stains. A: Cone gall stained with Mäule stain. Scale bar = 100 μm. B: Cone gall stained with Safranin O. Scale bar = 50 μm C: Cone gall stained with FastGreen. Scale bar = 50 μm D: Unstained cone gall. Scale bar = 50 μm



Supplemental Figure 10: Micrographs of cone galls and map of cell layers A: Transverse section of cone gall with cell layers labeled. There are 7 distinct cell layers, which can be seen by their morphology and differential uptake of stains. Arrows indicate bundles which are moderately lignified. Outer sclerenchyma and inner sclerenchyma are heavily lignified. B: Transverse section of cone gall imaged by autofluorescence using GFP filter. Scale bar = 50 µm C: 12 µm cryosection stained with toluidine blue O. The thick cell walls of the outer sclerenchyma are stained blue, and the bundles are also clearly visible as dark patches in the central sponge layer. Scale bar = 50 µm. D: 12 µm cryosection stained longer with toluidine blue O. The outer epidermis and parenchyma are stained purple, outer and inner sclerenchyma stained blue, and palisade parenchyma and inner epidermis are stained a dark violet. Scale bar = 50 µm



Supplemental Figure 11: Transverse sections of red cone galls. A: Cone gall stained with toluidine blue O. Scale bar = 100 µm (half-sized bar used to indicate 50 µm in Figure 4C). B: Cone gall stained with Weisner stain. Scale bar = 100 µm (half-sized bar used to indicate 50 µm in Figure 4C). C: Transverse section of cone gall immunostained with LM10. D: Transverse section of cone gall immunostained with LM10. Dashed rectangle goes beyond the edge of the micrograph, explaining the dark gray triangle in bottom right of the rightmost image in Figure 4C. Scale bar = 50 µm. For all micrographs, the dotted rectangle shows the portion of the image used in Figure 4C. OS = outer sclerenchyma, IS = inner sclerenchyma.

Weisner staining revealed large amounts of lignin, but histological studies cannot provide an accurate quantification of these chemical changes. To fill this gap, we used the thioglycolic acid assay to quantify lignin in leaf and gall tissue, comparing leaf tissue against young or mature cone galls, as shown in Figure 3C. Cone galls are substantially more lignified than leaf tissue (p=0.0025, Kruskal-Wallis test with Benjamini-Hochberg correction for multiple comparisons), but the two developmental stages are statistically indistinguishable from each other. In light of this substantial increase in lignin levels, we asked whether the lignin monomeric composition was altered as well using pyrolysis gas chromatography coupled to mass spectrometry (pyro-GC MS). Lignin polymers are composed of three subunits, namely syringyl (S), guaiacyl (G), and *p*-hydroxyphenyl

(H), which polymerize with a complex branched structure that is highly resistant to degradation (50, 51). Lignin associated with fiber cells tends to contain a higher fraction of S subunits, whereas vascular elements contain more G subunits (52). The S to G ratio was substantially lower in both stages of gall tissue (p = 0.011 and 0.0076 for early and mature galls respectively, Kruskal-Wallis test with Benjamini-Hochberg correction for multiple comparisons, full data for all lignin-derived fragments available in Dataset S8) compared to leaf tissue, which also supports generation of vasculature in the galls.

To our knowledge, this is the first description of de novo generation of vascular elements in insect galls, and our findings are in contrast to a detailed analysis of another cynipid-induced gall, where de novo production of vasculature was specifically ruled out (53), suggesting neovascularization only occurs in some types of cynipid galls. The importance of obtaining access to the plant vascular system has long been recognized as important for the growth and success of galling insects (54), but previous studies have shown modifications of existing vasculature rather than de novo vascularization (53, 55). In contrast, the histological evidence here demonstrates gall generation involves coordinated, spatially organized generation of de novo vasculature.

3.4.5 Cell wall remodeling is associated with gall formation

The cell wall plays an integral role in defining the form and function of plant cells. Although we had already observed changes in lignin composition and deposition, the majority of the cell wall is composed of polysaccharides (56). Changes in polysaccharide content can drastically alter the biochemical, physical, and ultimately physiological role of plant cells and tissues. A large portion of plant sugars are ultimately sequestered in the cell wall as polysaccharides, and in conjunction with lignin comprise the primary physical support structure of plant organs. Since metabolomics revealed differences in hexose phosphate concentrations, we reasoned that this could lead to changes in the monosaccharide composition of the cell walls. Indeed, cell wall composition varied wildly between galls and the leaf tissue from which they arise, as shown in Figure 4A. Notably, xylose residues were extremely abundant in gall tissue, to the extent that all other monosaccharide signals are largely suppressed, and surprisingly, xylose accounts for over 75 percent of all hydrolyzed cell wall monosaccharides in cone gall samples. It should be noted that the cell wall polysaccharide hydrolysis method employed leaves cellulose intact and measures the monosaccharide composition of all non-cellulosic cell wall polysaccharides.

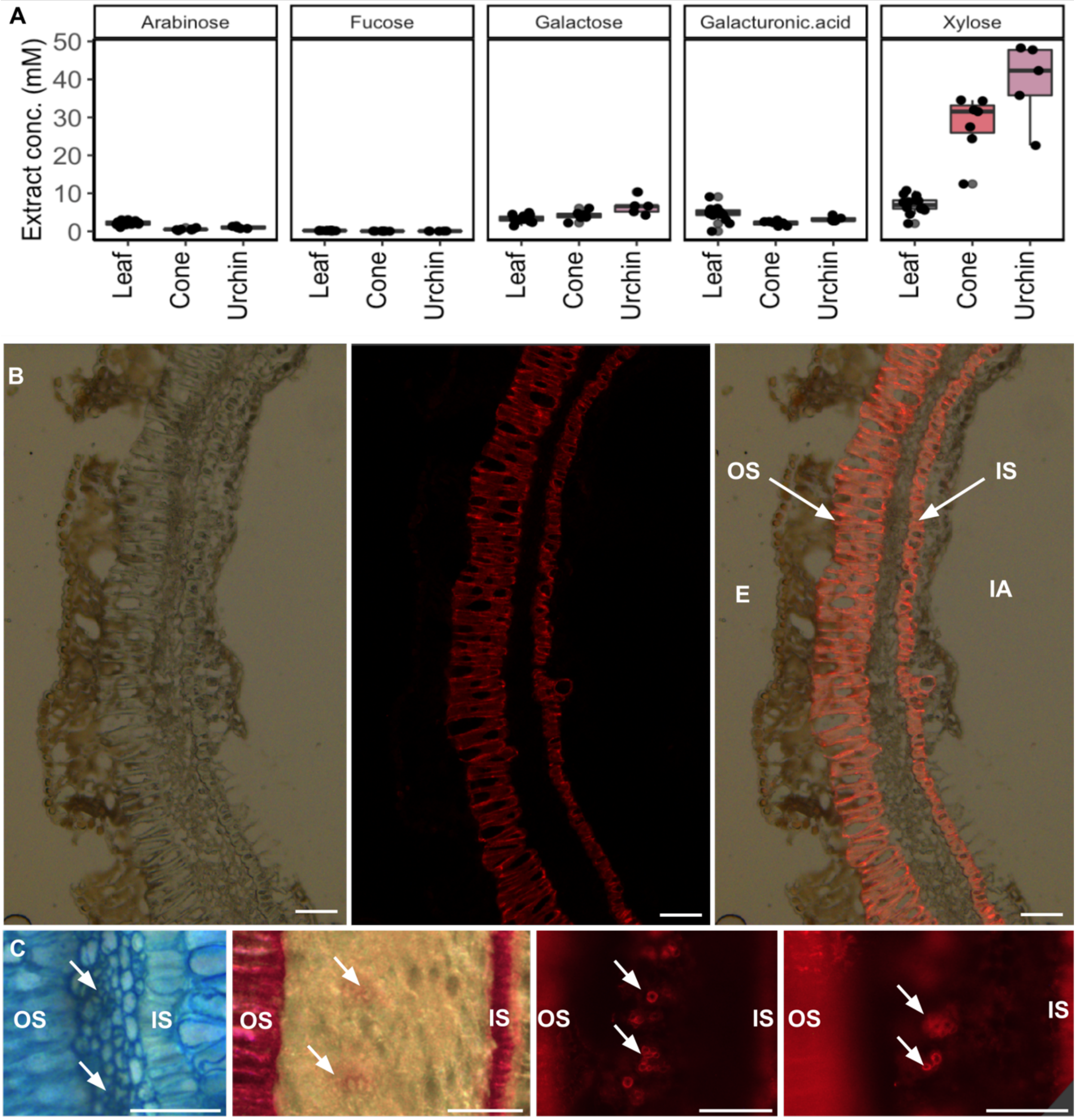


Figure 4: Composition of gall cell walls are altered to be highly enriched in xylan. A: Concentration of five sugars in cell wall residue hydrolysate. Glucose potentially derived from cell wall polymers cannot be accurately measured due to starch contamination. B: LM10 immunofluorescence staining signal for xylan. Left: differential interference contrast transmitted light. Center: Alexa-fluor 647 secondary antibody conjugated to LM10 primary antibody. Right: overlay. OS: outer sclerenchyma, IS: inner sclerenchyma, E: exterior, IA: interior airspace. Scale bar = 50 μm. C: Views of vascular bundles in sponge layer, from left to right: Toluidine Blue O, Weisner stain, LM10, LM10. In each case the outer sclerenchyma cells are shown on the left, sponge layer containing vascular bundles (arrows) in the middle, and inner sclerenchyma on the right

(mostly cropped out in LM10 images due to focus and saturation issues). All scale bars = 50 µm, uncropped source images available in Supplemental Figure 8.

The extraordinarily high levels of xylose suggest enrichment of a polymer composed largely of xylose in gall tissue. One natural candidate is xylan, which is named after and usually enriched in xylem tissue and other vasculature fibers (57, 58). The antibody LM10 selectively binds to and is used to detect xylan. We performed immunofluorescence microscopy with LM10 raised in mice as primary antibodies and anti-mouse IgG conjugated to Alexa-fluor 647 as a secondary antibody as shown in Figure 4B. This revealed high concentrations of xylan in the same two sclerenchyma cell layers which are highly lignified (Figure 3A, B). The colocalization of xylan and lignin deposition is suggestive of a mechanical defense role for these two cell layers. Furthermore, at higher magnification and exposure there were bundles of cells present in the sponge layer between these two, colocalizing with the lignified bundles revealed by Weisner stain (arrows in Figure 3). These bundles as viewed with toluidine blue O stain, Weisner stain, and two views of LM10 immunostain are shown in orientation-matched views in Figure 4C (uncropped source images available in Supplemental Figure 11). The colocalization of lignin and xylan in this particular bundled spatial pattern strongly suggests these are the vascular bundles of the gall, and their less-consistent organization compared to normal vascular bundles likely reflects imperfect control of plant developmental morphology on part of the gall-inducing wasps, as noted previously regarding the alteration of existing vasculature in galls (59). The cell-layer specific alteration of lignin and polysaccharide composition – the two primary constituents of plant cell walls – indicates that galling insects exert a large degree of control over plant growth and metabolism in the development of galls.

3.5 Conclusion

We leveraged metabolomics, three-dimensional light microscopy, lignin composition analysis, histology, and immunomicroscopy to study the biology of galls. In doing so, we revealed many similarities and several key differences in the metabolic and morphological changes associated with the gall-induction process between two types of gall-inducing wasps. We observe dramatic alteration in metabolite composition in two gall types produced from the same tissue of the same host. While many of the changes to the metabolome are consistent across both gall types, some such as abscisic acid and hexose phosphates are strikingly different. The metabolites with consistent increases in concentration (Dataset S7) are candidates for the shared induction mechanism of galls, whereas those with different concentration changes in the two gall types may be responsible for the specific gall morphology. We have further demonstrated that the cell wall lignin and polysaccharide composition of galls differs substantially from the normal plant tissue from which they arise.

We also present several lines of evidence for *de novo* vascularization in cone galls, a surprising finding given the leaf tissue from which the galls derive is terminally differentiated. It has long been known that gall-inducing insects modify and enlarge existing vasculature to deliver nutrients to the gall (28, 53, 54). Galls induced by *Agrobacterium* were long thought to lack vasculature (60) but eventually shown to contain a vascular system organized somewhat differently than that found in normal plant tissue (61, 62). Our finding of *de novo* vascularization in insect-induced galls suggests a similar slow-discovery process may be at play for insect-induced galls. Leafy galls induced by *Rhodococcus fascians* have also been shown to induce neovascularization (63), so while our report is to our knowledge the first describing neovascularization in insect-induced galls, neovascularization is widely accepted to occur in some galls, specifically bacteria-induced galls.

This detailed analysis of the morphological, metabolic, and structural changes found in cynipid galls invites comparison to better understood galls such as the crown gall induced by *Agrobacterium*. The key principle of crown gall induction by *Agrobacterium* is transfer and expression of a relatively short stretch of ‘T-DNA’ that comprises part of the tumor-inducing plasmid (64). This stretch of DNA encodes enzymes in the biosynthetic pathway for auxin (65) and cytokinin (66), which result in altered phytohormone levels and ratios in crown gall (67). The mechanism of gall induction in root knot nematodes is less well understood, but it is notable that the gall-inducing nematodes have been shown to synthesize auxin (68), which suggests synthesis of plant hormones is a common strategy to manipulate plant tissue to expand. Cynipid galls are much more morphologically complex than either of these better-characterized systems, and there is much more diversity in gall size, shape, location, and color. This diversity suggests that the mechanics of gall induction vary between different cynipid wasps, which is supported by our data demonstrating different changes to phytohormones. Nonetheless, the phylogenetic distribution of the galling habit within cynipid wasps suggests it is ancestral, and therefore at least some of the core mechanics are likely to be conserved (32).

The complex and colorful structures of galls have captured the imagination of naturalists for millennia, and demonstrate a mastery of inter-kingdom manipulation that remains unparalleled by current plant molecular biologists. Many practices used to modify and manipulate plants are still reliant on the same techniques adapted from natural plant engineers (i.e., *Agrobacteria*) several decades ago becoming the foundation of plant genetic transformations. Thus, looking for more examples in nature of non-model, non-traditional systems to expand our perspective on the degree to which plants can be reprogrammed may inspire novel approaches to engineering plants in general. Elucidating the molecular basis of the induction of complex galls may provide the blueprint to redefining the landscape to redesigning entirely new cellular, morphological, and physiological architectures in plants.

3.6 Methods

3.6.1 Gall collection

We monitored an arboretum collection of approximately one hundred species of oak trees for galls from spring to autumn. Dozens of gall types were found, of which two types of galls proved to be relatively abundant: the cone gall induced by *Andricus kingi* and the urchin gall induced by *Antron douglasii*, both on the valley oak *Quercus lobata*. Both of these galls were found on the abaxial and adaxial surfaces of leaves between June and August of 2019-2022, with the cone galls being more abundant and appearing somewhat earlier. Both were markedly concentrated in particular trees; one valley oak would often contain hundreds of galls while none could be seen on other valley oaks only a dozen meters away (Supplemental Figure 1). Furthermore, the cone galls in particular were found to cluster on particular branches - it was common to see one branch supporting many times more galls per leaf than an adjacent branch, a somewhat surprising finding given that the gall-inducing insects can fly. Galls were collected in the UC Davis arboretum (38°31'46.0"N 121°45'45.3"W) and Putah Creek Riparian Reserve (38°31'19.7"N 121°46'50.1"W) between May 2019 and September 2020. Initially, all galls identified on oak trees were collected, making use of the 89 species of oak trees in the Peter J Shields collection. Over 20 types of galls were initially collected. The cone gall and urchin gall induced by *Andricus kingii* and *Antron douglasii* on valley oak (*Quercus lobata)* were selected on the basis of morphological complexity and abundance in the study area for further collection and analysis, and 1000-2000 galls of these two species were gathered, at times individually divided into classes on the basis of mass / growth stage, at times in mass collections for large-scale metabolite analysis. Galls were removed from the tree and flash frozen in liquid nitrogen as quickly as possible. Date of collection and specific tree of origin were noted for each gall sample (Dataset S1, Supplemental Figure 1).

3.6.2 Laser Ablation Tomography

Fresh gall samples were sent to LATscan (State College, PA) to perform laser ablation tomography. In brief, samples are attached to a piece of pasta as a sacrificial supporting structure, then mounted in the beam path of a microscope from the front and a high-power flat-beam laser from the side. Rapid alternation of microscope image captures and laser pulses allows for rapid acquisition of several thousand serial ‘slice’ images through the entire sample.

3.6.3 Metabolite extraction

Metabolites were extracted using a protocol adapted from (69). Galls and leaves were flash-frozen in liquid nitrogen and stored at -80 °C until processing. Samples were lyophilized, then disrupted with a steel ball in a ball mill at 30 Hz for 20 minutes, yielding a fine powder. Powder was weighed, then 80 µL of methanol was added per mg. Samples were vortexed for 1 minute, then incubated at room temperature for 20 minutes with continuous mixing, centrifuged at 20,000 g for 5 minutes and the supernatant filtered through 0.45 µm PTFE filters.

3.6.4 Mass spectrometry

In preparation for LC-MS analysis, oak gall extracts were first dried in a SpeedVac (SPD111V, Thermo Scientific, Waltham, MA), then resuspended in 100% MeOH containing an internal standard mix of isotopically labeled compounds (~15 µM average of 5-50 µM of 13C,15N Cell Free Amino Acid Mixture, #767964, Sigma; 10 µg/mL 13C-trehalose, #TRE-002, Omicron; 10 µg/mL 13C-mannitol, ALD-030, Omicron; 2 µg/mL 13C-15N-uracil, CNLM-3917, CIL; 5.5 µg/mL 15N-inosine, NLM-4264, CIL; 4 µg/mL 15N-adenine, NLM-6924, CIL; 3 µg/mL 15N-hypoxanthine, NLM-8500, CIL; 5 µg/mL 13C-15N-cytosine, #294108, Sigma; 2.5 µg/mL 13C-15N-thymine, CNLM-6945, CIL;, 1 µg/mL 2-amino-3-bromo-5-methylbenzoic acid, R435902, Sigma), with resuspension volume of each varied to normalize by biomass for each sample group.

UHPLC normal phase chromatography was performed using an Agilent 1290 LC stack, with MS and MS/MS data collected using a QExactive HF Orbitrap MS (Thermo Scientific, San Jose, CA). Full MS spectra were collected from m/z 70 to 1050 at 60k resolution in both positive and negative ionization mode, with MS/MS fragmentation data acquired using stepped then averaged 10, 20 and 40 eV collision energies at 15,000 resolution. Mass spectrometer source settings included a sheath gas flow rate of 55 (au), auxiliary gas flow of 20 (au), spray voltage of 3 kV (for both positive and negative ionization modes), and capillary temperature or 400 degrees C. Normal phase chromatography was performed using a HILIC column (InfinityLab Poroshell 120 HILIC-Z, 2.1 × 150 mm, 2.7 µm, Agilent, #683775-924) at a flow rate of 0.45 mL/min with a 3 µL injection volume. To detect metabolites, samples were run on the column at 40 °C equilibrated with 100% buffer B (99.8% 95:5 v/v ACN:H2O and 0.2% acetic acid, w/ 5 mM ammonium acetate) for 1 minute, diluting buffer B down to 89% with buffer A (99.8% H2O and 0.2% acetic acid, w/ 5 mM ammonium acetate and 5 µM methylene-di-phosphonic acid) over 10 minutes, down to 70% over 4.75 minutes, down to 20% over 0.5 minutes, and isocratic elution for 2.25 minutes, followed by column re-equilibration by returning to 100% B over 0.1 minute and isocratic elution for 3.9 minutes. Samples consisted of 8 biological replicates each and extraction controls, with sample injection order randomized and an injection blank of 100% MeOH run between each sample.

Metabolite identification was based on exact mass and comparing retention time (RT) and fragmentation spectra to that of standards run using the same LC-MS method. LC-MS data was analyzed using custom Python code (Bowen, B. P. Analysis of Metabolomics Datasets with High-Performance Computing and Metabolite Atlases. 431–442 (2015). doi:10.3390/metabo5030431), with each detected peak assigned a level of confidence, indicated by a score from 0 to 3, in the compound identification. Compounds given a positive identification had matching RT and *m/z* to that of a standard, with detected m/z ≤ 5 ppm or 0.001 Da from theoretical as well as RT ≤ 0.5 minutes. A compound with the highest level of positive identification (score of 3) also had matching MS/MS fragmentation spectra. An identification was invalidated when MS/MS fragmentation spectra collected for the feature did not match that of the standard.

3.6.5 Molecular networking

The LC-MS files were run via MZmine2 version 2.39 workflow to generate a list of features, which were putatively annotated using the Global Natural Products Social Molecular Networking (GNPS) tool (70). This pipeline produced molecular networking files for positive (13918 features) and negative polarities (13562). Filtering accepted features with retention time > 0.6 min (post solvent front), maximum peak height > 1e6, and max peak height fold-change between sample and extraction control > 10, resulting in 8690 and 6305 features in negative and positive mode, respectively. The filtered features were merged into a single molecular network (14995 nodes) created in Cytoscape software version 3.9.1 (70, 71) following step-by-step procedure (72). The average peak height in Leaf control (n=16), Urchin (n=32) and Cone (n=40) galls was calculated and painted on each node as pie charts. This was followed by fold-change calculation between average peak height of Urchin or Cone divided by Leaf value; +1 was added to both numerator and denominator to avoid erroneous division by 0. Annotations with cosine score (MQScore) match to library compounds > 0.7 were (1886 nodes) labeled in the networks. NPClassifier was used to determine metabolite classifications of the annotations (33).

Molecular network of combined HILIC untargeted metabolomics without cosine thresholding:
https://www.ndexbio.org/viewer/networks/02f90a6c-dafd-11ed-b4a3-005056ae23aa
Molecular network of combined HILIC untargeted metabolomics cosine threshold 0.7 organized by mass feature cosine score:
https://www.ndexbio.org/viewer/networks/0d278c0e-dafd-11ed-b4a3-005056ae23aa
Molecular network of combined HILIC untargeted metabolomics cosine threshold 0.7 organized by NP Classifier class: https://www.ndexbio.org/viewer/networks/133ba370-dafd-11ed-b4a3-005056ae23aa

3.6.6 Lignin quantification

Lignin content was measured using the thioglycolic acid (TGA) method following Suzuki *et al* 2009 (73). 1 mL 3N HCl and 0.1 mL TGA were added to 15 mg of biomass. Samples were then incubated at 80 °C for 3 hours, centrifuged for 10 minutes at 16,100 g and the supernatant discarded. 1 mL sterile water was added to the pellet, and vortexed for 30 seconds, and the sample was again centrifuged with the same conditions. 1 mL 1N NaOH was added to the pellet and the sample was allowed to shake at 80 rpm at room temperature for 16 hours, then centrifuged with the same conditions. 1 mL supernatant was transferred to a new tube and 0.2 mL of 12N HCl was added in a fume hood. The samples were then incubated at 4 °C for 4 hours and centrifuged 10 minutes at 16,100 g. The supernatant was discarded and the pellet was dissolved in 1 mL 1N NaOH. Dilutions prepared in 1N NaOH were used to measure absorbance (A_{280}). Lignin concentrations were compared with Wilcoxon rank sum test using the Benjamini-Hochberg method for adjustment for multiple comparisons.

Alcohol-insoluble residue (AIR) preparation

AIR prep was adapted from (74). AIR extracts were prepared by adding ~15 mg of flash-frozen tissue to 1 mL 100% EtOH. The tissue was then ground in a ball mill at 20 Hz for 5 minutes, heated at 100 °C for 30 minutes with periodic shaking, cooled to room temperature and centrifuged at 21,000 g for 5 minutes. The supernatant was discarded and 1 mL 70% EtOH added and vortexed, then centrifuged at 20,000 g for 1 minute. These three steps were repeated until the supernatant was clear, and that clear supernatant discarded. 1 mL of acetone was then added and the samples vortexed, centrifuged at 20,000 g for 5 minutes, supernatant discarded, and the samples dried in a speed-vac overnight. The result was a fine powder which was stored at 4 °C.

Lignin monomeric composition

A small amount (~1 mg) of AIR extract was loaded into a quartz tube for Pyro-GC MS analysis using the methodology adapted from (75). Pyrolysis of biomass was performed with a Pyroprobe 5200 (CDS Analytical Inc., Oxford, PA, USA) connected with GC/MS (Thermo Electron Corporation with Trace GC Ultra and Polaris-Q MS) equipped with an Agilent HP-5MS column (30 m x 0.25 mm inner diameter, 0.25 μm film thickness). The pyrolysis was carried out at 650 °C. The chromatograph was programmed from 50 °C (1 min) to 300 °C at a rate of 20 °C/min; the final temperature was held for 10 min. Helium was used as the carrier gas at a constant flow rate of 1 mL/min. The mass spectrometer was operated in scan mode and the ion source was maintained at 300 °C. The compounds were identified by comparing their mass spectra with those of the NIST library. Peak molar areas were calculated for the lignin degradation products, and the summed areas were normalized.

Trifluoroacetic acid (TFA) hydrolysis

TFA hydrolysis and high-pressure anion exchange chromatography (HPAEC) was adapted from (76). 5-10 mg of AIR was transferred to a new tube using a +- 0.01 mg scale to record transferred mass. 1 mL 2M TFA was added to each sample in a screw-top tube and vortexed. Samples were heated to 120 °C for 1 hour, vortexing for 10 seconds every 15 minutes. After cooling to room temperature, samples were centrifuged at 20,000 g for 1 minute, as much supernatant as possible discarded, and the remainder removed by speed-vac overnight. The dried pellet was dissolved in 1 mL water and shaken at 1000 rpm 30 °C for 1 hour, then filtered through 0.45 μm nitrocellulose filters. Samples were then diluted in water for HPAEC coupled with pulsed amperometric detection. As described in text, several dilution ratios were ultimately required, ranging from 1/10 to 1/640. NaOH was used as needed to bring all samples within the range of 4-9 pH.

3.6.7 Microscopy

Samples were either kept at 4 °C and imaged within 1 week of collection or flash frozen in liquid nitrogen and stored at -80 °C. Sectioning was performed with a vibratome to generate ~50 μm sections or with a cryotome to generate ~12 μm sections. While several methods of sample fixation were performed, the best results were achieved with unfixed samples embedded in 7% agarose for vibratome sectioning or “Optimal cutting temperature” (Sakura Tissue-Tek OCT, part number 4583) cryotomy embedding fluid. For each stain, several concentrations and staining periods were attempted, and the most informative selected for further work. Imaging was performed with a fluorescent microscope also equipped with an RGB camera, all images except for the immunomicroscopy are real-color, with white-balance adjusted as well as possible to match printed images to the image in the eyepiece.

3.6.8 Data analysis

Data analysis was performed with Rstudio (Version 2022.07.0+548 macOS), primarily using the Tidyverse package for data manipulation and ggplot2 for visualization. Figures were assembled with Google Drawings.

3.7 Acknowledgements

This work was supported by the UC Davis Katherine Esau Junior Faculty Fellowship to P.M.S. This work was also conducted as part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy. We thank Denise Snichnes and Steve Ruzin at the UC Berkeley RCNR Berkeley Biological Imaging Facility for advice and assistance with microscopy. We thank LATscan for LAT imaging, Universitas 21 (U21) for funding of LAT. We thank Neelima Sinha and Lynn Kimsey for useful discussions and feedback. The work (proposal:505892 DOI:10.46936/10.25585/60000461)) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.

3.8 Contributions

KM conceived and performed the experiments and wrote the manuscript. LW assisted with bioinformatic analysis. PS supervised and provided funding. All authors have read and approved the manuscript.

3.9 References

Bemer M, van Mourik H, Muiño JM, Ferrándiz C, Kaufmann K, Angenent GC (2017) FRUITFULL controls SAUR10 expression and regulates Arabidopsis growth and architecture. Journal of Experimental Botany **68**: 3391–3403

Bodi Z, Zhong S, Mehra S, Song J, Li H, Graham N, May S, Fray R (2012) Adenosine Methylation in Arabidopsis mRNA is Associated with the 3′ End and Reduced Levels Cause Developmental Defects. Frontiers in Plant Science 3:

Chae K, Isaacs CG, Reeves PH, Maloney GS, Muday GK, Nagpal P, Reed JW (2012) Arabidopsis SMALL AUXIN UP RNA63 promotes hypocotyl and stamen filament elongation. The Plant Journal **71**: 684–697

Chia KF The Arabidopsis transcription factor ERF13 negatively regulates defense against Pseudomonas syringae. M.S. University of California, San Diego, United States -- California

Church C, Moir L, McMurray F, Girard C, Banks GT, Teboul L, Wells S, Brüning JC, Nolan PM, Ashcroft FM, et al (2010) Overexpression of Fto leads to increased food intake and results in obesity. Nat Genet **42**: 1086–1092

Ciftci-Yilmaz S, Morsy MR, Song L, Coutu A, Krizek BA, Lewis MW, Warren D, Cushman J, Connolly EL, Mittler R (2007) The EAR-motif of the Cys2/His2-type Zinc Finger Protein Zat7 Plays a Key Role in the Defense Response of Arabidopsis to Salinity Stress*. Journal of Biological Chemistry **282**: 9260–9268

Duan H-C, Wei L-H, Zhang C, Wang Y, Chen L, Lu Z, Chen PR, He C, Jia G (2017) ALKBH10B Is an RNA N6-Methyladenosine Demethylase Affecting Arabidopsis Floral Transition. Plant Cell **29**: 2995–3011

Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Brüning JC, Rüther U (2009) Inactivation of the Fto gene protects from obesity. Nature **458**: 894–898

Gao X, Shin Y-H, Li M, Wang F, Tong Q, Zhang P (2010) The Fat Mass and Obesity Associated Gene FTO Functions in the Brain to Regulate Postnatal Growth in Mice. PLOS ONE **5**: e14005

Guo T, Liu C, Meng F, Hu L, Fu X, Yang Z, Wang N, Jiang Q, Zhang X, Ma F (2022) The m6A reader MhYTP2 regulates MdMLO19 mRNA stability and antioxidant genes translation efficiency conferring powdery mildew resistance in apple. Plant Biotechnology Journal **20**: 511–525

Huang X, Lu Z, Zhai L, Li N, Yan H (2023) The Small Auxin-Up RNA SAUR10 Is Involved in the Promotion of Seedling Growth in Rice. Plants **12**: 3880

Huot B, Yao J, Montgomery BL, He SY (2014) Growth–Defense Tradeoffs in Plants: A Balancing Act to Optimize Fitness. Molecular Plant **7**: 1267–1287

Jiang L, Li R, Yang J, Yao Z, Cao S (2023) Ethylene response factor ERF022 is involved in regulating Arabidopsis root growth. Plant Mol Biol **113**: 1–17

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J **88**: 1058–1070

Lee S, Park JH, Lee MH, Yu J, Kim SY (2010) Isolation and functional characterization of CE1 binding proteins. BMC Plant Biology **10**: 277

Li K, Huang W, Wang Z, Nie Q (2022) m6A demethylase FTO regulate CTNNB1 to promote adipogenesis of chicken preadipocyte. Journal of Animal Science and Biotechnology **13**: 147

Luo G-Z, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, Liu J, Chen K, Jia G, Bergelson J, et al (2014) Unique features of the m6A methylome in Arabidopsis thaliana. Nat Commun **5**: 5630

Lv B, Wei K, Hu K, Tian T, Zhang F, Yu Z, Zhang D, Su Y, Sang Y, Zhang X, et al (2021) MPK14-mediated auxin signaling controls lateral root development via ERF13-regulated very-long-chain fatty acid biosynthesis. Molecular Plant **14**: 285–297

Merkestein M, Laber S, McMurray F, Andrew D, Sachse G, Sanderson J, Li M, Usher S, Sellayah D, Ashcroft FM, et al (2015) FTO influences adipogenesis by regulating mitotic clonal expansion. Nat Commun **6**: 6792

Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. Nat Protoc **8**: 1551–1566

Nagpal P, Reeves PH, Wong JH, Armengot L, Chae K, Rieveschl NB, Trinidad B, Davidsdottir V, Jain P, Gray WM, et al (2022) SAUR63 stimulates cell growth at the plasma membrane. PLOS Genetics **18**: e1010375

Nowak K, Wójcikowska B, Gaj MD (2015) ERF022 impacts the induction of somatic embryogenesis in Arabidopsis through the ethylene-related pathway. Planta **241**: 967–985

Ohta M, Matsui K, Hiratsu K, Shinshi H, Ohme-Takagi M (2001) Repression Domains of Class II ERF Transcriptional Repressors Share an Essential Motif for Active Repression. The Plant Cell **13**: 1959–1968

Parker MT, Soanes BK, Kusakina J, Larrieu A, Knop K, Joy N, Breidenbach F, Sherwood AV, Barton GJ, Fica SM, et al (2022) m6A modification of U6 snRNA modulates usage of two major classes of pre-mRNA 5' splice site. eLife **11**: e78808

Ren H, Gray WM (2015) SAUR Proteins as Effectors of Hormonal and Environmental Signals in Plant Growth. Mol Plant **8**: 1153–1164

Ronkainen J, Huusko TJ, Soininen R, Mondini E, Cinti F, Mäkelä KA, Kovalainen M, Herzig K-H, Järvelin M-R, Sebert S, et al (2015) Fat mass- and obesity-associated gene Fto affects the dietary response in mouse white adipose tissue. Sci Rep **5**: 9233

Roundtree IA, Evans ME, Pan T, He C (2017) Dynamic RNA Modifications in Gene Expression Regulation. Cell **169**: 1187–1200

Růžička K, Zhang M, Campilho A, Bodi Z, Kashif M, Saleh M, Eeckhout D, El-Showk S, Li H, Zhong S, et al (2017) Identification of factors required for m6A mRNA methylation in Arabidopsis reveals a role for the conserved E3 ubiquitin ligase HAKAI. New Phytologist **215**: 157–172

Sakamoto H, Maruyama K, Sakuma Y, Meshi T, Iwabuchi M, Shinozaki K, Yamaguchi-Shinozaki K (2004) Arabidopsis Cys2/His2-Type Zinc-Finger Proteins Function as Transcription Repressors under Drought, Cold, and High-Salinity Stress Conditions. Plant Physiology **136**: 2734–2746

Shen L, Liang Z, Gu X, Chen Y, Teo ZWN, Hou X, Cai WM, Dedon PC, Liu L, Yu H (2016) N6-Methyladenosine RNA Modification Regulates Shoot Stem Cell Fate in Arabidopsis. Developmental Cell **38**: 186–200

Shinde H, Dudhate A, Kadam US, Hong JC (2023) RNA methylation in plants: An overview. Frontiers in Plant Science 14:

Spartz AK, Lor VS, Ren H, Olszewski NE, Miller ND, Wu G, Spalding EP, Gray WM (2017) Constitutive Expression of Arabidopsis SMALL AUXIN UP RNA19 (SAUR19) in Tomato Confers Auxin-Independent Hypocotyl Elongation. Plant Physiology **173**: 1453–1462

Sun B, Bhati KK, Song P, Edwards A, Petri L, Kruusvee V, Blaakmeer A, Dolde U, Rodrigues V, Straub D, et al (2022) FIONA1-mediated methylation of the 3'UTR of FLC affects FLC transcript levels and flowering in Arabidopsis. PLOS Genetics **18**: e1010386

Wang X, Zhu L, Chen J, Wang Y (2015) mRNA m6A methylation downregulates adipogenesis in porcine adipocytes. Biochemical and Biophysical Research Communications **459**: 201–207

Wang Y, Jia G (2020) Detection methods of epitranscriptomic mark N6-methyladenosine. Essays in Biochemistry **64**: 967–979

Wang Z, Yang L, Liu Z, Lu M, Wang M, Sun Q, Lan Y, Shi T, Wu D, Hua J (2019) Natural variations of growth thermo-responsiveness determined by SAUR26/27/28 proteins in Arabidopsis thaliana. New Phytologist **224**: 291–305

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, et al (2019) Welcome to the Tidyverse. Journal of Open Source Software **4**: 1686

Wong CE, Zhang S, Xu T, Zhang Y, Teo ZWN, Yan A, Shen L, Yu H (2023) Shaping the landscape of N6-methyladenosine RNA methylation in Arabidopsis. Plant Physiology **191**: 2045–2063

Woodworth CM (1931) Breeding for Yield in Crop Plants ^{1}. Agronomy Journal **23**: 388–395

Yu Q, Liu S, Yu L, Xiao Y, Zhang S, Wang X, Xu Y, Yu H, Li Y, Yang J, et al (2021) RNA demethylation increases the yield and biomass of rice and potato plants in field trials. Nat Biotechnol **39**: 1581–1588

Zhao P, Zhang F, Liu D, Imani J, Langen G, Kogel K-H (2017) Matrix metalloproteinases operate redundantly in Arabidopsis immunity against necrotrophic and biotrophic fungal pathogens. PLOS ONE **12**: e0183577

Zheng H, Sun X, Zhang X, Sui N (2020) m6A Editing: New Tool to Improve Crop Quality? Trends in Plant Science **25**: 859–867

Zhong S, Li H, Bodi Z, Button J, Vespa L, Herzog M, Fray RG (2008) MTA Is an Arabidopsis Messenger RNA Adenosine Methylase and Interacts with a Homolog of a Sex-Specific Splicing Factor. Plant Cell **20**: 1278–1288

Zhou A, Kirkpatrick LD, Ornelas IJ, Washington LJ, Hummel NFC, Gee CW, Tang SN, Barnum CR, Scheller HV, Shih PM (2023) A Suite of Constitutive Promoters for Tuning Gene Expression in Plants. ACS Synth Biol **12**: 1533–1545

Zhou L, Tang R, Li X, Tian S, Li B, Qin G (2021) N6-methyladenosine RNA modification regulates strawberry fruit ripening in an ABA-dependent manner. Genome Biol **22**: 168

Chapter 4: FTO, the mammalian enhancer of plant yield

A modified version of this manuscript was preprinted to BioRxiv in February 2024, and another version is currently in review for formal publication. All authors have consented to it being re-published. Section 4.0 is entirely original just for this dissertation.

Kasey Markel, Lucas Waldburger, Patrick M. Shih

4.0 Chapter Preface

While the grand visions inspired by the plant bioengineering achieved by gall wasps remains far beyond the grasp of current plant biotechnologists, there are nonetheless fascinating and important new strategies for improving plant phenotypes being developed every decade. During the course of my PhD, a particularly novel and deeply surprising new strategy was reported: the expression of a mammalian RNA demethylase to alter the morphology of crop plants - most notably, to increase reproductive stem number and thereby to increase the yield. While a less radical change than that induced by gall wasps, this represents a larger structural and phenotypic change to the plants than most previous strategies, and is noteworthy for how unprecedented the strategy was in the literature. To my knowledge, no previous reports in plants have used expression of a RNA modifying enzyme to improve plant phenotypes for agriculture.

Because this report was both quite surprising and very interesting, I immediately set out to begin replication experiments. Early experiments using *Nicotiana benthamiana* proved tantalizing – notably, we found that 10-day-old seedlings expressing the RNA demethylase had larger cotyledons than wild-type plants. This led to increased focus and time-spend on the project, though ultimately we decided for the final paper to focus only on the *Arabidopsis* experiments. This project sits as an uncommonly unclear position on the spectrum between well-established plant biotechnology strategies with decades of historical use and blue-sky sci-fi biotechnology as typified by wasp galls. On the one hand, it is simple transgenic expression of a single enzyme, the sort of process that has been used for decades to enable GM traits like BT-based insect resistance and herbicide tolerance. On the other hand, the phenotypic improvement has no clear connection to the gene, and there is a lot of unknown biology between the sequence of the inserted DNA and the resulting improved phenotype. As a result of the novelty of the project, the promising results, and the key early timing in the global process of discovery, I found this to be the most interesting project of my PhD.

4.1 Abstract

RNA methylation plays a central regulatory role in plant biology and is a relatively new target for plant improvement efforts. In nearly all cases, perturbation of the RNA methylation machinery results in deleterious phenotypes. However, a recent landmark paper reported that transcriptome-wide use of the human RNA demethylase FTO substantially increased the yield of rice and potatoes. Here, we have performed the first

independent replication of those results and broader transferability of the trait, demonstrating increased flower and fruit count in the model species *Arabidopsis thaliana*. We also performed RNA-seq of our FTO-transgenic plants, which we analyzed in conjunction with previously-published datasets to detect several previously-unrecognized patterns in the functional and structural classification of the upregulated and downregulated genes. From these, we present mechanistic hypotheses to explain these surprising results with the goal of spurring more widespread interest in this promising new approach to plant engineering.

4.2 Introduction

RNA methylation plays a central regulatory role in all eukaryotes, and influences RNA processing, translation rate, and transportation (Roundtree et al., 2017). Over 200 types of RNA modification have been identified in plants, of which the most common is N6-methyladenosine (m^{6}A) (Shinde et al., 2023), which is found on approximately 0.5% of adenosines in mRNA (Luo et al., 2014). Most m^{6}A in eukaryotes is found within the consensus motif RRACH (R = A/G, H = A/U/C). Enzymes that add the methyl group to adenosine are often called m^{6}A writers, enzymes that remove the methyl group are called m^{6}A erasers, and proteins that recognize the methyl group are called m^{6}A readers. Many plant traits are known to be affected by m^{6}A status, including resistance to fungal pathogens (Guo et al., 2022) and fruit ripening (Zhou et al., 2021). Targeted m^{6}A editing has been proposed as a method to improve crop quality, with specific genes proposed to target flowering time and fruit maturation (Zheng et al., 2020).

Recently, Yu *et al.* reported that transgenic expression of the mammalian m^{6}A eraser FTO substantially increased yield in both potatoes (*Solanum tuberosum*) and rice (*Oryza sativa*) (Yu et al., 2021). They reported an approximately 50% increase in yield in the field, a remarkably large effect size for a single transgene. In addition to the increase in yield, Yu *et al.* reported an increase in shoot biomass and polyadenylated mRNA accumulation. The yield increase was explained through an increase in tillering, which suggests reduced apical dominance and increased branching.

FTO overexpression has been shown to increase biomass across a wide range of animal species. FTO has been intensely studied in humans because it is the QTL most strongly associated with obesity (Gao et al., 2010; Merkestein et al., 2015). Overexpression of FTO in mice leads to increased food intake and obesity (Church et al., 2010), whereas *fto* mice display reduced growth and lower body mass (Fischer et al., 2009; Gao et al., 2010). Overexpression of FTO also increases the mitosis rates of mouse fat cells *in vitro* (Merkestein et al., 2015), and *fto* mutant mice gain less weight even when forced to consume the same amount of food and water as wild-type mice (Ronkainen et al., 2015). In porcine cell lines, FTO overexpression results in higher accumulation of lipids while *fto* knockouts accumulate lower levels (Wang et al., 2015), and in chicken cells FTO overexpression increases adipogenesis while FTO silencing reduces adipogenesis (Li et al., 2022).

In contrast, there were no reports prior to Yu *et al.* that reduction in m^{6}A levels results in increased biomass in plants, though many mutants with altered m^{6}A levels have been characterized in the last two decades. In *Arabidopsis*, known m^{6}A writers include MTA, MTB, FIP37, VIR, FIONA1, and HAKAI. Complete knockout of MTA is embryo-lethal (Zhong et al., 2008), but plants rescued with embryo-specific expression display reduced apical dominance and an increase in trichome branch number (Bodi et al., 2012). Silencing of MTA with artificial microRNA results in overproliferation of shoot apical meristems (SAMs) (Shen et al., 2016). Another study found artificial microRNA interference against MTA or VIR resulted in expansion of the SAM onto the petioles of young leaves (Wong et al., 2023). RNA interference against MTB also resulted in a reduction in apical dominance and partial dwarfism (Růžička et al., 2017). *Fip37* knockout lines have reduced apical dominance and severe dwarfism (Růžička et al., 2017), and rescue with embryo-specific expression resulted in SAM overproliferation (Shen et al., 2016). *Fiona1* mutants have altered RNA splicing patterns (Parker et al., 2022), early flowering, and more branching in the floral stem (Sun et al., 2022). Knockout of the m^{6}A eraser ALKBH10B resulted in increased m^{6}A, late flowering, and reduced growth, whereas constitutive overexpression resulted in reduced m^{6}A and early flowering (Duan et al., 2017). While the details vary depending on which gene is perturbed, it is generally the case that reduction of genome-wide m^{6}A levels in plants results in reduced apical dominance, earlier flowering, and enlarged meristems. Taken together, there are several lines of evidence suggesting how controlling RNA methylation patterns may provide a means to tune agriculturally relevant traits.

Increasing yield potential through genetic improvement is a major goal of plant biology (Woodworth, 1931), and despite massive progress to date, more remains to be accomplished. Given the dramatic nature of the yield increase reported by Yu *et al.*, understanding the transferability of this approach into other plants has been of keen interest. Moreover, understanding how RNA methylation plays a role in this change will help elucidate new avenues in crop improvement. Here, we investigated transgenic expression of FTO in *Arabidopsis* and performed RNA-seq to generate hypotheses as to the underlying mechanism.

4.3 Results

4.3.1 Overexpression of FTO increases many yield-associated traits in *Arabidopsis*

We synthesized an *Arabidopsis* codon-optimized version of the human FTO gene under the control of the pCH3 promoter (At4G13930), which provides high constitutive expression (Zhou et al., 2023), in a binary plasmid as shown in Figure 1A. We generated *Arabidopsis* stable lines using floral dip and selecting for kanamycin resistance, then grew the plants for two generations and selected for lines with approximately 100% survival on kanamycin, indicating either multiple insertion or homozygosity for FTO. We grew the T3-generation FTO plants and found that they bolted earlier than Col-0 (Figure 1B). Plants were grown an additional 4 days after imaging to allow for maturation and for nearly all plants to bolt, then the plants were destructively phenotyped. Only plants that had bolted (defined as having a floral stem weighing over 100 mg and at least one flower) were included, for a total of 188 plants. Phenotyping consisted of counting the flowers and fruits (siliques), counting the number of shoot tips, and weighing the rosette and floral stems; raw data are available in Supplemental Table 1.

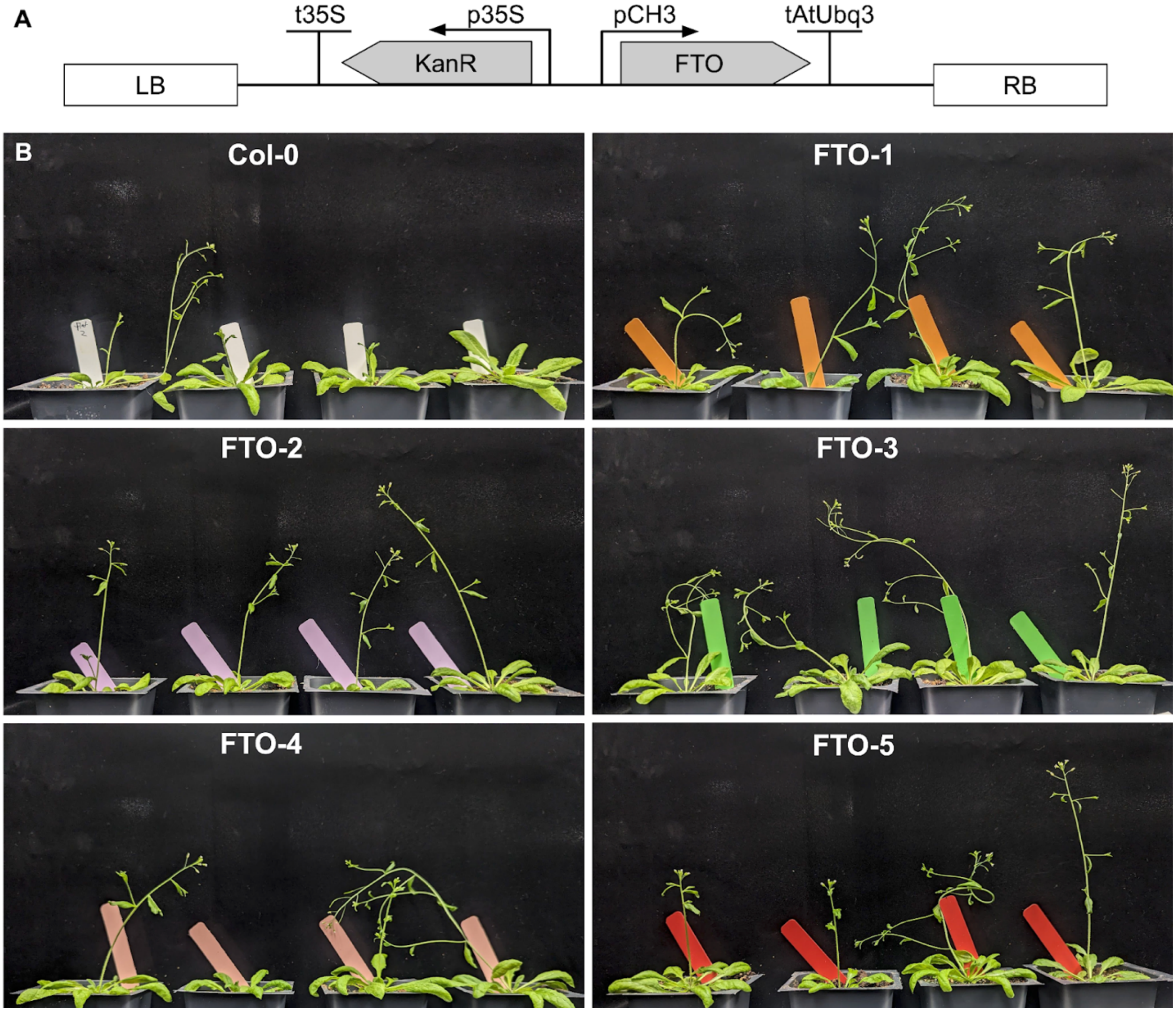


Figure 1: Transgenic expression of FTO causes early bolting in *Arabidopsis*. A) Genetic design of the synthetic FTO construct. B) Transgenic plants of the T3 generation. All plants were grown together and randomly-selected for imaging on the same day from one of the ten flats used in this experiment with no cherry picking.

Though Yu *et al.* claimed a significant increase in stem biomass for both potatoes and rice, we found no difference in *Arabidopsis* rosette mass between FTO plants and Col-0 wild-type (Figure 2A). However, the floral stem mass was significantly increased for all five independent FTO lines (Figure 2B). There was no change in total shoot mass (Figure 2C), suggesting the key difference between FTO and Col-0 plants was the allocation of growth towards sexual rather than vegetative tissue. Accordingly, there was a significant increase in percentage of stem mass apportioned in floral stem tissue (Figure 2D) for four out of five lines. All FTO lines had a larger number of flowers and siliques than Col-0, with an average increase of 100% (Figure 2E). Though *a priori* surprising, this result aligns well with Yu *et al*.'s finding of a ~200% increase in yield in rice in the greenhouse and ~50% in the field. Our result in a different species in growth chambers is within the range of their observed increases in yield. The effect was statistically significant for all lines, with a maximum p-value of 0.015 (Wilcoxon test with Benjamini-Hochberg correction for multiple comparisons).

Since Yu *et al*. identify an approximately 40% increase in the tiller number as the primary driver of increased yield, we quantified the number of floral stems per plant and the number of floral branch stem points. The total number of floral stem branches increased by an average of 62% (Figure 2F). There was also a significant increase in the number of floral stem branches (Figure 2G), but the number of branches per main floral stem remained constant (Figure 2H). This indicates that the effect was driven by the increase in the number of main stems. The total number of floral stem shoot apical meristems – the sum of main stems and stem branch points – also showed a significant increase in all five lines (Figure 2I).

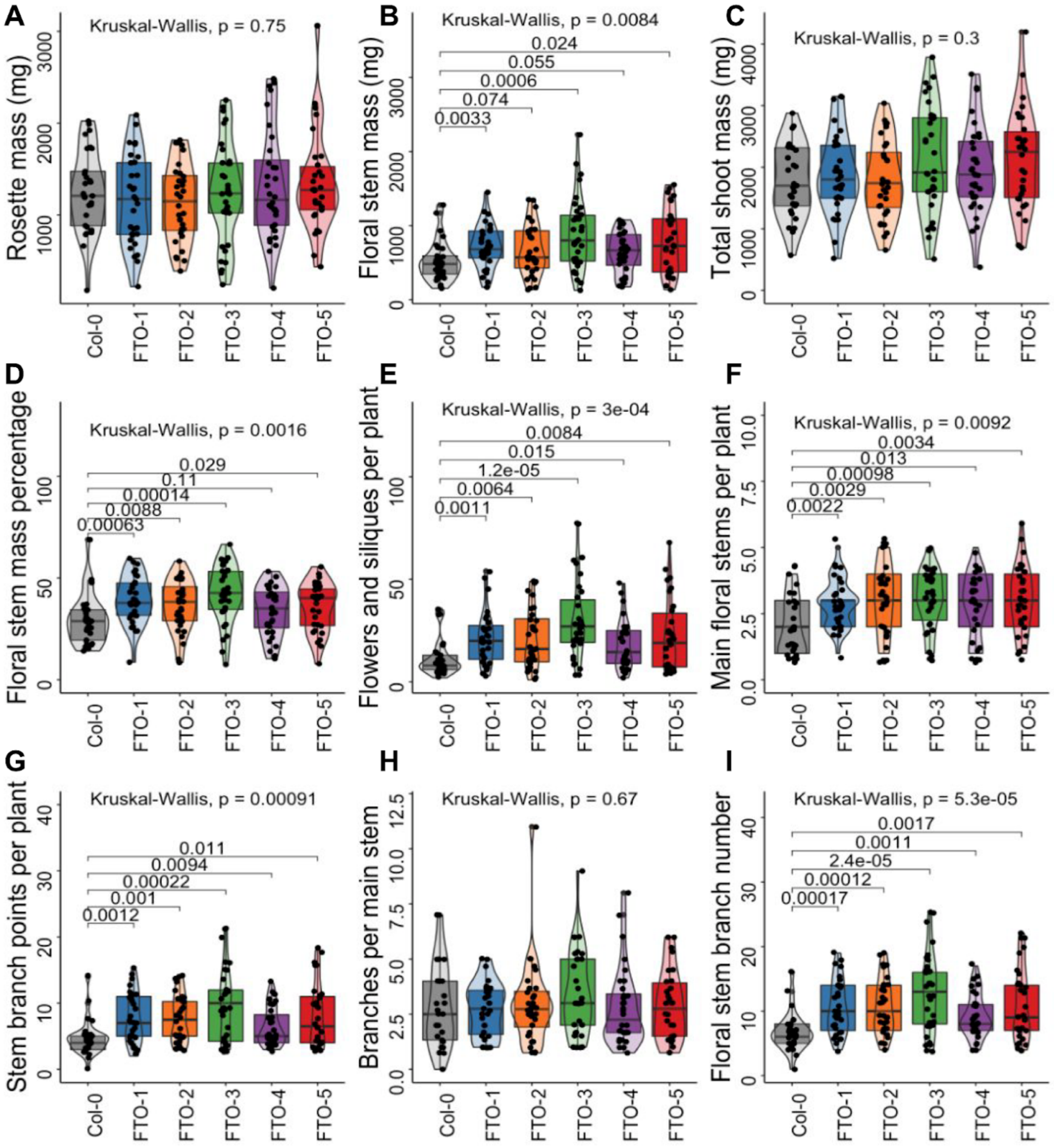


Figure 2: FTO increases many yield-associated traits in *Arabidopsis*. A) Rosette mass per plant. B) Floral stem mass per plant. C) Total shoot mass per plant. D) Floral stem mass percentage per plant. E) Number of flowers and siliques (fruits) per plant. F) Main floral stems per plant. G) Stem branch points per plant. H) Stem branch points per main stem. I) Floral stem branch number per plant. Boxplot central boxes cover the two central quartiles, points are raw data. Kruskal-Wallis p-value tests for significant difference between any of the groups, if Kruskal-Wallis result was significant then Wilcoxon test was performed between Col-0 and all FTO lines with Benjamini-Hochberg correction for multiple comparisons, pairwise p-values are displayed.

4.3.2 FTO expression alters global gene expression patterns

We performed RNA-seq using RNA extracted from seedling shoot tissue to generate mechanistic hypotheses as to why transgenic expression of FTO might have these surprising effects. To ensure robustness of data with respect to experimental variability, we opted for six biological replicates, which enabled us to sequence Col-0 and three FTO independent lines, twenty-four samples in total. An average of 34 million reads were mapped per sample, gene-by-gene count data is available in Supplemental Table 2, raw sequencing data are available through the JGI portal (link in methods). In comparisons to Col-0, we found FTO-1 has far fewer differentially expressed genes than FTO-2 and FTO-3, both upregulated (Figure 3A) and downregulated (Figure 3B). Interestingly, FTO-1 also had a smaller phenotypic difference compared to Col-0 in floral stems per plant (Figure 2F). Given the smaller molecular and phenotypic effects observed for FTO-1, we decided to focus our analysis on the genes differentially upregulated in either all three lines or FTO-2 and FTO-3, leading us to 1436 upregulated and 1014 downregulated genes, loci and additional details of which are included in Supplemental Table 3.

We performed Gene Ontology (GO) term enrichment analysis using the PANTHER (Mi et al., 2013) tool through TAIR, with selected GO categories shown in Figure 3C and Figure 3D for the upregulated and downregulated gene lists, respectively. The full list of significantly enriched or depleted GO categories along with their enrichment, number of genes, and significance can be found in Supplemental Table 4 and Supplemental Table 5 for upregulated and downregulated gene lists. For the upregulated gene list, the two most enriched categories are ‘protein maturation’ and ‘response to light intensity’. The next two upregulated categories both relate to porphyrin production, suggesting a change in chlorophyll metabolism. Several growth-related GO categories were also enriched, such as ‘peptide biosynthetic process’ and ‘macromolecule biosynthetic process’. ‘Defense response’ is depleted, suggesting the plants are growing at full speed, without regard to potential pathogen attack.

For the downregulated gene list, the most highly enriched GO categories were associated with organellar electron transport chain activity, such as ‘proton motive force-driven mitochondrial ATP synthesis’ and ‘ATP biosynthetic process’. Many defense related GO categories were enriched among the downregulated genes, such as ‘response to nematode’, ‘defense response to fungus’, and ‘response to bacterium’. Several GO terms for negative regulation of growth such as ‘negative regulation of biosynthetic process’ and ‘negative regulation of cellular biosynthetic process’ were also enriched, suggesting that FTO expression caused the plants to “turn off the brakes” in their growth process. Both the upregulated and downregulated gene lists were enriched for GO terms related to RNA metabolism, such as ‘RNA metabolic process’ and ‘mRNA metabolic process’, which is not surprising given the alteration to mRNA regulation reported in previous FTO expression papers.

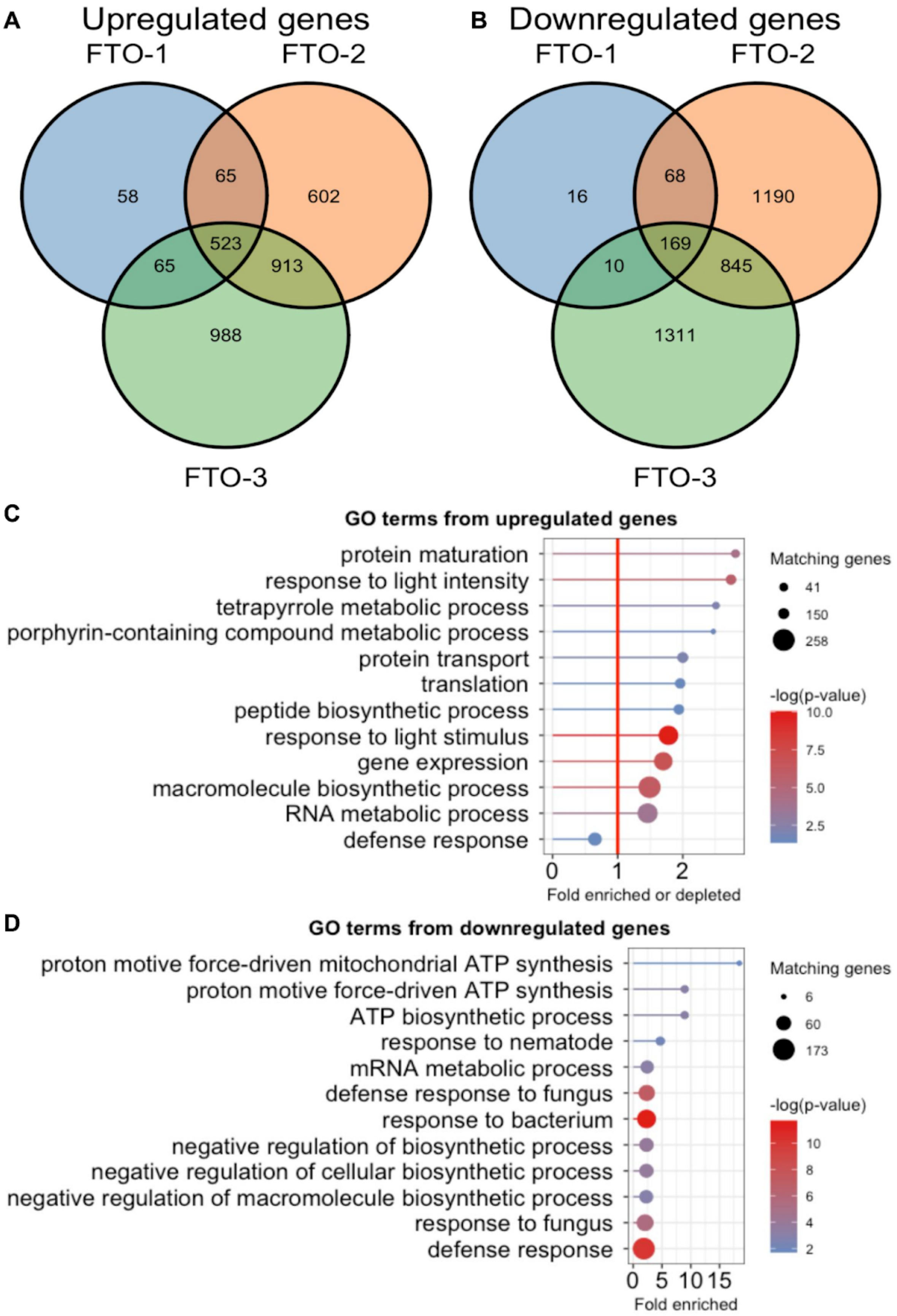

A
Upregulated genes
FTO-1
FTO-2
FTO-3
58
65
602
523
65
913
988
B
Downregulated genes
FTO-1
FTO-2
FTO-3
16
68
1190
169
10
845
1311
C
GO terms from upregulated genes
protein maturation
response to light intensity
tetrapyrrole metabolic process
porphyrin-containing compound metabolic process
protein transport
translation
peptide biosynthetic process
response to light stimulus
gene expression
macromolecule biosynthetic process
RNA metabolic process
defense response
0
1
2
Fold enriched or depleted
Matching genes
41
150
258
-log(p-value)
10.0
7.5
5.0
2.5
D
GO terms from downregulated genes
proton motive force-driven mitochondrial ATP synthesis
proton motive force-driven ATP synthesis
ATP biosynthetic process
response to nematode
mRNA metabolic process
defense response to fungus
response to bacterium
negative regulation of biosynthetic process
negative regulation of cellular biosynthetic process
negative regulation of macromolecule biosynthetic process
response to fungus
defense response
0
5
10
15
Fold enriched
Matching genes
6
60
173
-log(p-value)
10
8
6
4
2

Figure 3: FTO expression alters global gene expression. A) Venn diagram of significantly upregulated genes in our three transgenic lines compared to Col-0. Multiple-comparison adjusted p-value threshold = 0.01. B) Venn diagram of significantly downregulated genes in our three transgenic lines compared to Col-0. Multiple-comparison adjusted p-value threshold = 0.01. C) Lollipop plot of Gene Ontology (GO) terms enriched among the upregulated genes. Dot size indicates number of genes, line length indicates fold enrichment, color indicates statistical significance. GO categories with the circle to the right of the red vertical line are enriched, those to the left of the line are depleted. D) Lollipop plot of GO terms enriched among the downregulated genes. Dot size indicates number of genes, line length indicates fold enrichment, color indicates statistical significance. All GO categories are enriched, the lowest enrichment is 1.85x.

4.3.3 Growth-associated SAURs are overrepresented among the most highly upregulated genes

We identified the top 100 most upregulated and downregulated genes in each line to gain mechanistic insight into the function of FTO expression relative to Col-0. Similar to the broader search, we found there was more overlap between FTO-2 and FTO-3, and therefore chose to look at genes among the top 100 most-altered by fold change in either all three lines or just those two. This analysis yielded 41 shared most-upregulated genes and 36 shared most-downregulated genes (Supplemental Table 6). For both the highly upregulated and downregulated genes, there is much more overlap between lines than would be expected by chance (hypergeometric p-value = 2.53e-76 and 3.79e-64, respectively), confirming that the most highly-altered genes among all lines are strongly shared.

Among the 41 highly upregulated genes, 4 are SAUR genes which are known to regulate auxin-induced growth and development (Ren and Gray, 2015). This is approximately 32-fold higher than expected based on frequency in the genome (hypergeometric p-value = 7e-06). Overexpression of SAUR genes increases plant growth (Chae et al., 2012; Spartz et al., 2017), and microRNA silencing reduces growth (Chae et al., 2012). Specifically, we found SAUR10, SAUR28, SAUR63, and SAUR68 to be highly upregulated. SAUR10 has been shown to increase growth when overexpressed in *Arabidopsis* (Bemer et al., 2017), and was recently shown to be essential for normal growth and yield in rice (Huang et al., 2023). SAUR28, along with three other nearby SAUR genes, was identified as the largest effect QTL for leaf architecture changes at bolting stage in response to different temperatures (Wang et al., 2019). SAUR63 overexpression has also been shown to increase growth in *Arabidopsis* seedlings (Nagpal et al., 2022). The increased expression of SAUR genes is particularly interesting in light of the reduced apical dominance and increased branching we observed and suggests that auxin regulation may be altered in FTO-expressing plants.

4.3.4 Organellar genes and repressive transcription factors are overrepresented among the most highly downregulated genes

Among the 36 highly downregulated genes, we found 6 on the plastid and 6 on the mitochondrial genome, a ~37-fold and ~78-fold overrepresentation, respectively (hypergeometric p-values = 1.2e-08 and 1.3e-10). In contrast, 0 of the 41 highly upregulated genes were encoded in an organellar genome. Furthermore, among our 1436 significantly upregulated genes there were 0 mitochondrial and 0 plastid genes, whereas there were 21 mitochondrial and 34 plastid genes among the 1014 downregulated genes. This explains the enrichment of three different organellar GO terms - two variants of ‘proton motive force-driven ATP synthesis’ and ‘ATP biosynthetic process’ among our downregulated genes. Yu *et al*.’s claim that FTO expression increases nuclear genome-derived mRNA production offers one explanation, because one would expect to see a relative decrease in the abundance of mRNAs derived from organellar genomes. Yu *et al.* do not report RNA-seq data for organellar genes, and it should be noted that our library prep process using polyA selection should cause significant depletion of organellar compared to nuclear genome-derived transcripts, though there is no reason to expect this would differentially affect upregulated versus downregulated genes.

We also noticed several of the highly downregulated genes were transcription factors with known repression activity, including ERF13 (Lv et al., 2021), ERF22 (Nowak et al., 2015), ZAT7 (Ciftci-Yilmaz et al., 2007) and ZAT10 (Ohta et al., 2001). Interestingly, *erf22* mutants have approximately 5-fold increased expression of SAUR10 (Jiang et al., 2023), which concords well with our finding that FTO plants with highly reduced expression of ERF22 also have highly increased expression of SAUR10. Overexpression of ERF13 (Chia), ZAT7 (Ciftci-Yilmaz et al., 2007) and ZAT10 (Sakamoto et al., 2004) inhibit growth in *Arabidopsis*, and loss-of-function mutants of *erf22* (Jiang et al., 2023) and RNAi-silenced ERF13 (Lee et al., 2010) lines have enhanced growth rates. Taken together, these growth-reducing genes among the highly downregulated genes provide a plausible mechanism for the growth-enhancing effects of FTO. Downregulation of transcriptional repressors is also consistent with the claim by Yu *et al*. of an overall increase in nuclear-derived mRNA.

4.3.5 Genes expressed in senescent tissue and genes associated with defense are overrepresented among the genes downregulated by FTO

As another method of exploratory data analysis, we used the Klepikova atlas (Klepikova et al., 2016) through TAIR to identify the tissues in which each gene was most highly expressed. Of the 41 highly upregulated genes, 2 were most highly expressed in a senescent tissue (senescent leaf petiole and senescent leaf vein). In contrast, 11 of 36 highly downregulated genes were most highly expressed in a senescent tissue (all in senescent leaf petiole), a significant difference (Fisher’s exact test two-tailed p-value = 0.0046). Interestingly, all four repressive transcription factors were most highly expressed in senescent leaf petioles. It’s possible the downregulation of many genes expressed in senescent tissue is connected to the growth-enhancing phenotype.

We also collated GO terms for all genes, and noticed that 0 of 41 highly upregulated genes and 4 of 36 highly downregulated genes included ‘pathogen’ or ‘fungus’ in their

GO terms, a significant difference (Fisher's exact test two-tailed p-value = 0.044). Mutants of one of these genes, AT3-MMP, have been shown to be more vulnerable to the fungus *Botrytis cinerea* (Zhao et al., 2017). This supports our PANTHER GO-term analysis which found 'defense response' to be depleted among the upregulated genes (Figure 3C), and found the categories 'defense response to bacterium', 'regulation of defense response', and 'defense response to fungus' to all be enriched at least 2.3-fold among the downregulated genes (Supplemental Table 5). This reduction in expression level of pathogen-responsive genes is to be expected in light of long-established tradeoff between growth and defense (Huot et al., 2014).

4.3.6 Genes upregulated and downregulated by FTO expression differ in sequence properties

In general, RNA-seq data is best understood by considering genes and transcripts as units of biological function, which lends itself to methods of analysis like gene ontology and comparison of differentially expressed genes against other studies of the same gene. However, transcripts are also units of biological structure, and differences in structure such as length or nucleotide composition may be differentially affected by FTO expression. As such, we plotted the average nucleotide frequency within each transcript for three sets of genes: our 1436 upregulated genes, 1014 downregulated genes, and a set of 1200 genes chosen at random from our RNA-seq counts table as a reference point. The gene lists and sequences are available in Supplemental Table 3. We found the frequency of adenosine was significantly higher among our upregulated genes compared to both random and downregulated gene lists, with the majority of the gap being between upregulated and the other two categories (Figure 4A). Because m^{6}A modification is primarily destabilizing, it stands to reason that with a transcriptome-wide reduction of m^{6}A, we would expect to see adenosine-rich transcripts more often upregulated.

We next tested whether the length of transcripts differed in length between our three categories, and found the downregulated transcripts were much longer than either the random or upregulated transcripts: median length 90% and 81% longer, respectively (Figure 4B). FTO-mediated removal of m^{6}A and destabilization of those transcripts would cause an over-representation of longer genes among the downregulated gene list. Because m^{6}A is known to localize to the motif RRACH (R = A/G, H = A/U/C), we next examined the frequency of that motif within each category of transcripts. Because transcript length differed, we normalized by length, and found that downregulated genes had fewer RRACH motifs per kb (Figure 4C). With the naïve assumptions of equal base frequency and random sequence, the RRACH motif occurs about 12 times per kb, among transcribed genes in *Arabidopsis* we found a slightly higher but comparable frequency. If there is a transcriptome-wide reduction in a destabilizing epitranscriptomic mark at RRACH motifs, we would expect relative enrichment of transcripts with a high concentration of those motifs among the upregulated genes and a depletion in the downregulated genes, which we do.

We also tested the adenosine frequency, length, and RRACH motif concentration in the 5’ and 3’ UTRs and found the differences were either much smaller or nonexistent (Supplemental Figure 1), it seems to be the CDS in particular that displays these striking sequence-level differences between our different categories of transcript. These differences also don’t seem to be a result of average differences in sequence composition between different functional classes of genes as captured by GO categories - we tested the adenosine frequency, length, and RRACH motif density for all arabidopsis genes of three major GO categories differentially represented in our dataset, ‘growth’, ‘response to biological stimulus’, and ‘response to light’. We found minimal differences in these metrics between GO categories, suggesting the differences found here are due to the peculiarities of the effects of FTO expression, not simply a change in the abundance of genes of a particular GO category (Supplemental Figure 2).

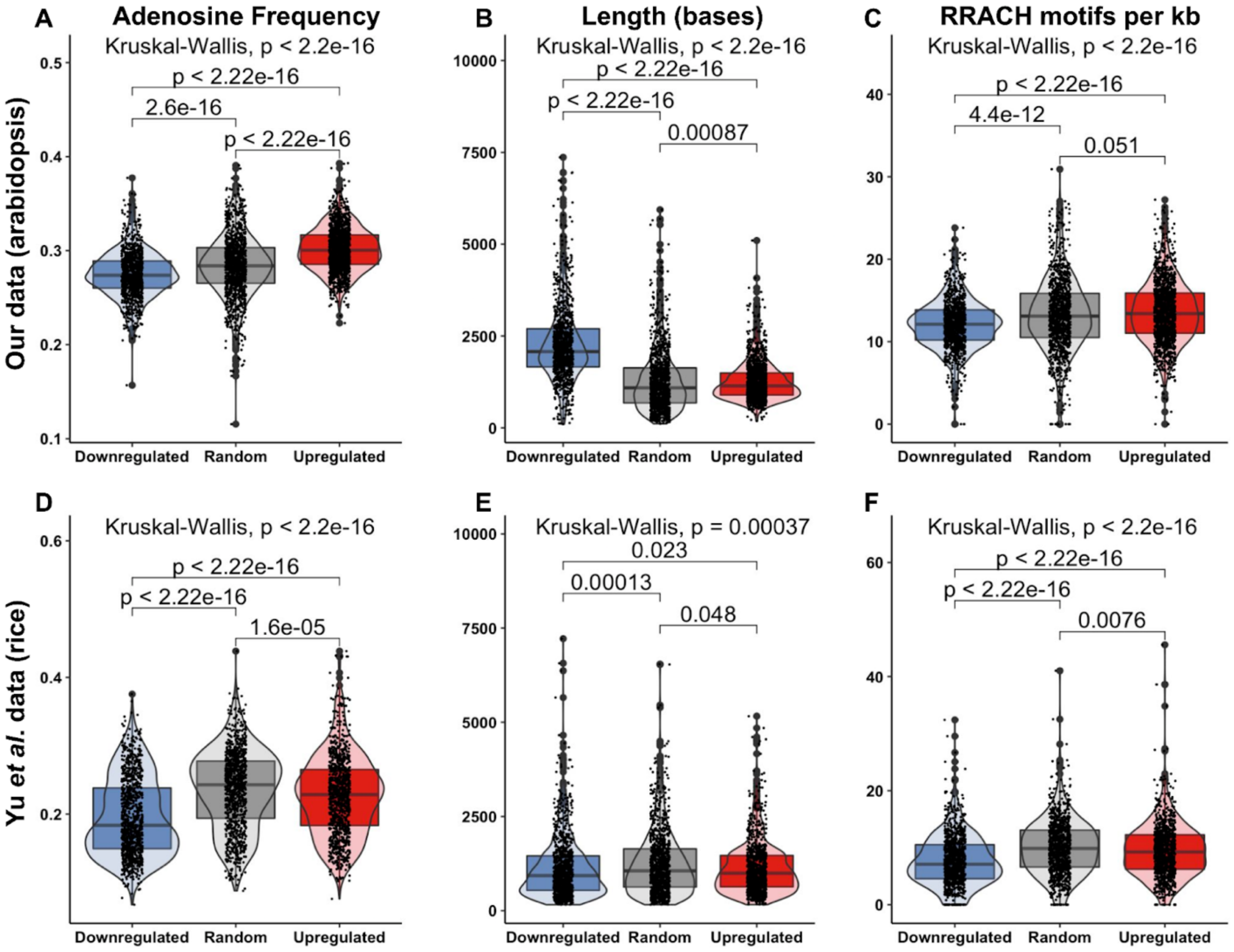
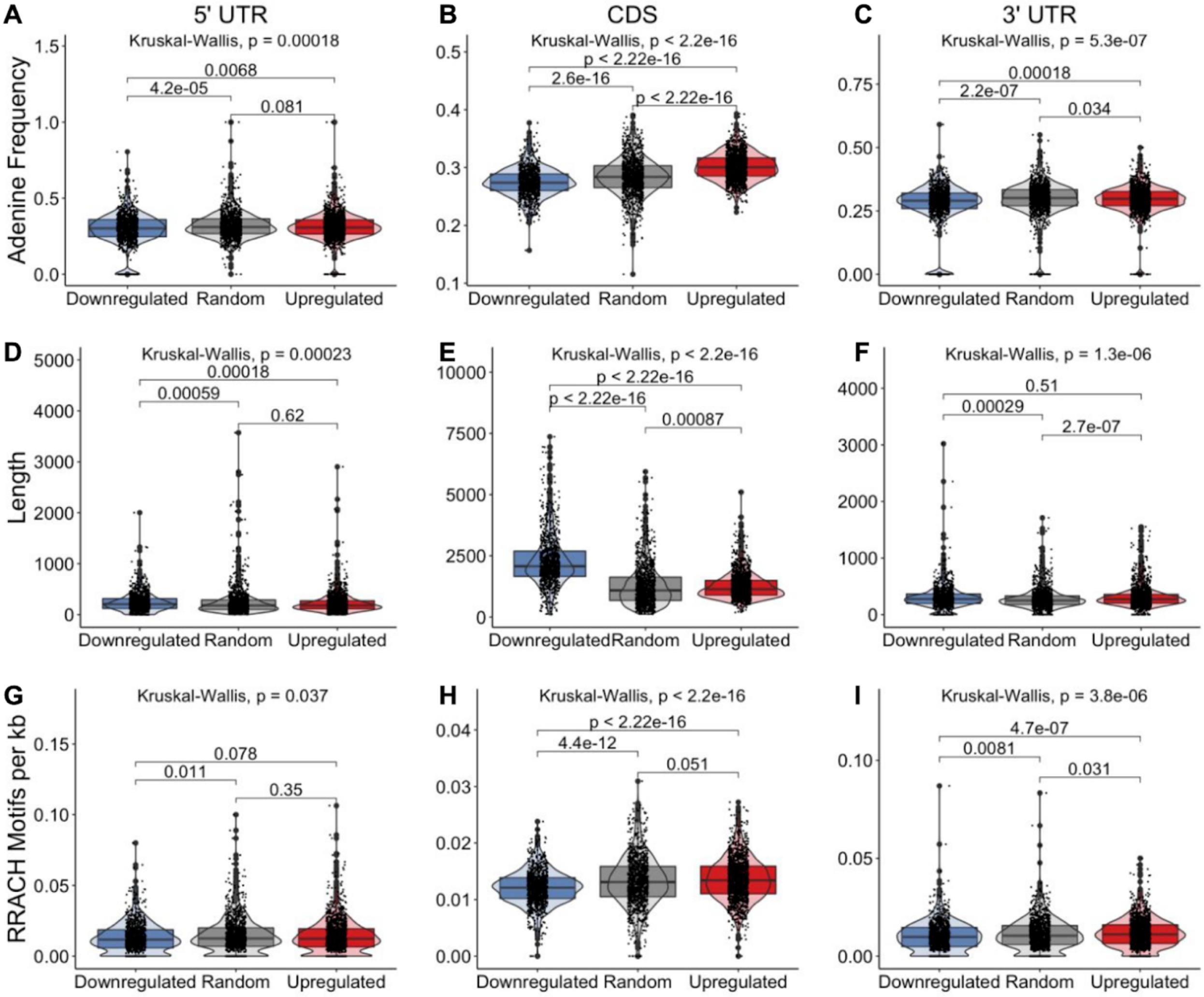
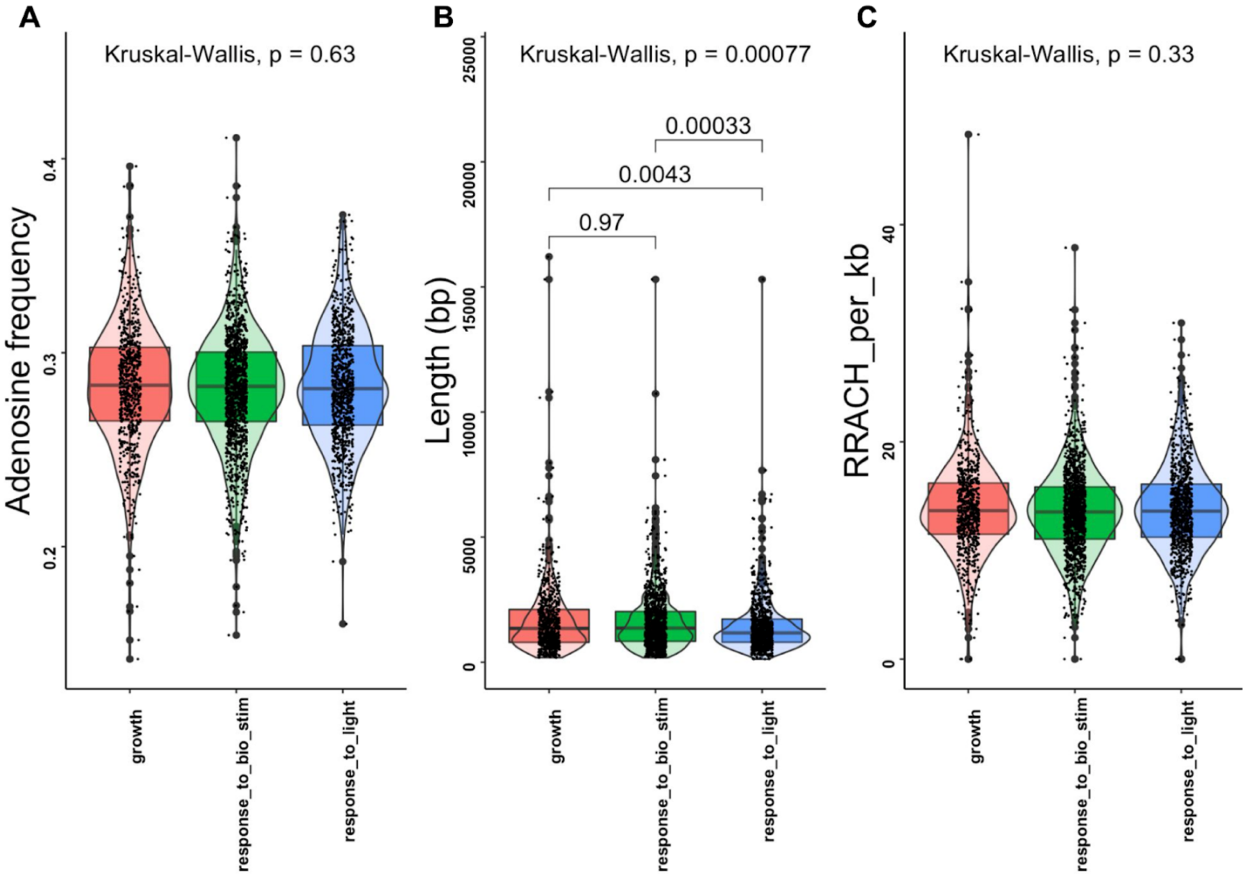


Figure 4: CDS composition of genes upregulated and downregulated by FTO expression are physically distinct. A) Frequency of adenosine in the CDS of transcripts downregulated and upregulated by FTO expression in our dataset. B) Length of the CDS of transcripts downregulated and upregulated by FTO expression in our dataset. C) Frequency of the RRACH (R = A/G, H = A/U/C) motif in the CDS of transcripts downregulated and upregulated by FTO expression in our dataset. D)

Frequency of adenosine in the CDS of transcripts downregulated and upregulated by FTO expression in Yu *et al*.'s dataset. E) Length of the CDS of transcripts downregulated and upregulated by FTO expression in Yu *et al*.'s dataset. F) Frequency of the RRACH (R = A/G, H = A/U/C) motif in the CDS of transcripts downregulated and upregulated by FTO expression in Yu *et al*.'s dataset. Boxplot central boxes cover the two central quartiles, points are raw data. Kruskal-Wallis p-value tests for significant difference between any of the groups, if Kruskal-Wallis result was significant then Wilcoxon test was performed between all three categories of transcript with Benjamini-Hochberg correction for multiple comparisons, pairwise p-values displayed.



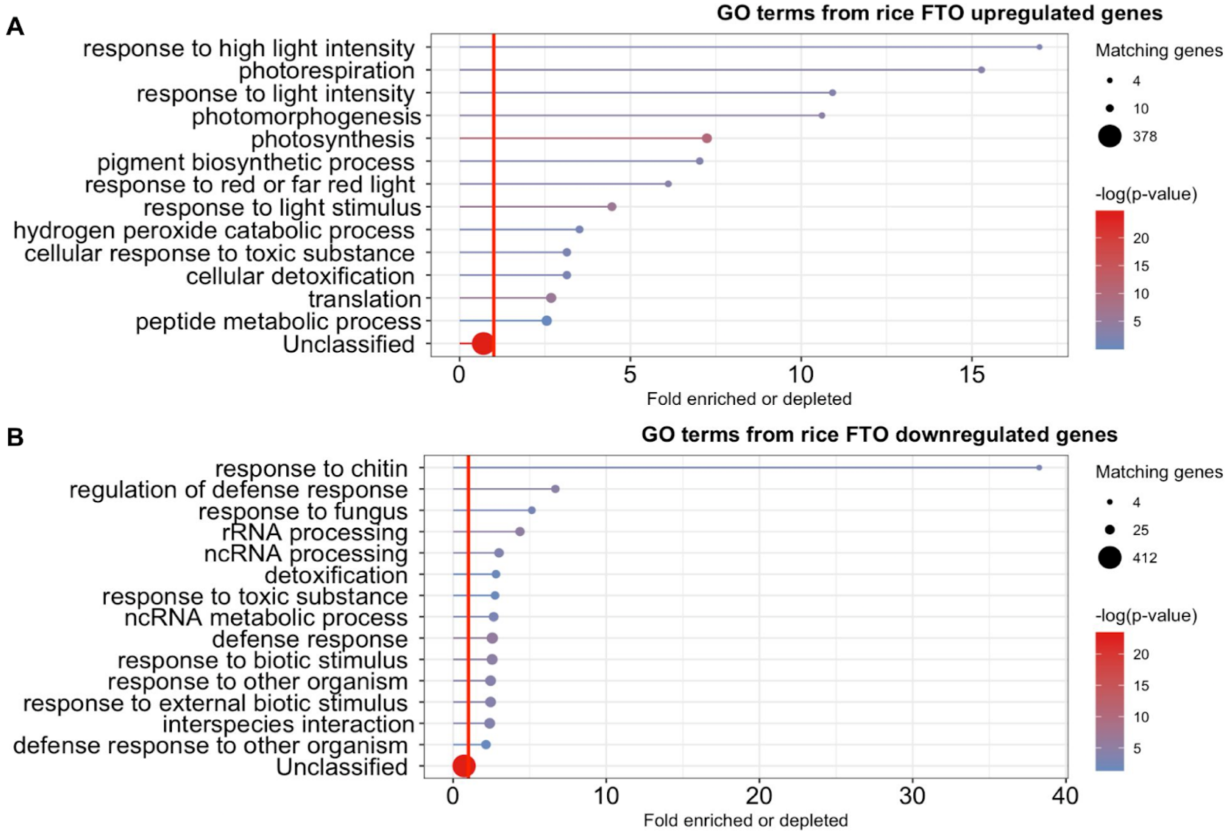
Supplemental Figure 1: Sequence composition differences of UTRs of genes upregulated and downregulated by FTO expression are smaller than those of CDSs. A) Frequency of adenosine in the 5' UTR of transcripts downregulated and upregulated by FTO expression. B) Frequency of adenosine in the CDS of transcripts downregulated and upregulated by FTO expression. C) Frequency of adenosine in the 3' UTR of transcripts downregulated and upregulated by FTO expression. D) Length of the 5' UTR of transcripts downregulated and upregulated by FTO expression. E) Length of the CDS of transcripts downregulated and upregulated by FTO expression. F) Length of the 3’ UTR of transcripts downregulated and upregulated by FTO expression. G) RRACH motifs per kb in the 5’ UTR of transcripts downregulated and upregulated by FTO expression. H) RRACH motifs per kb in the CDS of transcripts downregulated and upregulated by FTO expression. I) RRACH motifs per kb in the 3’ UTR of transcripts downregulated and upregulated by FTO expression. Boxplot central boxes cover the two central quartiles, points are raw data. Kruskal-Wallis p-value tests for significant difference between any of the groups, if Kruskal-Wallis result was significant then Wilcoxon test was performed between all three categories of transcript with Benjamini-Hochberg correction for multiple comparisons, pairwise p-values displayed.



Supplemental Figure 2: Sequence-level differences between GO categories are minimal. A) Adenosine frequency of genes for three GO terms. B) Length of genes for three GO terms. C) RRACH motif abundance per kb for three GO terms. All panels include the CDS of all genes annotated as within the GO category on TAIR. Boxplot central boxes cover the two central quartiles, points are raw data. Kruskal-Wallis p-value tests for significant difference between any of the groups, if Kruskal-Wallis result was significant then Wilcoxon test was performed between all three categories of transcript with Benjamini-Hochberg correction for multiple comparisons, pairwise p-values displayed.

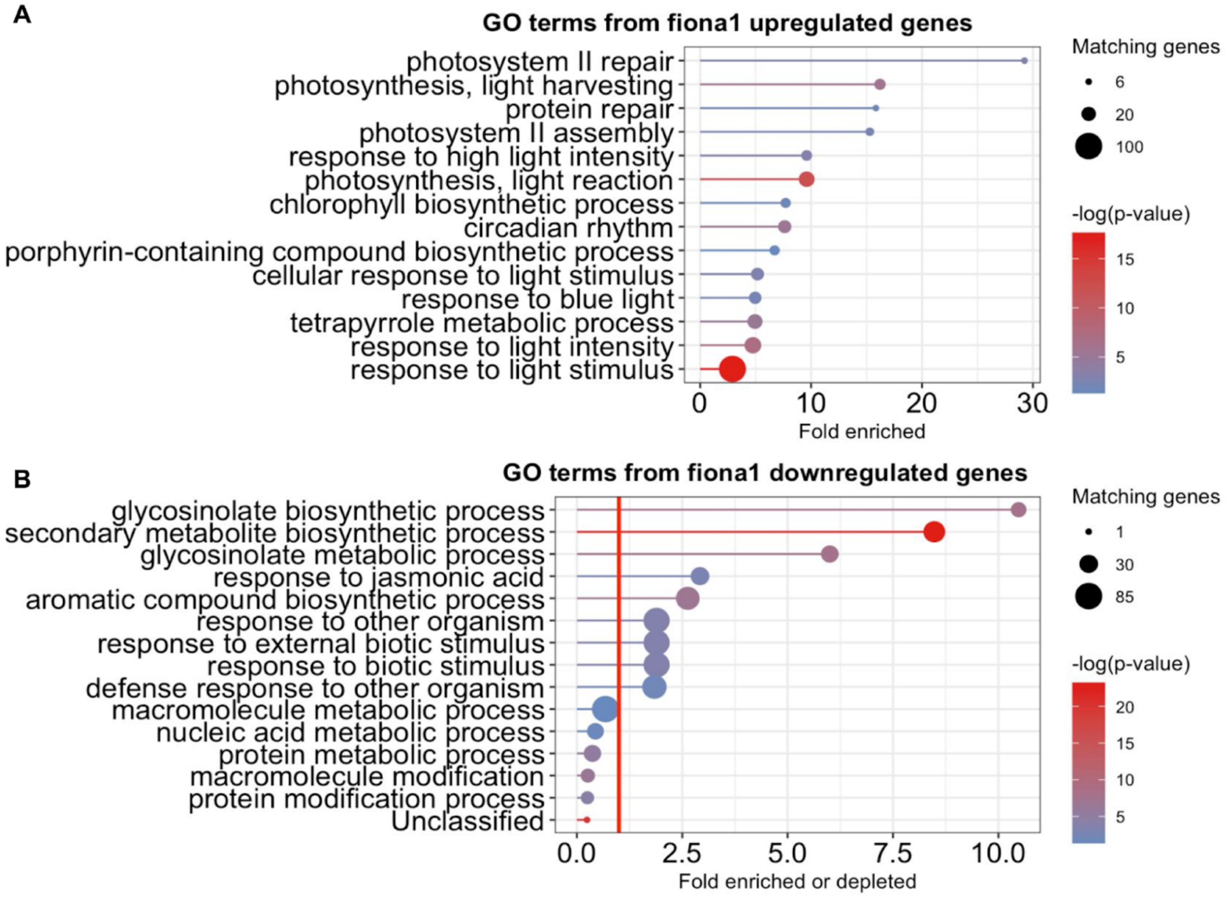
4.3.7 Trends identified in our dataset are also found in previously-published RNA-seq datasets of plants with altered m^{6}A deposition

We next investigated whether these structural differences between downregulated, randomly-selected, and upregulated genes were present in previous RNA-seq datasets of m^{6}A-altered plants. We begin with Yu *et al.*'s rice shoot RNA-seq dataset, which we used to generate even-sized groups of downregulated, randomly-selected, and upregulated genes in the FTO plants. We found genes upregulated by FTO to have higher CDS adenosine frequency (Figure 4D), matching our results and suggesting commonalities in the mechanism of action between rice and *Arabidopsis*. Median CDS length is only negligibly different between upregulated and downregulated genes in their dataset, though much like our dataset the upregulated gene list had fewer very long genes (Figure 4E), which is not an artifact of different size gene sets due to our balanced gene lists. Similar to our dataset, downregulated genes were depleted for RRACH motifs compared to randomly-selected or upregulated genes (Figure 4F). We performed GO analysis using PANTHER on the upregulated and downregulated gene lists, with selected GO terms plotted in Supplemental Figure 3 and full lists available as Supplemental Table 7 and Supplemental Table 8, respectively. Genes associated with response to light intensity are over-represented among their upregulated gene list, including 'response to high light intensity', 'response to light intensity', and 'response to red or far red light', matching findings from our data (Supplemental Figure 3, Supplemental Table 7). Protein synthesis related genes were also over-represented, including 'peptide metabolic process' and 'translation', also matching our results. Among the downregulated genes, the most enriched GO term was 'response to chitin', suggesting FTO-expressing plants are tuning down their defensive gene expression. 'Defense response', 'regulation of defense response', 'interspecies interaction', 'response to biotic stimulus', 'defense response to other organism', and 'response to fungus' are all overrepresented among their downregulated gene lists, matching our results (Supplemental Figure 3, Supplemental Table 8).



Supplemental Figure 3: GO analysis of Yu *et al*'s FTO rice RNA-seq reveals many of the same trends identified in our data. A) Lollipop plot of selected GO terms enriched among Yu *et al*'s FTO rice upregulated genes. Dot size indicates number of genes, line length indicates fold enrichment, color indicates statistical significance. GO categories with the circle to the right of the red vertical line are enriched, those to the left of the line are depleted. B) Lollipop plot of selected GO terms enriched among Yu *et al*'s FTO rice downregulated genes. Dot size indicates number of genes, line length indicates fold enrichment, color indicates statistical significance. GO categories with the circle to the right of the red vertical line are enriched, those to the left of the line are depleted. Full lists of GO categories and associated data are available in Supplemental Table 7 and Supplemental Table 8 for the upregulated and downregulated gene lists respectively.

To better understand how our results match previous *Arabidopsis* data, we also analyzed a publicly available RNA-seq dataset from *fiona1* mutant plants (Parker et al., 2022). We reason that mutants of an m^{6}A writer may behave similarly to our m^{6}A eraser overexpression lines. We chose mutants of this specific m^{6}A writer because the mutants of most others have severe phenotypes, whereas *fiona1* plants grow to maturity and obtain comparable size to wild-type plants. We performed GO analysis using PANTHER after identifying upregulated and downregulated genes between Col-0 and *fiona1* plants, selected GO terms of which are plotted in Supplemental Figure 4, full results of which are available as Supplemental Table 9 and Supplemental Table 10, respectively. GO terms 'response to high light intensity' and 'response to light' were enriched among the upregulated genes, matching results from our data and our analysis of Yu *et al.*'s data. We found 'response to biotic stimulus', 'defense response to other organism', and 'defense response' were all enriched at least 1.85-fold among the downregulated genes, whereas no defense categories were either depleted among the downregulated genes or enriched among the upregulated genes (Supplemental Figure 4, Supplemental Table 10). The organellar genes were also non-evenly distributed between upregulated and downregulated genes in *fiona1*, among the upregulated genes there were 10 plastid and 1 mitochondrial gene, whereas among the downregulated genes were 40 mitochondrial and 1 plastid gene. The strong enrichment of mitochondrial genes in the downregulated gene list matches our findings with FTO, whereas the distribution of plastid genes is opposite to our finding.



Supplemental Figure 4: GO analysis of published f*iona1* mutant data reveals many of the same trends identified in our data. A) Lollipop plot of selected Gene Ontology (GO) terms enriched among *fiona1* upregulated genes. Dot size indicates number of genes, line length indicates fold enrichment, color indicates statistical significance. All GO categories displayed are enriched, the lowest enrichment is 2.91x. B) Lollipop plot of selected GO terms enriched among *fiona1* downregulated genes. Dot size indicates number of genes, line length indicates fold enrichment, color indicates statistical significance. GO categories with the circle to the right of the red vertical line are enriched, those to the left of the line are depleted. Full lists of GO categories and associated data are available in Supplemental Table 9 and Supplemental Table 10 for the upregulated and downregulated gene lists respectively.

We investigated several additional published RNA-seq datasets specifically for the question of organellar genes being overrepresented among the downregulated genes, and did not find a clear pattern — microRNA knockdown of the m^{6}A writer MTA (Wong et al., 2023) revealed in an overrepresentation of mitochondrial genes among the downregulated genes, but plastid genes were overrepresented among upregulated genes, matching the *fiona1* results. Mutants of the m^{6}A writer FIP37 (Wong et al., 2023) show an enrichment for plastid genes among upregulated genes, whereas mitochondrial genes show no enrichment. Mutant plants deficient for the native m^{6}A eraser ALKBH10B also show neither enrichment nor depletion for plastid or mitochondrial genes (Duan et al., 2017).

4.4 Discussion

Our findings broadly support the findings of Yu *et al*., demonstrating how the FTO overexpression approach may be a broadly transferable trait and should encourage future research using modulation of RNA methylation status as a tool to improve plant yield. We find a significant increase in flower and fruit number, stem branching, and reproductive stem biomass. We present several potential hypotheses for the effect, such as the upregulation of SAUR genes and repression of repressive transcription factors. We also present two striking observations for further investigation: an inconsistent but strong pattern of disproportional representation of organellar genes among the lists upregulated and downregulated by FTO, and differences between upregulated and downregulated gene sets such as a reduction in adenosine frequency and RRACH motif abundance among downregulated genes. We believe our data in conjunction with that of Yu *et al*. are sufficient to warrant more widespread research into the effects of FTO expression on plant yield and present directions for future research.

Public opinion has been a major barrier to the widespread deployment of transgenic crops, and we recognize that the introduction of a transgene derived from humans into crops represents a significant public relations liability. To hedge these concerns, future studies pursuing FTO overexpression may benefit from avoiding the human version of the gene, and instead utilize orthologs from close mammalian relatives such as the porcine or bonobo orthologs, which are particularly similar in amino acid sequence. It may also be possible to achieve similar growth improvements with a transgene-free approach by reducing the level of RNA methylation through selective reduction of m^{6}A writers rather than the addition of a phylogenetically-distant m^{6}A eraser. However, the growth defects in reported m^{6}A writer mutants suggest this will need to be approached carefully, and will likely require more mechanistic understanding of the basis of the growth enhancement phenotype. Fortunately, new sequencing technologies are rapidly improving our ability to detect RNA modification (Wang and Jia, 2020), which may light the way towards knowing which RNA species in which tissues at which developmental stages are involved in this remarkable growth-enhancing phenotype.

4.5 Methods

4.5.1 Plant growth

Col-0 *Arabidopsis* were grown to flowering stage for transformation by floral dip and transformed with *Agrobacterium tumefaciens* strain GV3101 following standard methods, and the T0 progeny seeds selected for resistance to 50 mg/L kanamycin during axenic growth on agar plates. Over 50 independent lines were obtained, which were selected and propagated to increase seed and prepare for the main growth experiments with 5 lines of homozygous T3 plants. These plants were grown in a AR-95L3 chamber designed for *Arabidopsis* (Percival Scientific) under long day conditions with 12:12 light:dark, 25 °C, 60% relative humidity, “Professional Growing Mix” soil (Sungrow Horticulture) with Peter’s Professional 20-20-20 (ICL Growing Solutions) diluted into the water at 1 gram per Liter once per week. Plants were kept well watered in standard 1020 flats with 24 pots per flat.

4.5.2 Accession Numbers

Sequence data from this article can be found in the EMBL/GenBank database or the *Arabidopsis* Genome Initiative database under the following accession numbers: The human FTO enzyme used in this study is Genbank: KAI2578515, the porcine ortholog is GenBank: ADD69819. MTA is At4g10760, MTB is At4g09980, FIP37 is At3g54170, VIR is At3G05680, FIONA1 is AtMG00980, HAKAI is At5G01160, ECT2 is At3G13460, ECT3 is At5G61020, ECT4 is At1G55500, CPSF30 is At1G30460, and ALKBH10B is At4G02940. SAUR28 is At3G03830, SAUR10 is At2G18010. All other gene names are available in Supplemental Table 3. Raw sequencing data is available through the JGI portal: https://genome.jgi.doe.gov/portal/RNAofexinplants/RNAofexinplants.info.html

4.5.3 RNA extraction and library preparation

15-30 mg of shoot tissue from day-old seedlings grown under axenic conditions on 1/2 MS plates was excised and immediately flash-frozen in liquid nitrogen and stored at -80 °C. RNA extraction was performed using E.Z.N.A Total RNA kit (Omega BioTek) following manufacturer directions. Extracted RNA was then DNAse treated with TURBO DNA-Free kit (Thermo Scientific). mRNA was isolated from an input of 1 ug of total RNA with oligo dT magnetic beads and fragmented to 300 bp - 400 bp with divalent cations at a high temperature. Using TruSeq stranded mRNA kit (Illumina), the fragmented mRNA was reverse transcribed to create the first strand of cDNA with random hexamers and SuperScript™ II Reverse Transcriptase (Thermo Fisher Scientific) followed by second strand synthesis. The double stranded cDNA fragments were treated with A-tailing, ligation with JGI's unique dual indexed adapters (IDT) and enriched using 8 cycles of PCR. The prepared libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. Sequencing of the flowcell was performed on the Illumina NovaSeq sequencer using NovaSeq XP V1.5 reagent kits, S4 flowcell, following a 2x151 indexed run recipe.

4.5.4 RNA-seq data analysis

Read preprocessing: Raw fastq file reads were filtered and trimmed using the JGI QC pipeline resulting in the filtered fastq file (*.filter-RNA.gz files). Using BBDuk, raw reads were evaluated for artifact sequence by kmer matching (kmer=25), allowing 1 mismatch and detected artifact was trimmed from the 3' end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, following trimming, reads under the length threshold were removed (minimum length 25 bases or 1/3 of the original read length - whichever is longer). **Read Alignment and Counting:** Filtered reads from each library were aligned to the reference genome using HISAT2 version 2.2.1 (with -k 1 flag) (BAMs/ directory). Strand-specific coverage bigWig files (fwd and rev) were generated using deepTools v3.1(bigWigs/ directory). featureCounts was used to generate the raw gene counts (counts.txt) file using gff3 annotations. Only primary hits assigned to the reverse strand were included in the raw gene counts (-s 2 -p --primary options). Raw gene counts were used to evaluate the level of correlation between biological replicates using Pearson's correlation and determine which replicates would be used in the DGE analysis (replicate_analysis.txt, replicate_analysis_heatmap.pdf). In the heatmap view, the libraries were ordered as groups of replicates. The cells containing the correlations between replicates have a purple (or white) border around them. FPKM and TPM normalized gene counts are also provided (fpkm_counts.txt and tpm_counts.txt). **Strandedness:** Features assigned to the forward strand were also tabulated (-s 1 -p --primary options). Strandedness of each library was estimated by calculating the percentage of reverse-assigned fragments to the total assigned fragments (reverse plus forward hits). **Differential Gene Expression:** DESeq2 (version 1.30.0) was used to determine which genes were differentially expressed between pairs of conditions. The parameters used to call a gene differentially expressed between conditions were adjusted p-value < 0.01. The file DGE_summary.txt includes the log2 fold change, adjusted Pval and whether the gene is significantly differentially expressed (TRUE/FALSE/NA) for each pair of conditions specified. We also include shrunken log2 fold changes from DESeq2's "normal" method for visualization and ranking. Refer to the Sample Summary Table above to match SampleName with the ConditionNumber assigned for analysis. Individual results for each pairwise comparison are in the directory Pairwise DGE Results. Note: Raw gene counts (counts.txt), not normalized counts are used for DGE analysis. DESeq2 conducts its own internal normalization using a sophisticated sampling model. **Gene Set Enrichment Analysis:** Gene set enrichment analysis evaluates whether certain gene sets or pathways contain more differentially expressed genes than expected by chance. If KEGG annotations were available for the genes in this reference genome, PADOG (version 1.36.0) was used to calculate gene set enrichment of the KEGG pathways found in the genome. Gene annotations are summarized in gene_functional_annotations.tsv and KEGG pathways with at least three genes present in the reference genome are provided in KEGG_gene_sets.tsv. Figures were produced using RStudio with the Tidyverse package (Wickham et al., 2019). **Analysis of Yu *et al.* dataset:** We used the shoot standard RNA-seq (as opposed to m^{6}A-seq) data, and sorted genes by average fold-change between wild-type and FTO plants. We then generated our three gene lists by selecting the 1000 most-upregulated genes, 1000 most-downregulated genes, and 1000 genes selected using a random number generator.

4.6 Acknowledgements

We acknowledge support from the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The work (proposal: 10.46936/10.25585/60008943) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We thank Mitch Thompson, Sasilada Sirirungruang, and Simón Alamos for discussions during the preparation of this manuscript. We thank Daniel Peterson for comments and suggestions on our RNA-seq data analysis. We thank Peter Mellinger for watering plants.

4.7 Author contributions

KM conceived and performed the experiments and wrote the manuscript. LW assisted with bioinformatic analysis. PS supervised and provided funding. All authors have read and approved the manuscript.

4.8 Guide to supplemental files available online

Supplemental Table 1: Phenotypes of T3 generation plants
Supplemental Table 2: RNA-seq counts table
Supplemental Table 3: Gene IDs, sequence, and other parameters for upregulated, downregulated, and randomly-selected gene lists
Supplemental Table 4: PANTHER GO Term analysis of upregulated genes
Supplemental Table 5: PANTHER GO Term analysis of downregulated genes
Supplemental Table 6: Highly differentially expressed genes
Supplemental Table 7: PANTHER GO Term analysis of upregulated genes from Yu *et al*.'s dataset
Supplemental Table 8: PANTHER GO Term analysis of downregulated genes from Yu *et al*.'s dataset
Supplemental Table 9: PANTHER GO Term analysis of upregulated genes from publicly-available *fiona1* dataset
Supplemental Table 10: PANTHER GO Term analysis of downregulated genes from publicly-available *fiona1* dataset

4.9 References

Bemer M, van Mourik H, Muiño JM, Ferrándiz C, Kaufmann K, Angenent GC (2017) FRUITFULL controls SAUR10 expression and regulates

Arabidopsis growth and architecture. Journal of Experimental Botany **68**: 3391–3403

Bodi Z, Zhong S, Mehra S, Song J, Li H, Graham N, May S, Fray R (2012) Adenosine Methylation in Arabidopsis mRNA is Associated with the 3′ End and Reduced Levels Cause Developmental Defects. Frontiers in Plant Science 3:

Chae K, Isaacs CG, Reeves PH, Maloney GS, Muday GK, Nagpal P, Reed JW (2012) Arabidopsis SMALL AUXIN UP RNA63 promotes hypocotyl and stamen filament elongation. The Plant Journal **71**: 684–697

Chia KF The Arabidopsis transcription factor ERF13 negatively regulates defense against Pseudomonas syringae. M.S. University of California, San Diego, United States -- California

Church C, Moir L, McMurray F, Girard C, Banks GT, Teboul L, Wells S, Brüning JC, Nolan PM, Ashcroft FM, et al (2010) Overexpression of Fto leads to increased food intake and results in obesity. Nat Genet **42**: 1086–1092

Ciftci-Yilmaz S, Morsy MR, Song L, Coutu A, Krizek BA, Lewis MW, Warren D, Cushman J, Connolly EL, Mittler R (2007) The EAR-motif of the Cys2/His2-type Zinc Finger Protein Zat7 Plays a Key Role in the Defense Response of Arabidopsis to Salinity Stress*. Journal of Biological Chemistry **282**: 9260–9268

Duan H-C, Wei L-H, Zhang C, Wang Y, Chen L, Lu Z, Chen PR, He C, Jia G (2017) ALKBH10B Is an RNA N6-Methyladenosine Demethylase Affecting Arabidopsis Floral Transition. Plant Cell **29**: 2995–3011

Fischer J, Koch L, Emmerling C, Vierkotten J, Peters T, Brüning JC, Rüther U (2009) Inactivation of the Fto gene protects from obesity. Nature **458**: 894–898

Gao X, Shin Y-H, Li M, Wang F, Tong Q, Zhang P (2010) The Fat Mass and Obesity Associated Gene FTO Functions in the Brain to Regulate Postnatal Growth in Mice. PLOS ONE **5**: e14005

Guo T, Liu C, Meng F, Hu L, Fu X, Yang Z, Wang N, Jiang Q, Zhang X, Ma F (2022) The m6A reader MhYTP2 regulates MdMLO19 mRNA stability and antioxidant genes translation efficiency conferring powdery mildew resistance in apple. Plant Biotechnology Journal **20**: 511–525

Huang X, Lu Z, Zhai L, Li N, Yan H (2023) The Small Auxin-Up RNA SAUR10 Is Involved in the Promotion of Seedling Growth in Rice. Plants **12**: 3880

Huot B, Yao J, Montgomery BL, He SY (2014) Growth–Defense Tradeoffs in Plants: A Balancing Act to Optimize Fitness. Molecular Plant **7**: 1267–1287

Jiang L, Li R, Yang J, Yao Z, Cao S (2023) Ethylene response factor ERF022 is involved in regulating Arabidopsis root growth. Plant Mol Biol **113**: 1–17

Klepikova AV, Kasianov AS, Gerasimov ES, Logacheva MD, Penin AA (2016) A high resolution map of the Arabidopsis thaliana developmental transcriptome based on RNA-seq profiling. Plant J **88**: 1058–1070

Lee S, Park JH, Lee MH, Yu J, Kim SY (2010) Isolation and functional characterization of CE1 binding proteins. BMC Plant Biology **10**: 277

Li K, Huang W, Wang Z, Nie Q (2022) m6A demethylase FTO regulate CTNNB1 to promote adipogenesis of chicken preadipocyte. Journal of Animal Science and Biotechnology **13**: 147

Luo G-Z, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, Liu J, Chen K, Jia G, Bergelson J, et al (2014) Unique features of the m6A methylome in Arabidopsis thaliana. Nat Commun **5**: 5630

Lv B, Wei K, Hu K, Tian T, Zhang F, Yu Z, Zhang D, Su Y, Sang Y, Zhang X, et al (2021) MPK14-mediated auxin signaling controls lateral root development via ERF13-regulated very-long-chain fatty acid biosynthesis. Molecular Plant **14**: 285–297

Merkestein M, Laber S, McMurray F, Andrew D, Sachse G, Sanderson J, Li M, Usher S, Sellayah D, Ashcroft FM, et al (2015) FTO influences adipogenesis by regulating mitotic clonal expansion. Nat Commun **6**: 6792

Mi H, Muruganujan A, Casagrande JT, Thomas PD (2013) Large-scale gene function analysis with the PANTHER classification system. Nat Protoc **8**: 1551–1566

Nagpal P, Reeves PH, Wong JH, Armengot L, Chae K, Rieveschl NB, Trinidad B, Davidsdottir V, Jain P, Gray WM, et al (2022) SAUR63 stimulates cell growth at the plasma membrane. PLOS Genetics **18**: e1010375

Nowak K, Wójcikowska B, Gaj MD (2015) ERF022 impacts the induction of somatic embryogenesis in Arabidopsis through the ethylene-related pathway. Planta **241**: 967–985

Ohta M, Matsui K, Hiratsu K, Shinshi H, Ohme-Takagi M (2001) Repression Domains of Class II ERF Transcriptional Repressors Share an Essential Motif for Active Repression. The Plant Cell **13**: 1959–1968

Parker MT, Soanes BK, Kusakina J, Larrieu A, Knop K, Joy N, Breidenbach F, Sherwood AV, Barton GJ, Fica SM, et al (2022) m6A modification of U6 snRNA modulates usage of two major classes of pre-mRNA 5' splice site. eLife **11**: e78808

Ren H, Gray WM (2015) SAUR Proteins as Effectors of Hormonal and Environmental Signals in Plant Growth. Mol Plant **8**: 1153–1164

Ronkainen J, Huusko TJ, Soininen R, Mondini E, Cinti F, Mäkelä KA, Kovalainen M, Herzig K-H, Järvelin M-R, Sebert S, et al (2015) Fat mass- and obesity-associated gene Fto affects the dietary response in mouse white adipose tissue. Sci Rep **5**: 9233

Roundtree IA, Evans ME, Pan T, He C (2017) Dynamic RNA Modifications in Gene Expression Regulation. Cell **169**: 1187–1200

Růžička K, Zhang M, Campilho A, Bodi Z, Kashif M, Saleh M, Eeckhout D, El-Showk S, Li H, Zhong S, et al (2017) Identification of factors required for m6A mRNA methylation in Arabidopsis reveals a role for the conserved E3 ubiquitin ligase HAKAI. New Phytologist **215**: 157–172

Sakamoto H, Maruyama K, Sakuma Y, Meshi T, Iwabuchi M, Shinozaki K, Yamaguchi-Shinozaki K (2004) Arabidopsis Cys2/His2-Type Zinc-Finger Proteins Function as Transcription Repressors under Drought, Cold, and High-Salinity Stress Conditions. Plant Physiology **136**: 2734–2746

Shen L, Liang Z, Gu X, Chen Y, Teo ZWN, Hou X, Cai WM, Dedon PC, Liu L, Yu H (2016) N6-Methyladenosine RNA Modification Regulates Shoot Stem Cell Fate in Arabidopsis. Developmental Cell **38**: 186–200

Shinde H, Dudhate A, Kadam US, Hong JC (2023) RNA methylation in plants: An overview. Frontiers in Plant Science 14:

Spartz AK, Lor VS, Ren H, Olszewski NE, Miller ND, Wu G, Spalding EP, Gray WM (2017) Constitutive Expression of Arabidopsis SMALL AUXIN UP RNA19 (SAUR19) in Tomato Confers Auxin-Independent Hypocotyl Elongation. Plant Physiology **173**: 1453–1462

Sun B, Bhati KK, Song P, Edwards A, Petri L, Kruusvee V, Blaakmeer A, Dolde U, Rodrigues V, Straub D, et al (2022) FIONA1-mediated methylation of the 3'UTR of FLC affects FLC transcript levels and flowering in Arabidopsis. PLOS Genetics **18**: e1010386

Wang X, Zhu L, Chen J, Wang Y (2015) mRNA m6A methylation downregulates adipogenesis in porcine adipocytes. Biochemical and Biophysical Research Communications **459**: 201–207

Wang Y, Jia G (2020) Detection methods of epitranscriptomic mark N6-methyladenosine. Essays in Biochemistry **64**: 967–979

Wang Z, Yang L, Liu Z, Lu M, Wang M, Sun Q, Lan Y, Shi T, Wu D, Hua J (2019) Natural variations of growth thermo-responsiveness determined by SAUR26/27/28 proteins in Arabidopsis thaliana. New Phytologist **224**: 291–305

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, et al (2019) Welcome to the Tidyverse. Journal of Open Source Software **4**: 1686

Wong CE, Zhang S, Xu T, Zhang Y, Teo ZWN, Yan A, Shen L, Yu H (2023) Shaping the landscape of N6-methyladenosine RNA methylation in Arabidopsis. Plant Physiology **191**: 2045–2063

Woodworth CM (1931) Breeding for Yield in Crop Plants [1]. Agronomy Journal **23**: 388–395

Yu Q, Liu S, Yu L, Xiao Y, Zhang S, Wang X, Xu Y, Yu H, Li Y, Yang J, et al (2021) RNA demethylation increases the yield and biomass of rice and potato plants in field trials. Nat Biotechnol **39**: 1581–1588

Zhao P, Zhang F, Liu D, Imani J, Langen G, Kogel K-H (2017) Matrix metalloproteinases operate redundantly in Arabidopsis immunity against necrotrophic and biotrophic fungal pathogens. PLOS ONE **12**: e0183577

Zheng H, Sun X, Zhang X, Sui N (2020) m6A Editing: New Tool to Improve Crop Quality? Trends in Plant Science **25**: 859–867

Zhong S, Li H, Bodi Z, Button J, Vespa L, Herzog M, Fray RG (2008) MTA Is an Arabidopsis Messenger RNA Adenosine Methylase and Interacts with a Homolog of a Sex-Specific Splicing Factor. Plant Cell **20**: 1278–1288

Zhou A, Kirkpatrick LD, Ornelas IJ, Washington LJ, Hummel NFC, Gee CW, Tang SN, Barnum CR, Scheller HV, Shih PM (2023) A Suite of Constitutive Promoters for Tuning Gene Expression in Plants. ACS Synth Biol **12**: 1533–1545

Zhou L, Tang R, Li X, Tian S, Li B, Qin G (2021) N6-methyladenosine RNA modification regulates strawberry fruit ripening in an ABA-dependent manner. Genome Biol **22**: 168

Chapter 5: Plant transcriptional repressors – sometimes you need to switch it off

Kasey Markel, Jean Sabety, Shehan Wijesinghe, Patrick M. Shih

A modified version of this manuscript is currently in review for formal publication. All authors have consented to it being re-published here (or published here originally if the timing works out that way). Section 5.0 is entirely original just for this dissertation.

5.0 Chapter Preface

As demonstrated by several decades of widespread use of single-gene genetically engineered traits across billion acres of agricultural land as well as the much more recent experiments showing RNA demethylases can modify plant morphology and increase yield, it is not necessary to use a complex suite of finely-tuned genetic elements to achieve breathtaking effects. Despite this, for the roughly two decades that synthetic biology has been around as a discipline, synthetic biologists have worked hard to expand the maximum number of genes that can be inserted into genomes and to build well-characterized libraries of parts that enable tunable expression. Especially in the world of microbial synthetic biology, these advances have born major fruit, with ever-larger metabolic pathways being ported into production chassis over the years, the process-development of which clearly demonstrates that titers are massively improved by a combinatorial exploration of different promoters, terminators, CDS alternate coding schemes, and other genetic tricks.

The generation and characterization of such genetic parts has been one of the core functions of synthetic biology research over the last two decades, and it is only natural that such research would comprise one element of a PhD working on plant synthetic biology. This particular branch of synthetic biology is more properly characterized as engineering than science, for the goal is not to learn about life as it is but instead to develop parts and technologies to change life into new and desired forms. As an engineering discipline, the benefit at the end of the day is not knowledge gained but products developed, and this fact in combination with the downward trajectory of the political palatability and therefore economic viability of novel transgenic crop traits ultimately resulted in this project becoming less of interest over time. A large suite of transcriptional repressor protein motifs could enable many interesting plant metabolic engineering efforts, but the nature of the motifs is that they could only be used in a transgenic context.

Unlike the expression of RNA demethylases to increase yield or alternate biosynthetic enzymes in key metabolic pathways to circumvent herbicide toxicity, repression of native genes is unlikely to offer the massive improvements in plant phenotype required to justify the massive capital expenses of getting a novel transgenic crop approved. This is doubly true in light of the additional alternative of simply using RNA-guided nucleases to knock out the gene to be repressed, resulting in the lower gene-edited regulatory burden. This is not to say that a library of transcriptional repressors has no scientific or engineering value – indeed, as the rest of this chapter describes, this library of repressor parts could enable new types of plant engineering that would not be possible with other technologies such as knockout. However, it does mean that the business case for the development of transgenic crops using transcriptional repressors is rather weak, and I predict that no such crops will reach market anywhere in the world in the next two decades, if they ever do, which certainly reduced the glamor of the project in my eyes. Nonetheless, this chapter details a large suite of experiments carried out across many years, and will likely be of interest to future scientists, even if the commercial prospects that originally attracted me to the area are slim.

5.1 Abstract

Regulation of gene expression is essential for all life. Tools to manipulate gene expression level have therefore proven to be very valuable in efforts to engineer biological systems. However, there are few well-characterized genetic parts that reduce gene expression in plants, commonly known as transcriptional repressors. We characterized the repression activity of a library consisting of approximately 25% of the members of the largest known family of repressors. Combining sequence information with our trans-regulatory function data, we next generated a library of synthetic transcriptional repressors, with function predicted in advance. After characterizing our synthetic library, we demonstrate that not only are many of our synthetic constructs functional as repressors, but our advance predictions of repression strength were better than random guesses. We also assessed the functionality of known transcriptional repressors from a wide range of eukaryotes. Our study represents the largest plant repressor library experimentally characterized to date, providing unique opportunities for tuning transcription in plants.

5.2 Introduction

Tuning of expression levels of genes is among the most powerful methods for controlling the phenotype of a cell, and is extensively used by both nature and biotechnologists[1–3]. For plants in particular, the modulation of the expression of existing genes has led to some of the greatest successes in the history of plant biotechnology, such as reduced browning in apples[2] and resistance to viruses via silencing of viral genes[4]. Despite these standout successes, plant biotechnologists have relatively few tools for gene expression modulation, which limits our capacity to engineer plants.

One highly customizable approach for controlling plant phenotype level involves targeted repression of the genes encoding pathway enzymes that result in metabolites not desired in a particular tissue type or developmental time point. Because transcriptional repression is mediated by proteins, this approach allows for spatial and temporal control of pathway flux through the application of tissue-, developmental stage-, or physiological state-specific promoters to drive transgenes encoding the repressor proteins. A substantial amount of research has been invested in development of RNA-guided dead nuclease-mediated activation, but less research has focused on repression. Of the existing repression systems, a key limitation is a lack of diversity of well-characterized repressor elements.

The largest known family of repressor motifs in plants are the Ethylene-responsive element binding-factor associated Amphiphilic Repression (EAR) domains, characterized by the motifs LxLxL or DLNxxP, where x can be any amino acid. EAR motif-containing proteins have substantially increased in number in the lineage leading to land plants, and comprise 0.5-2% of protein-encoding genes in angiosperms[5]. The best-characterized EAR repressor is SRDX, a 12 amino-acid motif identified through mutational modifications of the repression domain from the *Superman* transcription factor[6]. While this tool has proven useful for modulation of hormone signaling[7], improving salt tolerance[8], and inducing male sterility for breeding[9], it is not suitable for all transcriptional repression projects because it only offers one level of down-tuning. Here, we aimed to expand the range of characterized repressors to enable in plant synthetic biology applications.

5.3 Results and discussion

5.3.1 Characterization of natural EAR repressors

We initially mined a previously-published bioinformatic search of the arabidopsis genome for putative EAR repressors[10]. From this list, we identified all putative EAR repressors located on the C-terminal end of the gene, which we believed might be most accurately assessed in a medium-throughput assay that depends on C-terminal fusions of trans-elements[11]. In brief, our assay relies upon fusing each trans-element on the C-terminus of the well-characterized Gal4 DNA-binding domain, which acts in concert with a GFP driven by a promoter which contains a Gal4 binding site, as has been previously described[11,12]. The fluorescence of GFP thereby enables us to determine the trans-regulatory behavior of the putative repressor. We identified 90 unique EAR motifs in that set, and added SRDX and Gal4 alone as key controls - a positive control for a strong known repressor, and a negative control to indicate the baseline GFP fluorescence level in the absence of a repressor. We successfully cloned fusions of 84 EAR motifs to Gal4 and infiltrated them all one by one into *Nicotiana benthamiana* leaves for analysis (see Figure 1A).

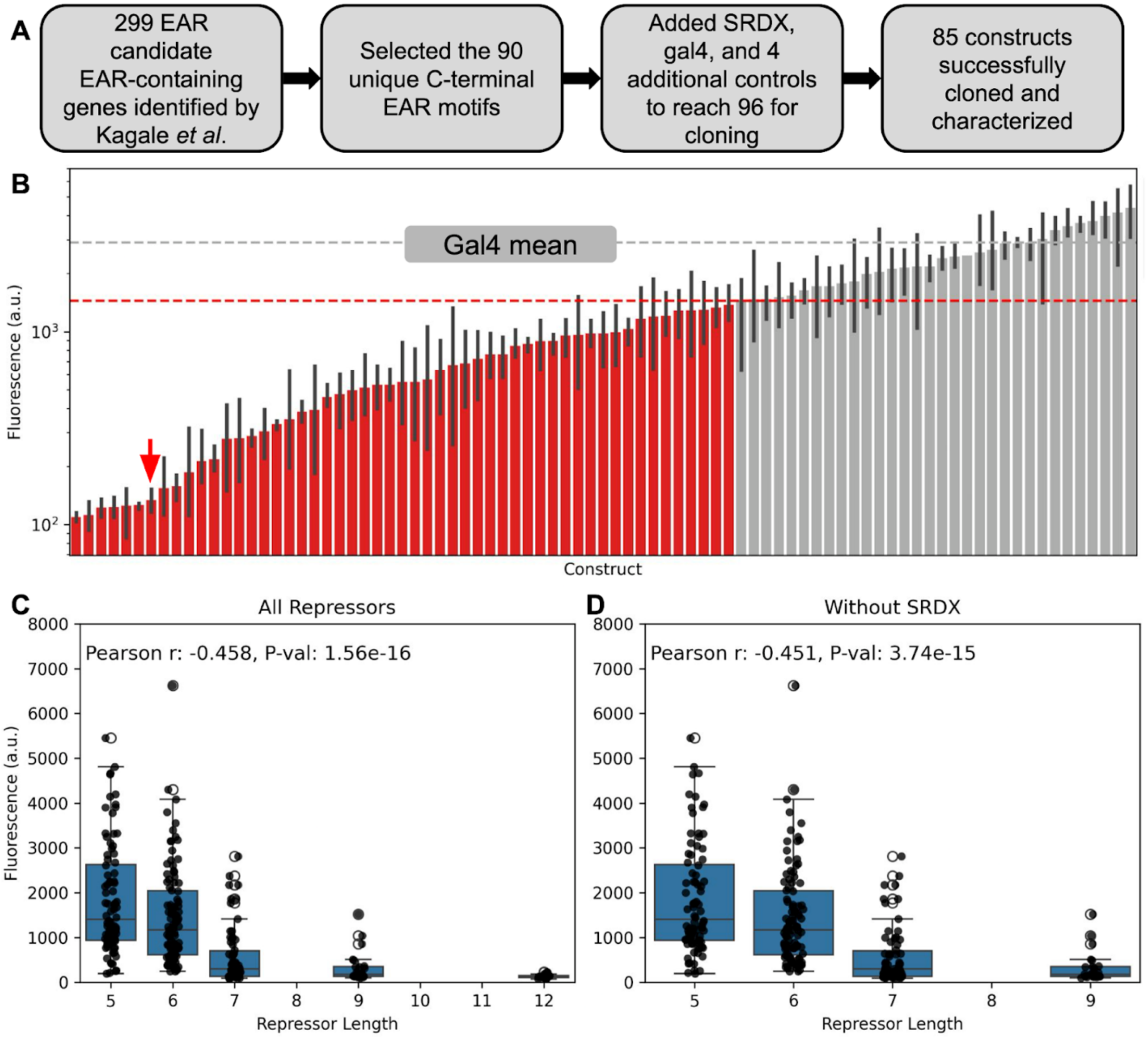
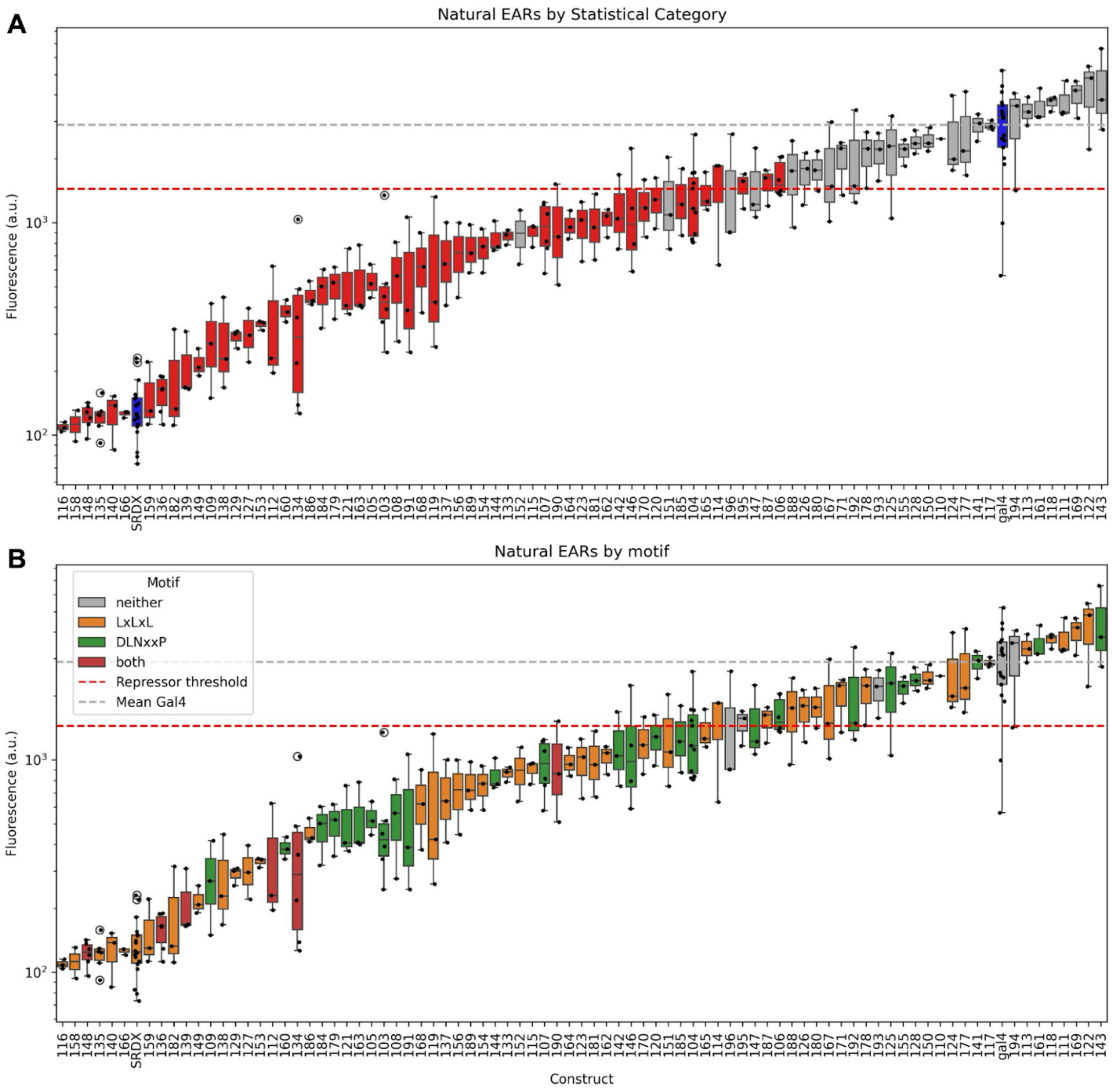


Figure 1: Natural EAR repressors span a broad range of activity and repressor length correlates with strength. A) Flowchart of our selection and cloning of candidate EAR repressors. B) Fluorescence data of repressor assay. Constructs in gray display minimal activity, constructs in red are repressors. Horizontal lines indicate mean activity of Gal4 (grey line) and the threshold for a construct being classified as a repressor (red line). Red arrow indicates SRDX, error bars indicate the standard error of the mean (SEM). C) Strength of repressor constructs by length, measuring all constructs in this library. D) Strength of repressor constructs by length with SRDX, which is one of the strongest and the longest construct, removed. Boxes indicate 25th, 50th, and 75th percentile, raw data are plotted as points.

The majority of the putative EAR domains indeed act as repressors in this system, validating the bioinformatic analysis of Kagale *et al*[10]. (Figure 1B, Supplemental Figure 1, data in Supplemental Table 1). We found that 53 of our 84 constructs passed our threshold of 50% GFP reduction to be classified as repressors. The constructs covered a wide dynamic range from 51% increase to 96% reduction of GFP expression. These repressors cover a wide range of repression activity from barely distinguishable from the negative control up to somewhat stronger than SRDX, the strongest previously-described plant transcriptional repressor. While some repressors had stronger activity than SRDX in our screen, a Dunn test with Bonferroni correction for multiple comparison revealed the differences to not be statistically significant (see Supplemental Code). When we compared the length of the repressors to their average repression strength, we discovered that longer repressors were on average stronger than shorter ones (Figure 1C). This effect is not an artifact of SRDX's effect as the longest and one of the strongest repressors, the correlation only drops minorly when we remove it from the dataset (Figure 1D). We also discovered that repressor constructs containing both the DLNxxP and LxLxL motif were more likely to act as repressors - of the 6 constructs containing both motifs, all were repressors and all were in the stronger half of the constructs, a statistically significant enrichment (Fisher's exact test P = 0.0276, Supplemental Figure 1).



Supplemental Figure 1: Natural EAR library plotted with additional characteristics. A) EAR library plotted with a statistical significance threshold difference from Gal4 rather than a mean repression difference. EAR constructs are colored red if the mean GFP fluorescence is lower than that of Gal4 as determined by a Mann-Whitney U test with a threshold of p = 0.05. Gal4 and SRDX are highlighted in blue. B) EAR library data plotted by whether the construct contains the classic EAR motifs LxLxl, DLNxxP, both, or neither. Boxes indicate 25th, 50th, and 75th percentile, raw data are plotted as points.

5.3.2 Design and characterization of SynEAR (Synthetic EAR) repressors

In an effort to further understand the relationship between the sequence composition of the EAR repressors and their transcriptional trans-regulatory activity, we applied the n-gram analysis method first developed for natural language processing[13], which has more recently been successfully applied to analyze DNA[14] and protein[15] sequences. This method breaks down linear sequences – like words or amino acids – into small 'grams' of various integer lengths n, and thereby can encode not just the relative frequencies of amino acids but also the spatial relationships. As depicted in Figure 2A, all sequences were decomposed into all possible 2, 3, and 4-amino acid long n-grams. These n-grams were then sorted into three categories: those overrepresented among the strongest repressor constructs, those without any overrepresentation, and those overrepresented among the weakest constructs (see methods for detailed thresholds, all n-grams in each category are available in Supplemental Table 2). These n-grams were then randomly concatenated within the length distribution of the original EAR library to create a set of SynEAR (Synthetic EAR) constructs. SynEARs were generated in 4 predicted categories of strength: weak, moderate, strong, and strongest. The first three were constructed of their respective associated n-grams with the same length distribution as the natural EAR library, while the 'strongest' was generated from strong n-grams with a longer length distribution in light of the finding that natural EAR repression strength correlated with length (sequences of all SynEARs available in Supplemental Table 3).

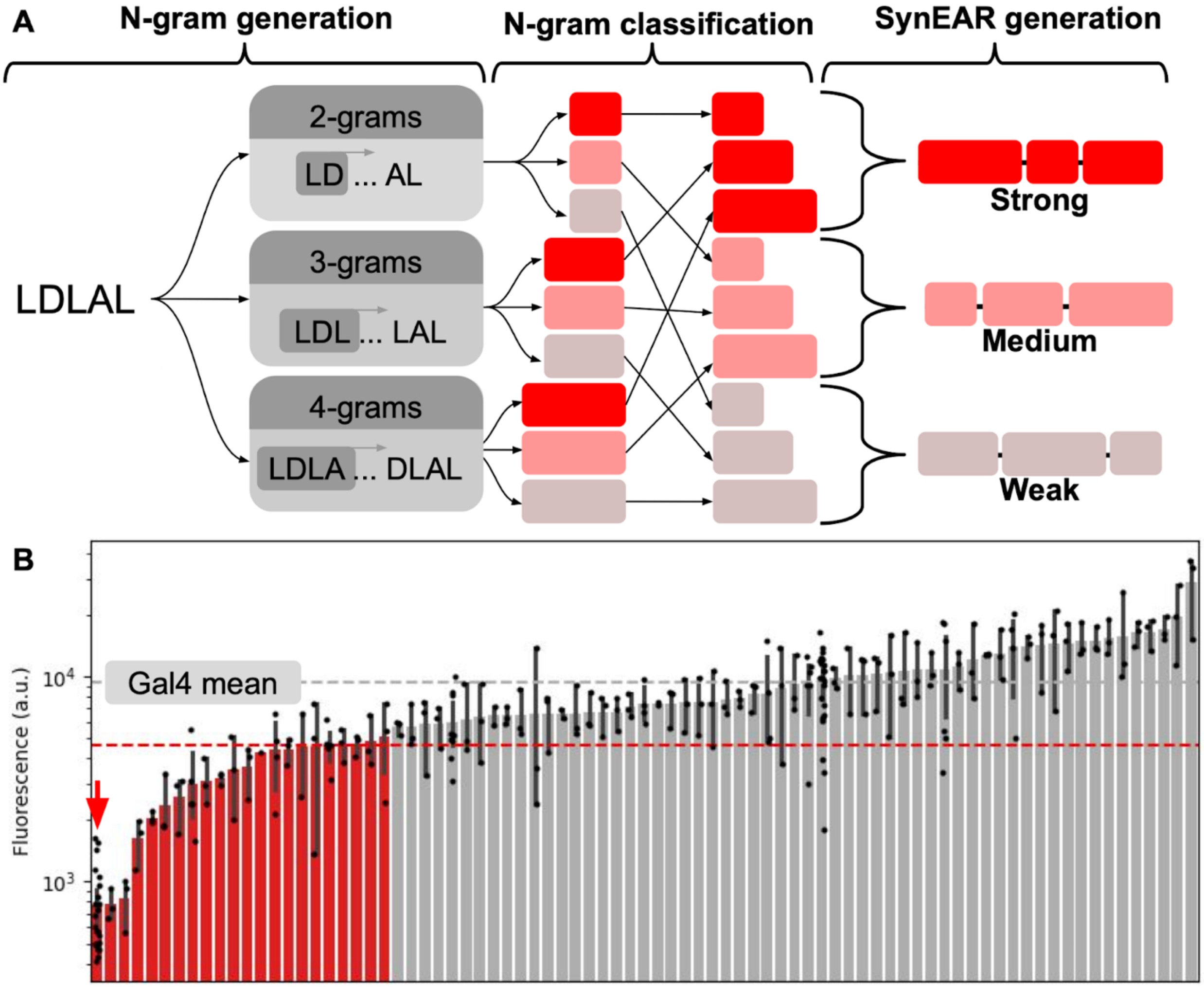
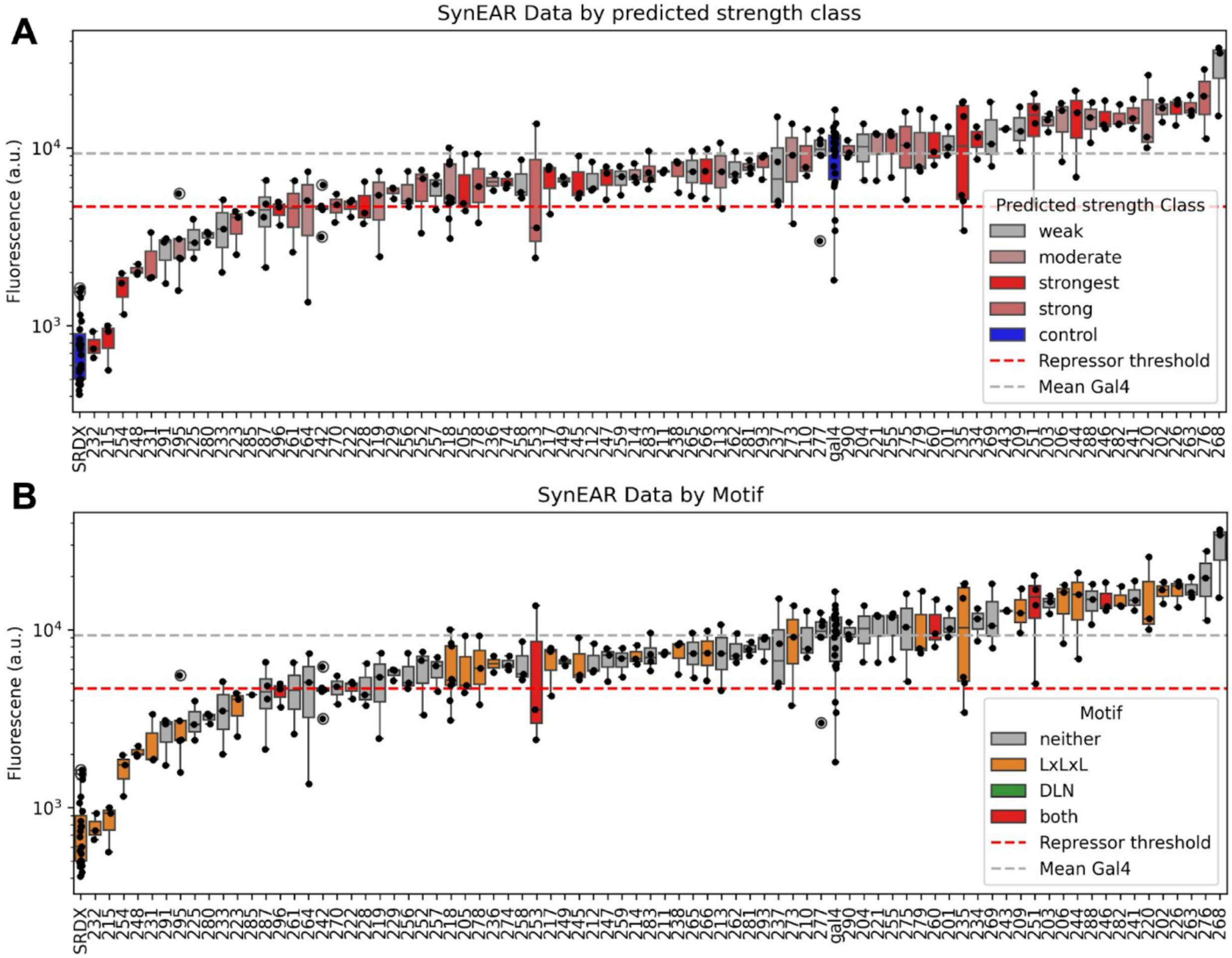


Figure 2: Natural language processing model facilitates generation of SynEAR repressors. A) Cartoon diagram of the n-gram based process used to generate SynEARs. The process is depicted from left to right. B) Repression performance data of SynEARs. Grey dashed line indicates the mean value of Gal4, the red line indicates the threshold for repressors. Arrow indicates SRDX. Error bars indicate the SEM, points indicate raw data.

These SynEARs were cloned and characterized *in planta* in similar fashion to the natural EAR library (Figure 2B, data in Supplemental Table 3). Despite being composed of n-grams from EAR repressors, a smaller fraction of the SynEARs met our threshold for clear repression, only 18 of 80 (Figure 2B). We suspect that this overall lower repression activity of the SynEAR library compared to the natural EAR library is due to a failure to recapitulate the EAR motifs, which are larger than the length of our longest n-grams. Indeed, the majority of SynEARs contain neither of the two classic EAR motifs (Supplemental Figure 2), highlighting a shortcoming in this method for the generation of synthetic repressor constructs. Nonetheless, this SynEAR library did generate and characterize many synthetic trans-regulatory elements, some of which are statistically indistinguishable in repression strength from SRDX, the strongest previously-characterized repressor, despite significantly different sequences, with all pairwise identities for statistically-indistinguishable repressors under 45% (see Supplemental Code for calculations).



Supplemental Figure 2: SynEAR library plotted with additional characteristics. A) EAR library plotted with a statistical significance threshold difference from Gal4 rather than a mean repression difference. SynEAR constructs are colored red if the mean GFP fluorescence is lower than that of Gal4 as determined by a Mann-Whitney U test with a threshold of $p = 0.05$. Gal4 and SRDX are highlighted in blue. B) EAR library data plotted by whether the construct contains the classic EAR motifs LxLxl, DLN, both, or neither. Boxes indicate 25th, 50th, and 75th percentile, raw data are plotted as points.

5.3.3 Evaluation of predicted function of SynEARs

Given that we generated and characterized this library of SynEARs, we next sought to determine whether the measured trans-regulatory activity matched our advance predictions. From simply plotting the SynEARs by predicted strength class (Supplemental Figure 3A), it was clear that the prediction was not overwhelmingly accurate. Nonetheless, we were interested in whether the predictions were better than random guesses, and settled on using the confusion matrix method, which among other use cases is widely used in assessment of classification systems[16]. A confusion matrix is a table or heatmap in which predicted categories are plotted against observed real categories; a perfect classifier only has nonzero values at the y=x diagonal. In this case the four predicted categories of SynEAR trans-regulatory activity were plotted against the actual repression level. Since repression level is a continuous numerical value, it was necessary to establish thresholds to bin our numerical data into four categories of observed repression activity. We determined optimized thresholds in an unbiased fashion through an iterative approximation algorithm implemented in the Python package scikit-learn[17] (see Supplemental Code for precise implementation details). The resulting confusion matrix is shown in Figure 3A.

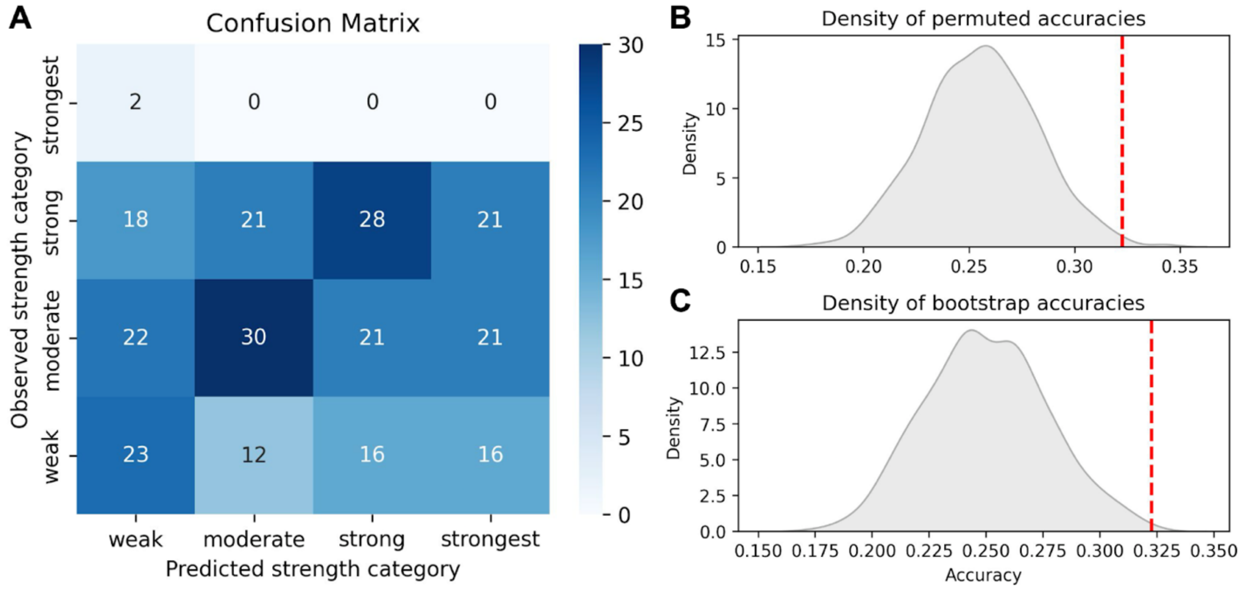
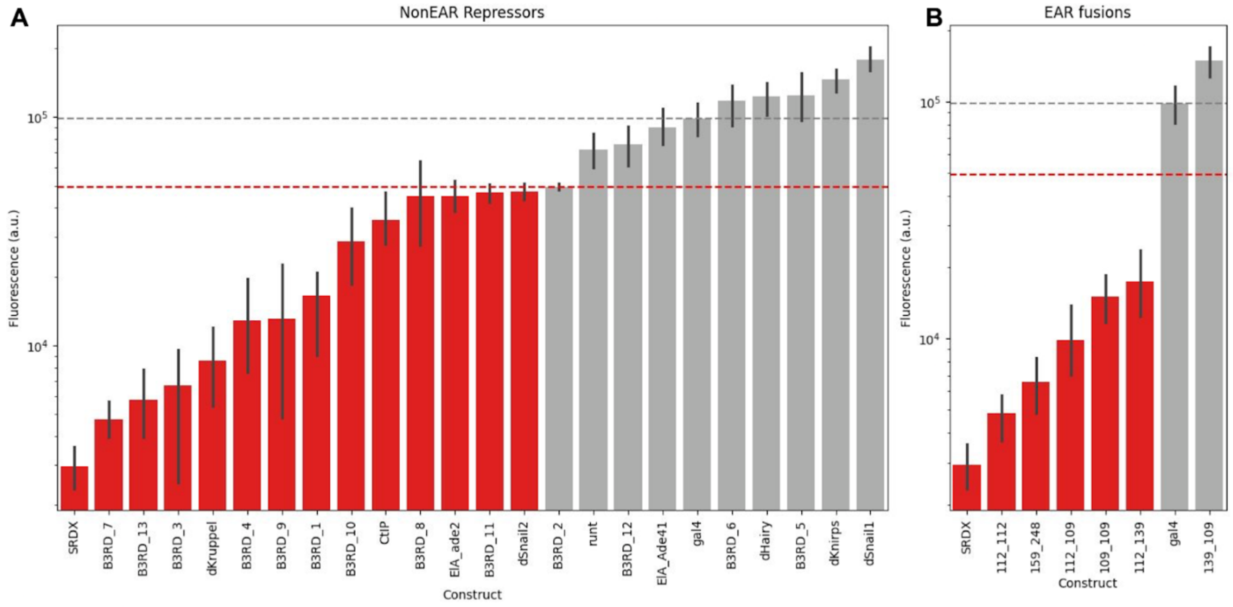


Figure 3: Predictions of the activity of synthetic repressors is better than chance. A) Confusion matrix of predicted vs actual activity strength class.Numbers in heatmap indicate number of constructs with that combination of observed and predicted strength. B) Model accuracy robustness check comparing our model's accuracy to 1000 randomly-permuted model predictions. C) Model accuracy robustness check comparing our model's accuracy to 1000 randomly-generated models with randomized bootstrap predictions.

Our model prediction accuracy was 32.3%. To determine whether this was better than chance, we randomly shuffled the labels of the data and ran 1000 permutations of modeling using Python (see Supplemental Code). These thousand randomly-shuffled versions of our predicted strength levels had an average accuracy of 25.7% with a standard deviation of 2.49%. Our actual model outperformed 99.4% of these shuffled permuted models (Figure 3B), for a two-tailed p-value of 0.013 for the hypothesis that our model is either significantly better or significantly worse than a randomly-shuffled version of our model's guesses. As an additional robustness check, we compared our model against 1000 bootstrapped runs of a random guessing model. Our model accuracy of 32.3% compared well to the 25.0% mean performance from 1000 bootstrapped runs of random-choice, with a standard deviation of 2.66% (Figure 3C). Our model performed 2.66 standard deviations above the mean for random guessing, meaning we outperformed 99.6% of random-choice models, for a two-tailed p-value of 0.0073 for the hypothesis that our model is statistically different from a randomly-selected model.

5.3.4 Characterization of diverse repressors from distant eukaryotic lineages

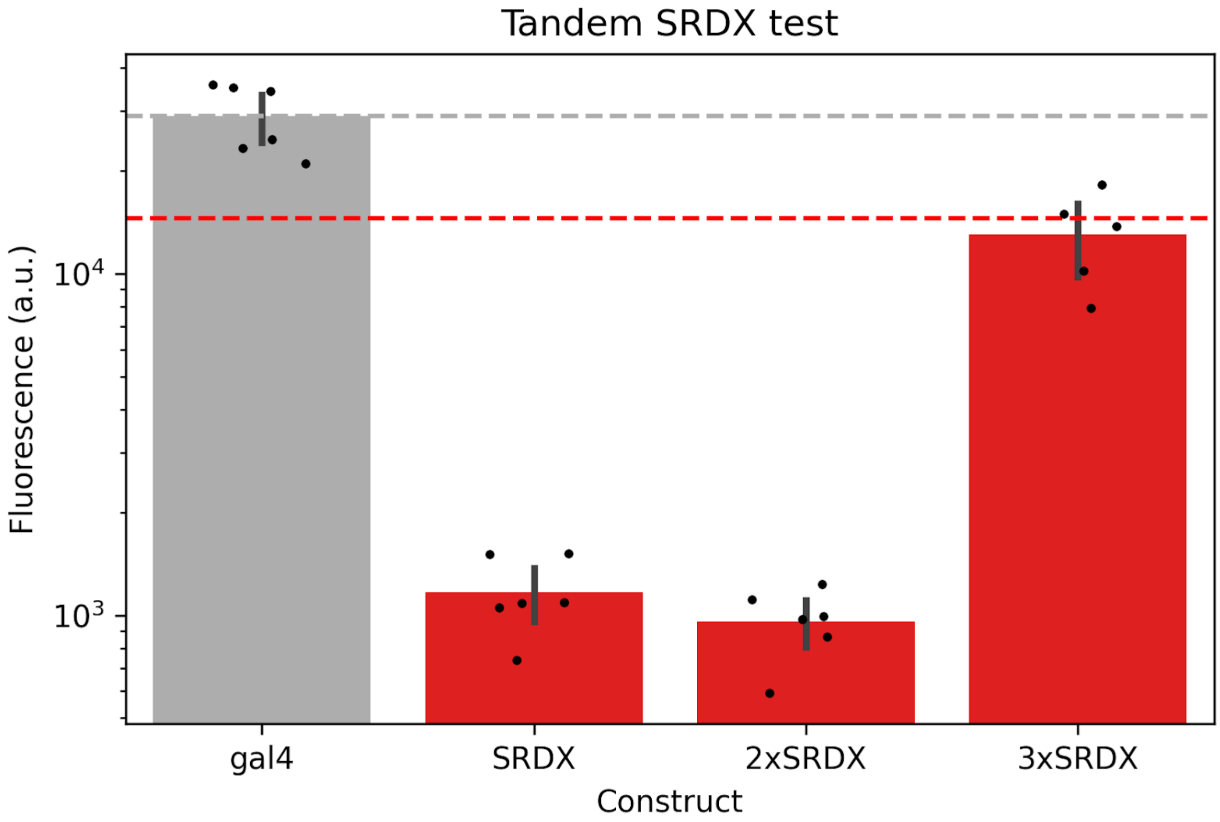
Having characterized a library of natural EARs and SynEARs, we next turned to other classes of repressors, including representative members from other families of plant repressors such as the B3 family[18] as well as known non-plant eukaryotic repressor domains such as the drosophila Hairy[19], Knirps[20], and Runt[21] domains (full list, data, and amino acid sequences available in Supplemental Table 4). We characterized these repression domains in similar fashion to the EAR and SynEAR libraries, and rather surprisingly found that the majority of these domains act as repressors in plants, despite substantial phylogenetic distance (Supplemental Figure 3A). However, all tested non-EAR repressor domains had substantially lower repression activity than SRDX.



Supplemental Figure 3: Non-EAR repressors and EAR fusion repressors. A) Trans-regulatory activity of repressor constructs generated from reported transcriptional repressors across Eukarya. Data and sequence information are available in Supplemental Table 4. B) Trans-regulatory activity of fusions of EAR and SynEAR repressors (EARs have ID numbers between 100 and 196, SynEARs have ID between 200 and 296, all sequence compositions and data are available in Supplemental Table 1 and Supplemental Table 3 for EARs and SynEARs respectively). Error bars indicate the SEM.

5.3.5 Analysis of tandem concatenation of repressors

Since concatenation of genetic parts often results in an increase in their functionality[22], we also tested concatenated fusions of some of our previously-characterized EAR and SynEAR repressor constructs. All but one of which retained repression activity (Supplemental Figure 3B, data in Supplemental Table 4). However, none of these fused repressor constructs was stronger than the strongest non-fused candidates, some of which were contained as components. Surprised by this finding that concatenation resulted in weaker overall repression level than the component parts, we devised a simpler experiment comparing single, double and triple tandem repeats of SRDX, following previous a report which tested 3xSRDX but did not compare it to single or doubly-repeated versions[23]. To our surprise, SRDX and 2xSRDX displayed statistically indistinguishable levels of repression, while 3xSRDX was noticeably weaker (Supplemental Figure 4, data in Supplemental Table 5). This suggests that concatenation of repressor parts does not reliably increase repression strength, and may even weaken it.



Supplemental Figure 4: Concatenation of SRDX reduces repression strength. Reporter fluorescence for gal4 and 1x, 2x, and 3x SRDX. Error bars indicate the SEM, raw data are plotted as points

5.3.6 Novel repressors can convert an activating transcription factor into a repressor in *Arabidopsis* stable lines

To verify the activity of our repressors in a more natural context, we fused one of our EAR repressors to a transcription factor and performed RTqPCR on the target of that transcription factor. We performed this experiment in stably-transformed *Arabidopsis thaliana*. For our test-case transcription factor we selected NTL8, a regulator of trichome development[24]. We made a fusion of NTL8 and one of our repressors, LDLNLPP, driven under the constitutive promoter pCH3[25] (Figure 4A). As expected, overexpression of NTL8 resulted in high trichome density whereas NTL8-LDLNLPP resulted in low trichome density (Figure 4B). As a target of NTL8, we selected TCL1, which is activated by wild-type NTL8[24]. While there was substantial variation in expression levels between independent lines, NTL8-LDLNLPP appears to have lower expression of TCL1 than NTL8 alone (Figure 4C). This suggests that not only do our repressors function in a different species from the one in which they were initially characterized, but also that they are capable of modulating the activity of transcription factors and therefore tuning natural gene regulatory networks.

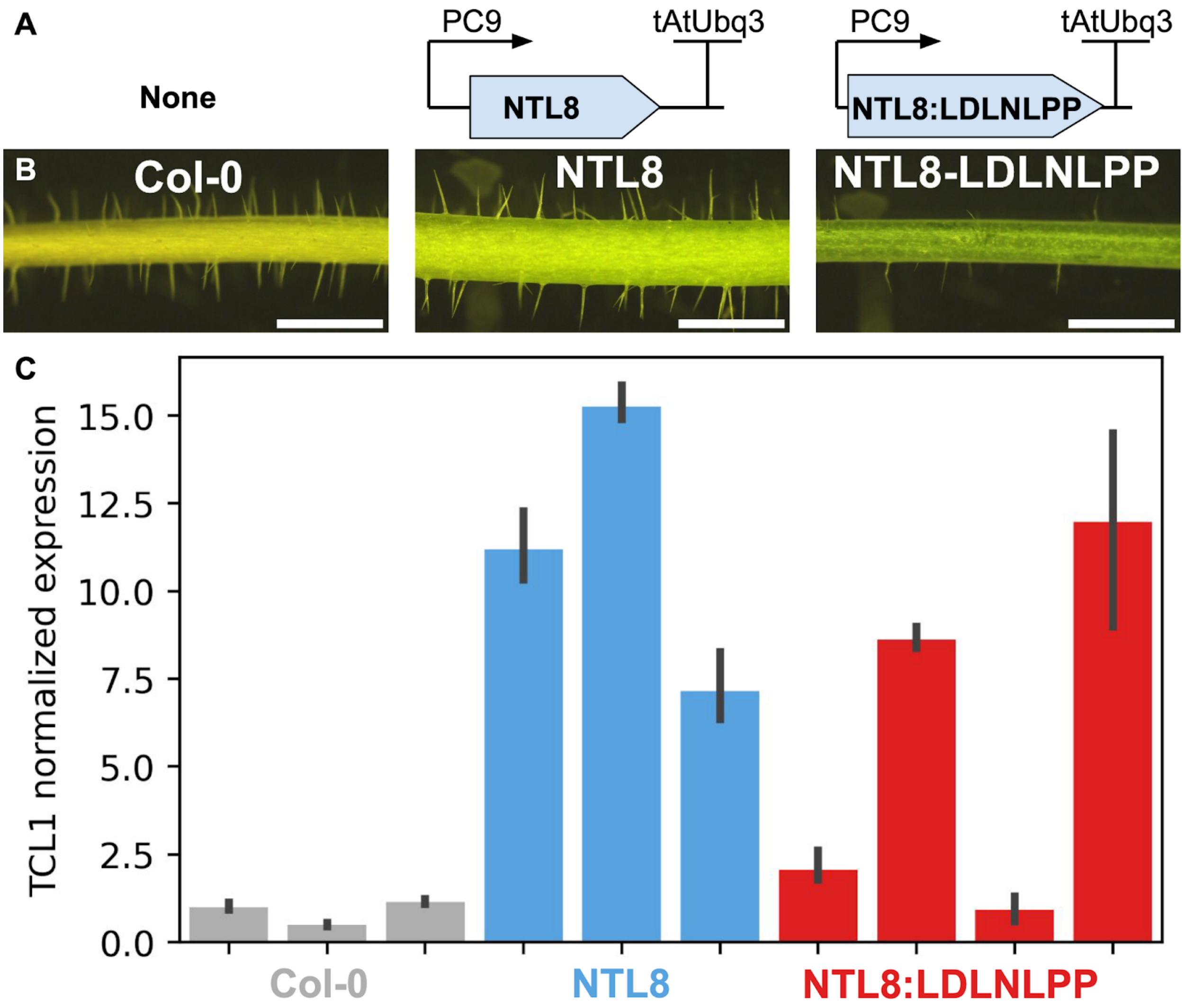


Figure 4: NTL8-repressor fusions can modulate trichome density. A) Details of the constructs used for *Arabidopsis* stable-line transformation. B) Images of the floral stem trichome density on T0 stably-transformed plants. Scale bar = 2 mm. C) RTqPCR data for the expression level of TCL1 normalized by Ef1α, with the average Col-0 expression set to 1. Primers are available in Supplemental Table 6, data are available in Supplemental Table 7. Error bars indicate the SEM.

5.4 Conclusion

With limited exceptions, the genome sequences of every cell of a given organism are identical. The massive variation in cellular structure, metabolism, and function between different tissue types within an organism is primarily generated through differences in the expression level of genes. Our set of transcriptional repressors enable more fine-grained control of the expression level of plant genes when expression levels lower than the natural baseline are required. These repressors will also likely function as targetable repressors when fused to catalytically-dead RNA-guided endonucleases, as SRDX has already been shown to do[23], enabling a wide dynamic range of gene expression tuning.

Ideally, it would be possible to compare the activity of all parts within a transcription expression tuning toolkit. To facilitate comparisons between repressor constructs tested in different datasets, we normalized all constructs within each of the three libraries by their two shared common elements, Gal4 and SRDX (Supplemental Table 8). While there are inherent limitations in making comparisons between constructs tested as part of different libraries, this dataset will serve as a simple combined resource to compare the approximately 200 constructs characterized in this study.

To our knowledge, this is the most comprehensive characterization of plant transcriptional repressors to date, with a particular focus on short repressors motifs which are especially useful for plant engineering and synthetic biology efforts where length can be a restriction. Indeed, these motifs are short enough that they can be cloned as primer tails, simplifying plasmid construction and enabling higher throughput testing. These repressors have broad value for plant synthetic biology, including applications such as the repression of native genes, modulating the function of native transcription factors, or the construction of synthetic genetic circuits. We hope that this transcriptional repression toolkit will enable more precise control of gene expression level in engineered plants.

5.5. Materials and methods

5.5.1 Plant growth conditions

Nicotiana benthamiana was grown following a previously published lab protocol[25]. Plants were grown in an indoor growth room at 25 °C and 60% humidity using a 16/8 hour light/dark cycle with a daytime PPFD of ~120 µmol/m^{2}s. Soil consisted of Sunshine Mix #4 (Sungro) supplemented with Osmocote 14-14-14 fertilizer (ICL) at 5mL/L and agroinfiltrated 29 days after seed sowing. *Arabidopsis thaliana* Col-0 were germinated and grown in Sunshine Mix #1 soil (Sungro) in a Percival growth chamber at 22 °C and 60% humidity using a 8/16 hour light/dark cycle with a daytime PPFD of ~200 µmol/m^{2}s.

5.5.2 N. benthamiana agroinfiltration assay

Each construct was infiltrated in conjunction with the GFP reporter into one leaf per plant, three plants per construct. For each leaf, 8 technical replicates of leaf disk were removed for analysis on a plate reader, which were averaged to form one biological data point.

Tobacco agroinfiltration was performed according to previously published standardized lab method[11], with the minor modification that a reporter construct with a stronger GFP promoter was used (pSynUAS19_WUS instead of pSynUAS17_WUS) in order to optimize the dynamic range of the assay for repressors. Fluorescence values recorded on an Omega Biotek plate reader were obtained for 8 leaf disks per plant, and those technical replicate values averaged to generate the by-biological-replicate values available in the supplemental tables.

5.5.3 Transformation of *Arabidopsis thaliana*

Floral dip transformation was performed according to a previously described protocol[26]. Seeds from floral dipped plants were selected on 50 mg/L Kanamycin in plates sealed with micropore tape, plate composition 1.5% plant TC agar, 1/2 MS with nutrients vitamins at pH 5.6. Seedlings were allowed to grow for 2 weeks in a Percival growth chamber dedicated to axenic plant growth with constant light at 24°C. After 2 weeks, transformant and non-transformant plants were easily differentiable via size (transformants larger), root length (transformants longer), cotyledon color (bleached yellow versus green), and the presence of true leaves (only present in transformant plants). At that point, transformant plants as well as Col-0 grown on plates without selection were transplanted onto Sunshine Mix #1 soil and grown as described in “Plant growth conditions”.

5.5.4 N-gram classification

Each repressor construct was broken into all 2, 3, and 4 grams, and the number of each of these n-grams was tallied within each quartile of the constructs organized by repression strength. In order to qualify as a “strong” n-gram, an n-gram must be at least 2x overrepresented in the strongest quartile of the constructs and underrepresented in the weakest quartile. To qualify as ‘weak’, the opposite must be true - 2x overrepresentation in the weakest quartile of repression constructs, and some amount of underrepresentation in the strongest quartile. To qualify as a “moderate” n-gram, an n-gram must not be over nor under represented by more than 25% in any of the four quartiles.

5.5.5 Microscopy

Arabidopsis floral stem trichomes were imaged using a Leica MZ16F stereo-microscope equipped with an Infinity 3 real-color camera (Teledyne Lumenera) with image capture into Infinity Analyze software.

5.5.6 RTqPCR

Total mRNA was extracted using E.Z.N.A. plant RNA kit (Omega Bio-tek) following manufacturer directions using the RB lysis buffer variation and on-column DNase digestion, cDNA synthesis was achieved with SSIV Vilo IV kit using random hexamers (Thermo Fisher). Quantitative PCR was performed using a CFX96 Real-Time thermocycler (Bio-Rad) programmed for detection of SYBR intercalating dye with the following temperature programming: 95 °C for 3 minutes, then 95 °C for 30 seconds, 60 °C for 45 seconds repeated 34 times, then a gradual increase from 65 °C to 95 °C at 0.5 °C / minute to generate melt curves. Sso-Advanced Universal SYBR Green Supermix (Bio-Rad) was used for qPCR amplification. A previously-validated primer set was used to amplify EF1α for internal normalization, three sets of primers for target gene TCL1 (AT2G30432) were designed with Benchling's qPCR primer design wizard and the best was selected on the basis of consistent amplification of template DNA, primer sequences are available in Supplemental Table 6. Melt curves for the product of all primer sets were unimodal and steep, suggesting only a single product was formed for each primer set. No reverse-transcriptase controls showed no amplification within the dynamic range of samples, confirming the efficacy of DNAse treatment, and no template controls instituted at the beginning of RNA extraction with no plant matter and kept in parallel with real samples throughout all molecular steps didn't amplify, confirming lack of contamination with extraneous DNA. Normalized relative expression was calculated using the delta-delta-Cq method, and normalized by setting the average level of amplification in the wild-type samples as 1.

5.5.7 Data analysis

Data was recorded in Google Sheets, all analyses were performed using Python implemented in Google Collab, data and code are available as supplements. Figures were assembled in Google Drawings.

5.8 Acknowledgements and funding

We thank Liam D. Kirkpatrick, Shuying Liu and Ruoming Cui for plant growth and maintenance. We thank Jutta Dalton for maintenance of plant growth facilities. This work was also conducted as part of the DOE Joint BioEnergy Institute (http://www.jbei.org) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.

5.9 Author contributions

KM and PS conceived of the experiments. KM performed the experiments and wrote the manuscript. JS and SW assisted with plant growth and agroinfiltration experiments. PS supervised, advised, and provided funding. All authors have read and approved the manuscript.

5.10 List of additional supplemental files available on request or online

Supplemental Table 1: Natural EAR repressor data

Supplemental Table 2: Ngrams associated with strong, medium, and weak repression in natural EAR dataset

Supplemental Table 3: SynEAR repressor data

Supplemental Table 4: Non-EAR repressor data

Supplemental Table 5: Tandem SRDX data

Supplemental Table 6: Primers used for RTqPCR

Supplemental Table 7: *Arabidopsis* stable line TF-repressor fusion RTqPCR data

Supplemental Table 8: Normalized merged averaged data for all repressors characterized in this study

Supplemental Code: Python notebook to run all analyses and produce all figures in this manuscript is available through Github: https://github.com/KaseyMarkel/Plant-Transcriptional-Repression-Toolkit/tree/main

5.11 References

(1) Watson-Lazowski, A.; Lin, Y.; Miglietta, F.; Edwards, R. J.; Chapman, M. A.; Taylor, G. Plant Adaptation or Acclimation to Rising CO2? Insight from First Multigenerational RNA-Seq Transcriptome. *Global Change Biology* **2016**, *22* (11), 3760–3773. https://doi.org/10.1111/gcb.13322.

(2) Stowe, E.; Dhingra, A. Development of the Arctic® Apple. In *Plant Breeding Reviews*; John Wiley & Sons, Ltd, 2021; pp 273–296. https://doi.org/10.1002/9781119717003.ch8.

(3) Lu, Z.; Yang, S.; Yuan, X.; Shi, Y.; Ouyang, L.; Jiang, S.; Yi, L.; Zhang, G. CRISPR-Assisted Multi-Dimensional Regulation for Fine-Tuning Gene Expression in Bacillus Subtilis. *Nucleic Acids Research* **2019**, *47* (7), e40. https://doi.org/10.1093/nar/gkz072.

(4) Souza, T. L. P. O.; Faria, J. C.; Aragão, F. J. L.; Del Peloso, M. J.; Faria, L. C.; Wendland, A.; Aguiar, M. S.; Quintela, E. D.; Melo, C. L. P.; Hungria, M.; Vianello, R. P.; Pereira, H. S.; Melo, L. C. Agronomic Performance and Yield Stability of the RNA Interference-Based Bean Golden Mosaic Virus-Resistant Common Bean. *Crop Science* **2018**, *58* (2), 579–591. https://doi.org/10.2135/cropsci2017.06.0355.

(5) Yang, J.; Liu, Y.; Yan, H.; Tian, T.; You, Q.; Zhang, L.; Xu, W.; Su, Z. PlantEAR: Functional Analysis Platform for Plant EAR Motif-Containing Proteins. *Front Genet* **2018**, *9*, 590. https://doi.org/10.3389/fgene.2018.00590.

(6) Hiratsu, K.; Mitsuda, N.; Matsui, K.; Ohme-Takagi, M. Identification of the Minimal Repression Domain of SUPERMAN Shows That the DLELRL Hexapeptide Is Both Necessary and Sufficient for Repression of Transcription in Arabidopsis. *Biochem Biophys Res Commun* **2004**, *321* (1), 172–178. https://doi.org/10.1016/j.bbrc.2004.06.115.

(7) Heyl, A.; Ramireddy, E.; Brenner, W. G.; Riefler, M.; Allemeersch, J.; Schmülling, T. The Transcriptional Repressor ARR1-SRDX Suppresses Pleiotropic Cytokinin Activities in Arabidopsis. *Plant Physiology* **2008**, *147* (3), 1380–1395. https://doi.org/10.1104/pp.107.115436.

(8) Cen, H.; Ye, W.; Liu, Y.; Li, D.; Wang, K.; Zhang, W. Overexpression of a Chimeric Gene, OsDST-SRDX, Improved Salt Tolerance of Perennial Ryegrass. *Sci Rep* **2016**, *6* (1), 27320. https://doi.org/10.1038/srep27320.

(9) Figueroa, P.; Browse, J. Male Sterility in Arabidopsis Induced by Overexpression of a MYC5-SRDX Chimeric Repressor. *The Plant Journal* **2015**, *81* (6), 849–860. https://doi.org/10.1111/tpj.12776.

(10) Kagale, S.; Links, M. G.; Rozwadowski, K. Genome-Wide Analysis of Ethylene-Responsive Element Binding Factor-Associated Amphiphilic Repression Motif-Containing Transcriptional Regulators in Arabidopsis. *Plant Physiology* **2010**, *152* (3), 1109–1134. https://doi.org/10.1104/pp.109.151704.

(11) Hummel, N. F. C.; Zhou, A.; Li, B.; Markel, K.; Ornelas, I. J.; Shih, P. M. The Trans-Regulatory Landscape of Gene Networks in Plants. *Cell Syst* **2023**, *14* (6), 501-511.e4. https://doi.org/10.1016/j.cels.2023.05.002.

(12) Belcher, M. S.; Vuu, K. M.; Zhou, A.; Mansoori, N.; Agosto Ramos, A.; Thompson, M. G.; Scheller, H. V.; Loqué, D.; Shih, P. M. Design of Orthogonal Regulatory Systems for Modulating Gene Expression in Plants. *Nat Chem Biol* **2020**, *16* (8), 857–865. https://doi.org/10.1038/s41589-020-0547-4.

(13) Suen, C. Y. N-Gram Statistics for Natural Language Understanding and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **1979**, *PAMI-1* (2), 164–172. https://doi.org/10.1109/TPAMI.1979.4766902.

(14) Osmanbeyoglu, H. U.; Ganapathiraju, M. K. N-Gram Analysis of 970 Microbial Organisms Reveals Presence of Biological Language Models. *BMC Bioinformatics* **2011**, *12* (1), 12. https://doi.org/10.1186/1471-2105-12-12.

(15) Islam, S. M. A.; Heil, B. J.; Kearney, C. M.; Baker, E. J. Protein Classification Using Modified N-Grams and Skip-Grams. *Bioinformatics* **2018**, *34* (9), 1481–1487. https://doi.org/10.1093/bioinformatics/btx823.

(16) Visa, S.; Ramsay, B.; Ralescu, A.; Knaap, E. Confusion Matrix-Based Feature Selection.; 2011; Vol. 710, pp 120–127.

(17) Kramer, O. Scikit-Learn. In *Machine Learning for Evolution Strategies*; Kramer, O., Ed.; Springer International Publishing: Cham, 2016; pp 45–53. https://doi.org/10.1007/978-3-319-33383-0_5.

(18) Ikeda, M.; Ohme-Takagi, M. A Novel Group of Transcriptional Repressors in Arabidopsis. *Plant and Cell Physiology* **2009**, *50* (5), 970–975. https://doi.org/10.1093/pcp/pcp048.

(19) Fisher, A. L.; Ohsako, S.; Caudy, M. The WRPW Motif of the Hairy-Related Basic Helix-Loop-Helix Repressor Proteins Acts as a 4-Amino-Acid Transcription Repression and Protein-Protein Interaction Domain. *Molecular and Cellular Biology* **1996**. https://doi.org/10.1128/MCB.16.6.2670.

(20) Payankaulam, S.; Arnosti, D. N. Groucho Corepressor Functions as a Cofactor for the Knirps Short-Range Transcriptional Repressor. *Proceedings of the National Academy of Sciences* **2009**, *106* (41), 17314–17319. https://doi.org/10.1073/pnas.0904507106.

(21) Fisher, A. L.; Caudy, M. *Groucho proteins: transcriptional corepressors for specific subsets of DNA-binding transcription factors in vertebrates and invertebrates*. https://doi.org/10.1101/gad.12.13.1931.

(22) *Toward controlling gene expression at will: Specific regulation of the erbB-2/HER-2 promoter by using polydactyl zinc finger proteins constructed from modular building blocks*. https://doi.org/10.1073/pnas.95.25.14628.

(23) Lowder, L. G.; Zhang, D.; Baltes, N. J.; Paul, J. W.; Tang, X.; Zheng, X.; Voytas, D. F.; Hsieh, T.-F.; Zhang, Y.; Qi, Y. A CRISPR/Cas9 Toolbox for Multiplexed Plant Genome Editing and Transcriptional Regulation. *Plant Physiol.* **2015**, *169* (2), 971–985. https://doi.org/10.1104/pp.15.00636.

(24) Tian, H.; Wang, X.; Guo, H.; Cheng, Y.; Hou, C.; Chen, J.-G.; Wang, S. NTL8 Regulates Trichome Formation in Arabidopsis by Directly Activating R3 MYB Genes TRY and TCL1. *Plant Physiol* **2017**, *174* (4), 2363–2375. https://doi.org/10.1104/pp.17.00510.

(25) Zhou, A.; Kirkpatrick, L. D.; Ornelas, I. J.; Washington, L. J.; Hummel, N. F. C.; Gee, C. W.; Tang, S. N.; Barnum, C. R.; Scheller, H. V.; Shih, P. M. A Suite of Constitutive Promoters for Tuning Gene Expression in Plants. *ACS Synth Biol* **2023**, *12* (5), 1533–1545. https://doi.org/10.1021/acssynbio.3c00075.

(26) Zhang, X.; Henriques, R.; Lin, S.-S.; Niu, Q.-W.; Chua, N.-H. Agrobacterium-Mediated Transformation of Arabidopsis Thaliana Using the Floral Dip Method. *Nat Protoc* **2006**, *1* (2), 641–646. https://doi.org/10.1038/nprot.2006.97.

Chapter 6: Conclusion

Plants comprise the majority of Earth's biomass, and are the base of all terrestrial ecosystems. The capacity to control plant growth is extremely narrowly distributed among the animal kingdom, and by far most strongly expressed among humans. In human prehistory, the development of agriculture – fundamentally, the emergence of the capacity to manipulate the distribution and properties of plants – catalyzed the emergence of civilization.

Advances in agricultural technology have been one of the most prominent driving forces in shaping the history of civilization, and have enabled the exponential rise in the human population. At present, improvements at the level of agronomic practices as well as plant genetics are driving a massive reduction in the rate of human malnutrition. If present trends continue, widespread malnutrition may well be eliminated within the century, as was mandated by the UN sustainable development goals for the year 2030.

While plants are a grand mover and shaker, a key shaping element of the human historical story, each individual human is merely one small pebble in the river of history, making the slightest diversion to the overall flow of events. The contribution I aspire to achieve throughout my career is to marginally increase humanity's capacity to engineer plants to our benefit, and to that effort I have engaged in research at several different levels of abstraction aimed towards that goal.

In Chapter 1, I aimed to contribute a conceptual framework widespread in classical breeding but, in my opinion, insufficiently well known in the plant biotechnology world. In Chapter 2, I analyzed the ability of one of the relatively few species other than humanity that has a clear capacity to intentionally modulate the growth forms of plants, gall inducing wasps. This project yielded many insights into that naturally-occurring plant engineering process, though the human application of similar principles may yet be far off. In Chapter 3, I present the first replication of what is perhaps the most exciting research finding in plant biology during my lifetime: the potential of a mammalian gene to massively increase the yield of crop plants. While the mechanism behind this action remains unclear, we present the first plausible mechanistic hypotheses, at least some of which are likely at least partially explanatory for this remarkable finding. This project has the most direct potential for application to continue humanity's quest to improve plant yield and agronomic properties, indeed no further elaboration is required to spell out how that might be achieved. Chapter 5 presents a large toolkit developed for the control of plant gene expression levels, which may facilitate future plant genetic engineering efforts and enable more complex synthetic biology such as metabolic engineering or genetic circuits.

I hope that these contributions I have made to plant biology as a small pebble in the streambed will ripple down the flow of history, and aspire to continue to have impact as long as this pebble may last.