Effects of genetic variants on bolting asynchrony in Arabidopsis using eQTL analysis

Miguel Alburo
University of York

6th May, 2024

Abstract

Asynchrony in bolting time causes problems in agriculture by preventing an optimal harvest which maximises crop yield and quality. Utilizing results and data from Redmond et al. 2023, we demonstrate that subtle genetic differences can influence the stochastic phenotype of bolting time in Arabidopsis thaliana. Employing a novel methodology of performing eQTL analysis in a single-plant-omics background with a near-isogenic population, we determined variants associated with bolting through differential expression of key bolting regulators. This study emphasises accounting for genetic diversity, even with highly inbred lines; and the potential of single-plant-omics combined with eQTL analysis to scout for functional variants.

1 Introduction

1.1 Expression quantitative trait loci influence gene expression

Genetic variants, or variants for short, are specific variants/versions of loci in the genome. These include single nucleotide polymorphisms (SNPs) for the majority, but also insertions/deletions (indels) as well as some other, rarer types (Fig. 1). Variants that are significantly associated with the differential expression of a gene are referred to as expression-quantitative-trait-loci (eQTL) for that gene.

Biological mechanisms behind eQTL interactions are complex and diverse (Nica and Dermitzakis 2013) and are usually characterised based on the proximity of the variant to its target gene. Associated variants/eQTL within close proximity to the target gene are *cis*-acting whilst eQTL distant from their target genes are termed *trans*-acting. *cis*-eQTL located in regulatory regions such as promoters, have been found to influence the binding of transcription factors resulting in the differential expression of genes located nearby (Michaelson, Loguercio, and Beyer 2009). Mechanisms for *trans*-eQTL are more complex and context-dependent though *cis*-acting eQTL are frequently disguised as *trans*-acting. This happens because a *cis*-eQTL targets a regulatory gene, which subsequently forms a domino of correlations on downstream genes (Fig. 2; Nica and Dermitzakis 2013). Many studies, such as this one, use a 1 megabase threshold to distinguish between *cis*/*trans*-eQTL (Bryois et al. 2021; Yan Liu et al. 2020). However, it is important to note that using a distance-based threshold only estimates whether a specific eQTL was *cis*/*trans*-acting. Validating the type of interaction requires a detailed investigation into the loci itself, as will be demonstrated later.

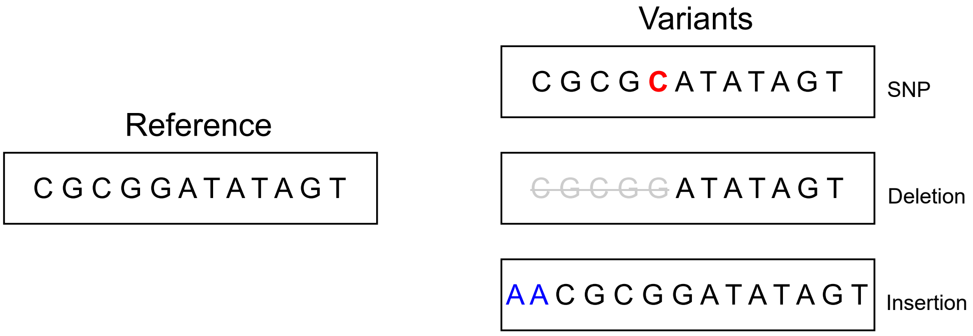


Figure 1: *Diagram explaining 3 main types of variants: single nucleotide polymorphisms (SNPs) refer to single base changes, deletions and insertions refer to one or more bases being deleted or inserted into a sequence. Note that we are comparing these changes against a reference genome, which may not always be the most representative genome of a population.*

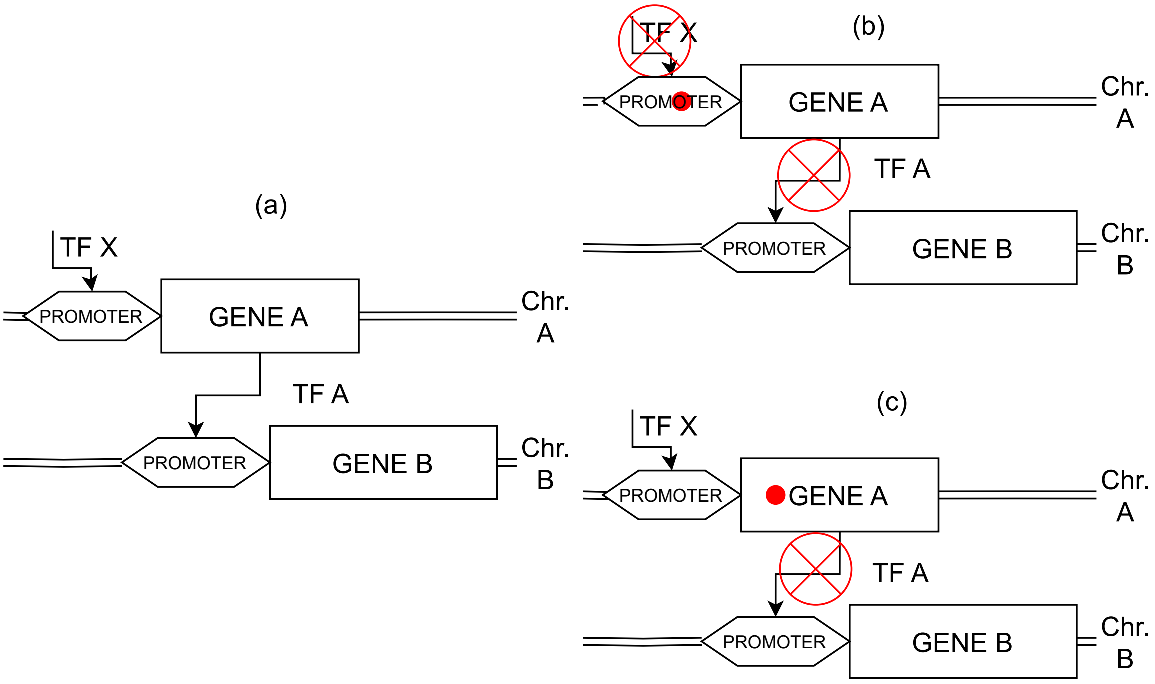


Figure 2: *Example mechanisms of trans-eQTL interactions. (a) In a given module of a gene regulatory network, transcription factor X binds to the promoter region, stimulating the expression of gene A which then produces transcription factor A which binds to the promoter of gene B, activating gene B. (b) A variant located in the promoter of gene A prevents or slows TF X binding, thus resulting in less/no TF A production. (c) A variant located in the coding sequence of gene A produces an altered TF A, which may inhibit or reduce the expression of gene B.*

eQTL analyses are performed in order to map variants significantly correlated with expression of genes. The analysis workflow starts with the collection and preparation of data. This requires genotyping and RNA sequencing every sample so that there is genotype data for all variants and gene expression values across all genes. Then the analysis itself is performed which involves computing linear regressions, taking variant genotype as the explanatory variable and gene expression as the response variable to derive significant variant-gene pairs.

eQTL and similar concepts have been studied since microarrays but are on the rise following innovation in modern RNA sequencing technologies, making genomics studies cheaper, faster, and better (Slatko, Gardner, and Ausubel 2018). Massive efforts have been made to characterise functional variants in humans, which have demonstrated the potential of eQTL analysis in the biomedical field (GTEx Consortium 2020; Bryois et al. 2021) such as characterization of disease-causing eQTL (Y. Chen et al. 2008).

1.2 Single-plant-omics for investigating developmental asynchrony

The problem that most studies have when performing eQTL analysis is that there is often a huge number of variants in their sample population. Plant eQTL and QTL studies (Galpaz et al. 2018; Keurentjes et al. 2007; Sonah et al. 2015) usually involve crossing 2 genetically diverse parents from different accessions. Subsequent generations are then crossed together which eventually leads to the distribution of different variants into different recombinant inbred lines (RILs). This approach, though done intentionally to incorporate variants for analysis, can generate tens of thousands to millions of variants, resulting in lower statistical power (Nica and Dermitzakis 2013). Furthermore, to account for the effects of linkage disequilibrium, either a genetic map would need to be constructed which bins linked variants into groups, or only a smaller subset of all variants must be selected instead (Galpaz et al. 2018; Keurentjes et al. 2007; Sonah et al. 2015). These steps make it difficult to identify the exact functional variants causing expression variation.

In this project, we demonstrate that the novel single-plant-omics approach is an effective platform for exploring plant developmental asynchrony (Redmond et al. 2023) and performing eQTL analysis. Single-plant-omics has been used previously to study inter-plant transcriptional variation in Arabidopsis (Cortijo et al. 2019) and phenotype and function characterisation of genes in maize (Cruz et al. 2020). In single-plant-omics, plant samples come from a highly inbred population originating from a single accession/ecotype (Cortijo et al. 2019; Cruz et al. 2020), resulting in very low genetic diversity and a small number of variants. These studies have used highly inbred population and assume that these are practically isogenic and so do not consider the effects of the small number of variants in their experimental outcome. However, we show that this makes it far more effective for performing an eQTL study since the statistical power is higher and associated variants identified through regressions are more likely to be truly causal rather than just correlated.

1.3 Bolting asynchrony prevents an optimal harvest time

One such example of a developmental process subject noise is bolting. Bolting refers to the initial, swift growth of the flowering stem. More importantly, it acts as the visual indicator of the plant's transition from vegetative to reproductive phases C. Chen et al. 2019. Within a population, bolting occurs asynchronously, with the stochastic phenotype of bolting time being a function of environmental and genetic variables, as well as inherent randomness (Klingenberg 2019).

This asynchronous nature poses real issues in agriculture by affecting crop yields and quality. This problem is further complicated by research showing climate change increases the degree of flowering asynchrony (Zohner, Mo, and Renner 2018) resulting in potential risks to agriculture and food security. Lettuce, for example, has the tendency to turn bitter and stop growing when bolted. Scheduling for an ideal harvest time is challenging since a farmer would also like to wait for the lettuce to grow as large as possible. The ability to predict bolting time, and even better still control bolting time, would be valuable to a farmer for optimising crop yields and quality. These points highlight the importance of researching to better understand the nature of developmental processes such as bolting. In fact, recent studies were already effective in modelling and predicting flowering time in maize (Azodi et al. 2020). It may become a possibility to screen and select for bolting genotypes, especially with modern sequencing technologies such as NGS which continue to become cheaper, faster, and more accurate (Slatko, Gardner, and Ausubel 2018).

1.4 Variants are linked to bolting through pseudotime

During the onset of bolting, the expression across many genes change (Redmond et al. 2023), aligning with the switch from vegetative to reproductive states. With such a sudden, almost-binary change in the transcriptome, it is difficult to resolve a transcriptional time series as usually done. Recently, however, Redmond et al. 2023 was successful in deconstructing the series of transcriptional events that transpired during bolting, and further identifying the gene regulatory network governing it. Using single-plant-omics, samples were RNA sequenced once half the population had bolted, and using pseudotime inference, they were able to derive a biological age metric for each plant sample. This metric, termed pseudotime, was calculated based on an independent subset of genes in which expression was proportional to gradually developing biomass and leaf size traits. By ordering plants based on pseudotime, analogous to a time-series, they were able to identify genes in which the activity had shifted at the moment of bolting.

As a brief section at the end, they consider if genetic variation had any influence on developmental noise through correlations with pseudotime, biomass or leaf size. Furthermore, they found different clusters of variants where alleles transitioned at the point of bolting. These results were quite unexpected since, by using a near-isogenic population with only few variants, genetic variation was not expected to show such an effect on developmental noise. The main takeaway from their results is that there exist multiple independent complex biological mechanisms in which functional genetic variants influence pseudotime/developmental rates leading to bolting.

1.5 Identifying bolting-associated variants with eQTL analysis.

The identification of variants/loci associated with a trait of interest would typically be done in a genome-wide association study (GWAS) or quantitative-trait-loci (QTL) analysis which performs regressions directly between genotype and trait. This method can successfully locate trait-associated variants as well as model how specific genotypes influence the phenotype. However, it is difficult to interpret how variants biologically interact to control the complex phenotype (Nica and Dermitzakis 2013). We instead aim to find functional variants that control the regulatory genes involved in the bolting process, which we will call bolting regulators for the rest of the paper (Redmond et al. 2023). This same workflow of characterising candidate genes in a TWAS before performing eQTL analysis for variants was already performed in Arabidopsis (Chien et al. 2023) and also cotton (Li et al. 2020). Using this method, we will achieve the same goal of identifying bolting-related variants. Furthermore, by performing eQTL analysis we also paint a better picture of the variant-bolting interaction through our extended knowledge of the intermediary transcriptome phenotype (Fig. 3).

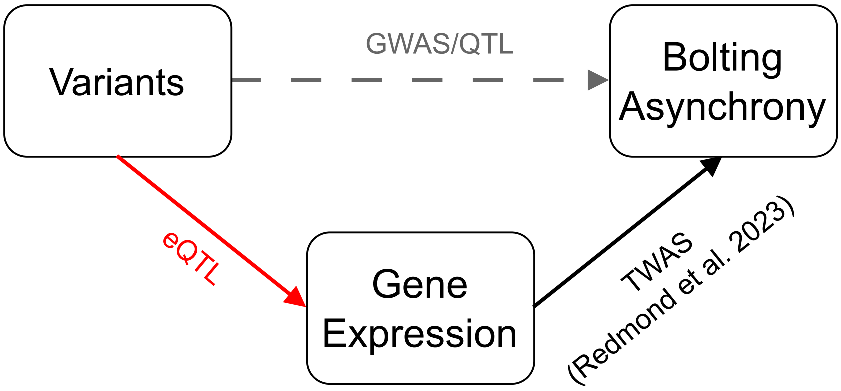


__Figure 3:__ Diagram detailing the pathway in which bolting asynchrony occurs along with different analysis techniques used to explore causal relationships between the genome, transcriptome and phenotype.

In the upcoming sections of this report, we perform eQTL analysis on a near-isogenic population of Arabidopsis (Redmond et al. 2023). Despite a near-isogenic population obtained in a single-plant-omics setting, we reveal a surprising number of eQTL. Afterward, we determine which specific variants are related to the bolting process through control of bolting regulators and additional clustering reveals clusters of variants which would lead to early or later bolting. In analysing the spatial distribution of these eQTL across the genome, we determine that our eQTL are all trans-acting and discover an eQTL void followed by a hotspot on chromosome 3. We also notice how distally located variants exhibit haplotypes akin to patterns caused by linkage disequilibrium. Finally, we zoom in on some of our top variants and confirm they represent nonsynonymous changes within exon regions of genes which are shown to be involved in stress. ORA of the genes which the bolting-related variants are within produce overrepresented terms also for stress responses. Our findings indicate that the bolting-related variants alter more principal developmental pathways which, as a side effect, leads to asynchrony in bolting.

2 Results

2.1 Vast number of eQTL in near-isogenic population

It was previously shown that several variants had strong correlations with pseudotime (Redmond et al. 2023). Following from this, we wanted to determine if the effects of variants on plant samples were statistically significant. Specifically, we wanted to identify variants which were significantly associated with differential expression of genes. Obtained through single-plant-omics methods, our population dataset included the expression levels of 19,208 pre-filtered genes and the genotypes of only 2012 protein-coding variants. eQTL analysis using the MatrixEQTL R package identified a surprising 202,539 eQTL pairs in the system. These included 377 variants associated with the differential expression of 9395 genes which had fallen under an FDR-adjusted p-value of 1e-5. These results account for 19% of all variants and 49% of all genes. This high number of interactions underscores the sensitivity of plant transcriptomes to genetic differences, even when restricted to protein-coding regions. Additionally, we have demonstrated the efficacy of single-plant-omics in generating robust data from a highly inbred population, enabling the identification of numerous significant eQTL despite the small sample size (n=65).

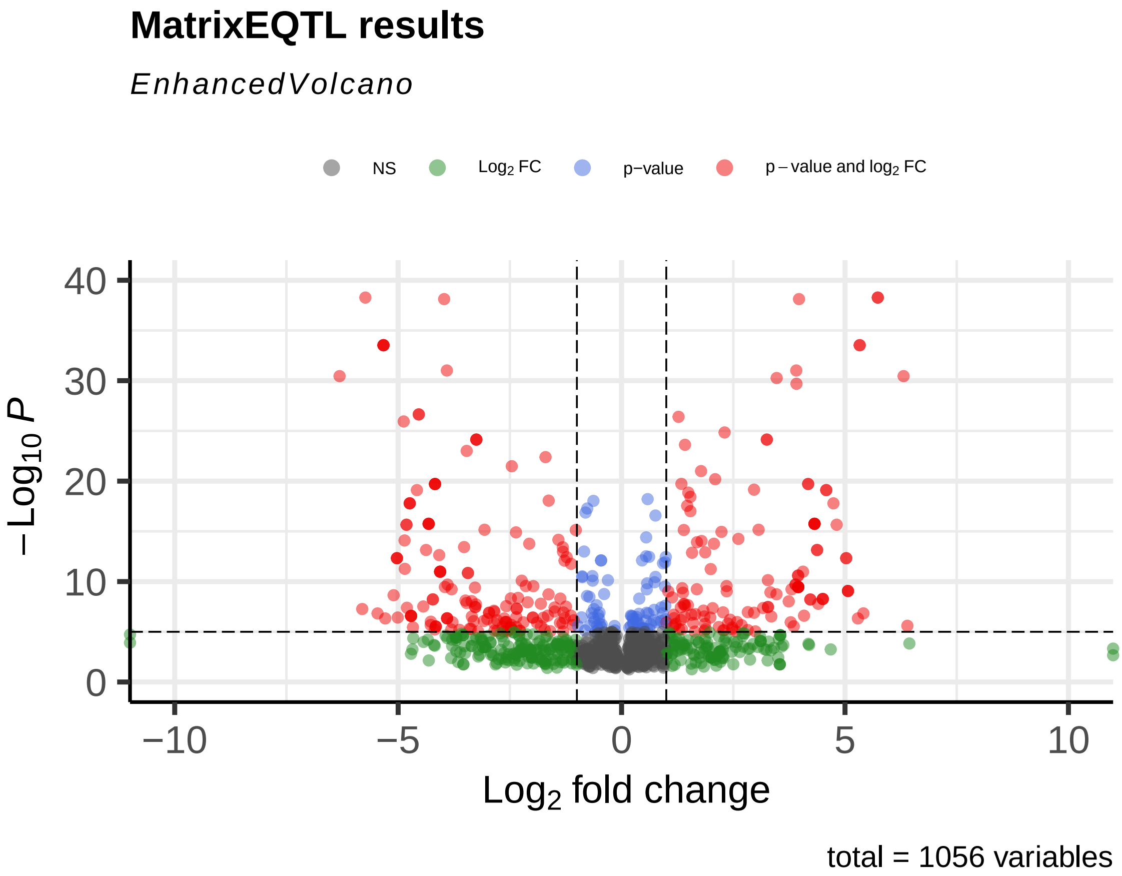
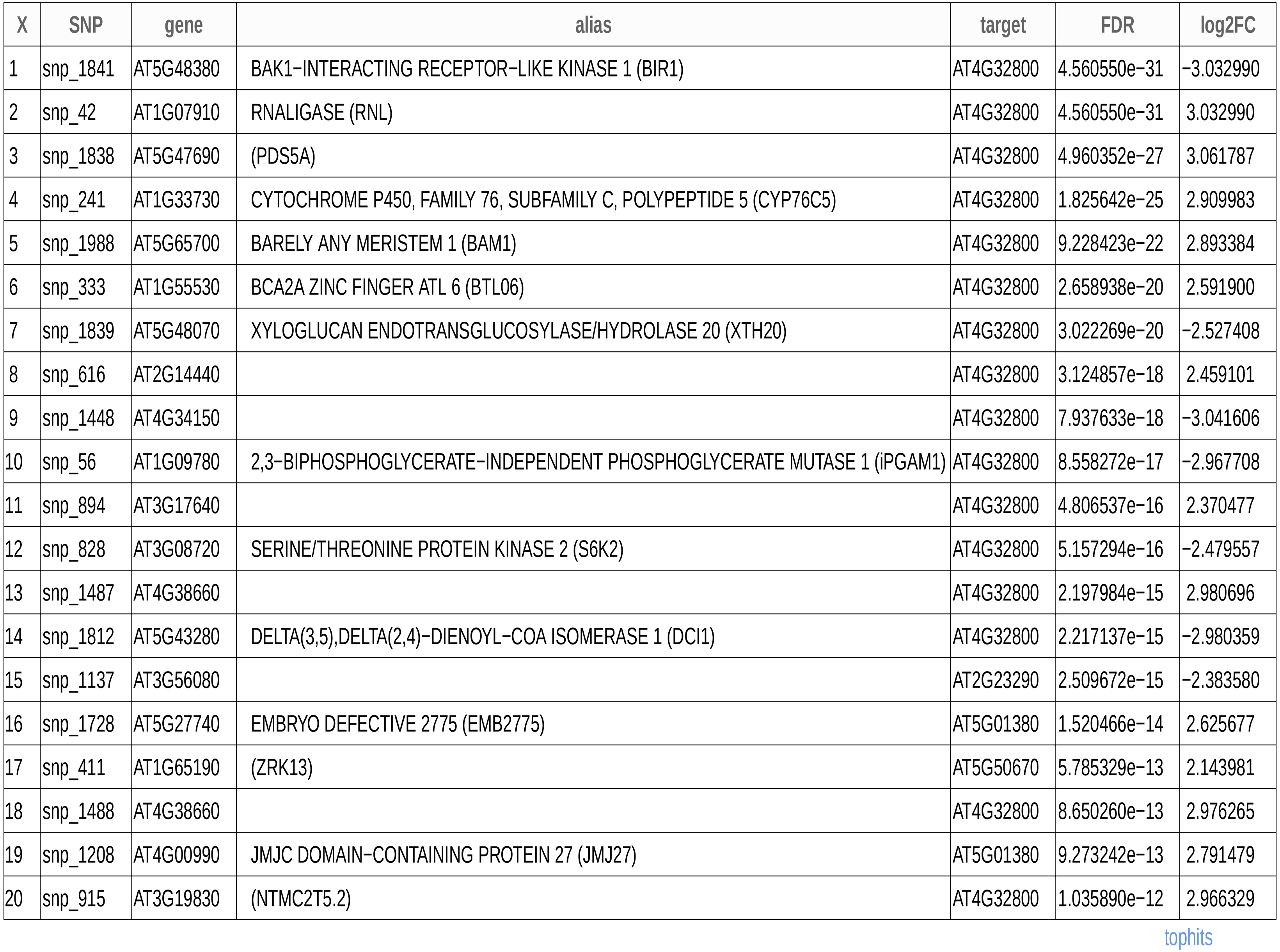


Figure 4: *Volcano plot of general eQTL interactions. For each unique variant, only the most significant eQTL interaction was plotted.*



X	SNP	gene	alias	target	FDR	log2FC
1	snp_1841	AT5G48380	BAK1−INTERACTING RECEPTOR−LIKE KINASE 1 (BIR1)	AT4G32800	4.560550e−31	−3.032990
2	snp_42	AT1G07910	RNALIGASE (RNL)	AT4G32800	4.560550e−31	3.032990
3	snp_1838	AT5G47690	(PDS5A)	AT4G32800	4.960352e−27	3.061787
4	snp_241	AT1G33730	CYTOCHROME P450, FAMILY 76, SUBFAMILY C, POLYPEPTIDE 5 (CYP76C5)	AT4G32800	1.825642e−25	2.909983
5	snp_1988	AT5G65700	BARELY ANY MERISTEM 1 (BAM1)	AT4G32800	9.228423e−22	2.893384
6	snp_333	AT1G55530	BCA2A ZINC FINGER ATL 6 (BTL06)	AT4G32800	2.658938e−20	2.591900
7	snp_1839	AT5G48070	XYLOGLUCAN ENDOTRANSGLUCOSYLASE/HYDROLASE 20 (XTH20)	AT4G32800	3.022269e−20	−2.527408
8	snp_616	AT2G14440		AT4G32800	3.124857e−18	2.459101
9	snp_1448	AT4G34150		AT4G32800	7.937633e−18	−3.041606
10	snp_56	AT1G09780	2,3−BIPHOSPHOGLYCERATE−INDEPENDENT PHOSPHOGLYCERATE MUTASE 1 (iPGAM1)	AT4G32800	8.558272e−17	−2.967708
11	snp_894	AT3G17640		AT4G32800	4.806537e−16	2.370477
12	snp_828	AT3G08720	SERINE/THREONINE PROTEIN KINASE 2 (S6K2)	AT4G32800	5.157294e−16	−2.479557
13	snp_1487	AT4G38660		AT4G32800	2.197984e−15	2.980696
14	snp_1812	AT5G43280	DELTA(3,5),DELTA(2,4)−DIENOYL−COA ISOMERASE 1 (DCI1)	AT4G32800	2.217137e−15	−2.980359
15	snp_1137	AT3G56080		AT2G23290	2.509672e−15	−2.383580
16	snp_1728	AT5G27740	EMBRYO DEFECTIVE 2775 (EMB2775)	AT5G01380	1.520466e−14	2.625677
17	snp_411	AT1G65190	(ZRK13)	AT5G50670	5.785329e−13	2.143981
18	snp_1488	AT4G38660		AT4G32800	8.650260e−13	2.976265
19	snp_1208	AT4G00990	JMJC DOMAIN−CONTAINING PROTEIN 27 (JMJ27)	AT5G01380	9.273242e−13	2.791479
20	snp_915	AT3G19830	(NTMC2T5.2)	AT4G32800	1.035890e−12	2.966329

Figure 5: *Table containing results of our top bolting eQTL. Each row is a unique variant along with their most significant eQTL interaction. Rows are ordered in increasing FDR.*

2.2 Regulators of bolting correlate with variant genotypes

To find which variants were most associated with bolting, we filtered our results to include only eQTL which targeted the expression of bolting regulators whilst also taking the effect size into account. This left us with 647 eQTL interactions comprising 75 variants controlling the expression of 20 bolting regulators. In these results, we found that a subset of 11 regulators represented targets for a large portion (84%) of eQTL pairs (Fig. 6). This suggests that many bolting-related eQTL interactions result from indirect downstream effects, potentially mediated by trans-eQTL affecting principal developmental processes. Furthermore, since these bolting regulators act in order, a causal variant affecting an upstream regulator would therefore result in correlations with downstream regulators. This complex web of interactions illustrates the difficulty in determining causality from correlation within our eQTL data, which will be a theme observed across our results.

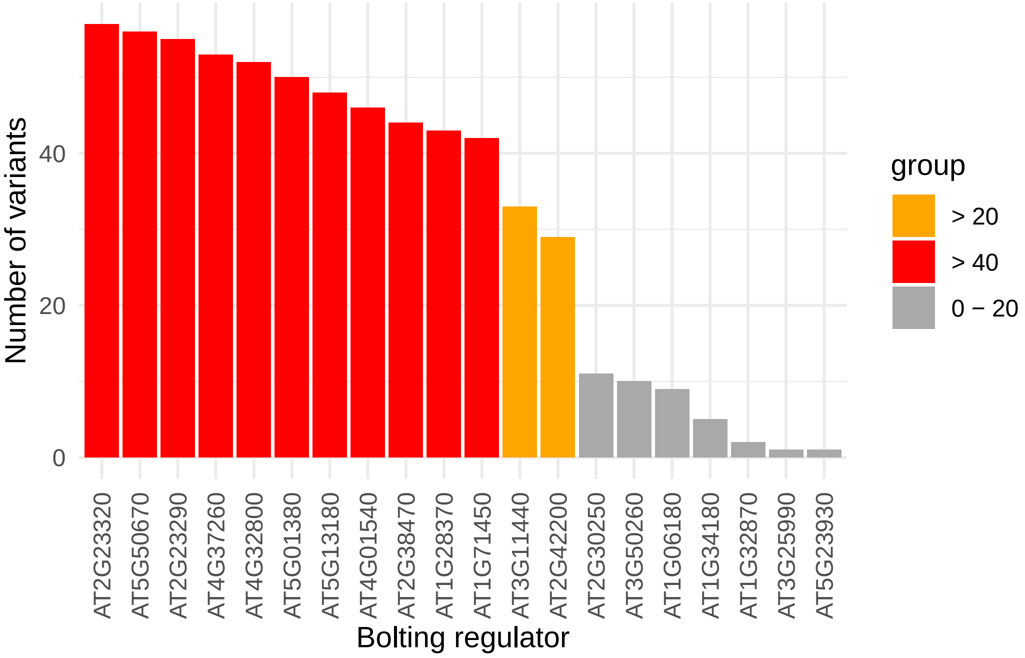


Figure 6: *Ordered histogram of genes and the number of variants which they are the target of.*

2.3 eQTL interactions are dominated by *trans*-acting loci

Given that our data included only coding region variants, we hypothesized that all identified eQTL were trans-acting since cis-eQTL are commonly located in regulatory regions. To test this claim, we constructed a cis-trans graph which plots the location of the variant against the location of its target gene (Fig. 7). We can observe that data points do not align with the x = y slope which would determine a cis-interaction. Coupled with an extremely low R-squared statistic, these results imply the absence of cis-acting eQTL in our data. We can still observe both horizontal and vertical bands which add to the complexity of our eQTL interaction network. As mentioned previously, these bands are likely a product of downstream effects.

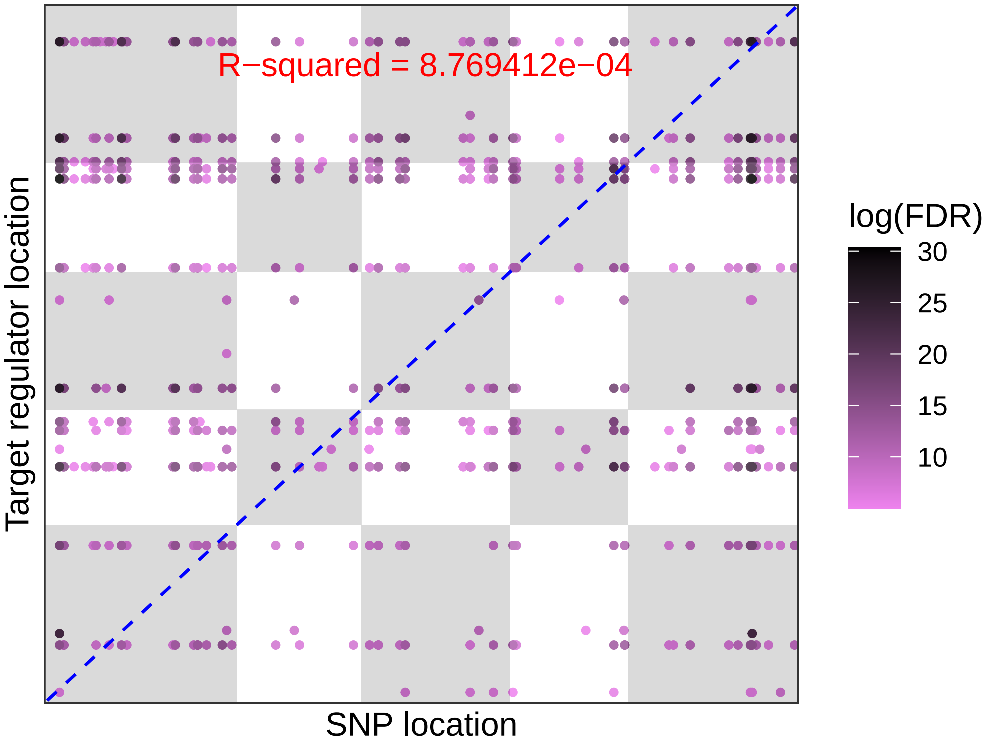


Figure 7: *Cis-trans plot of eQTL. Variant positions are plotted against their target gene position. The y = x slope broadly indicates where cis-eQTL data points should be located.*

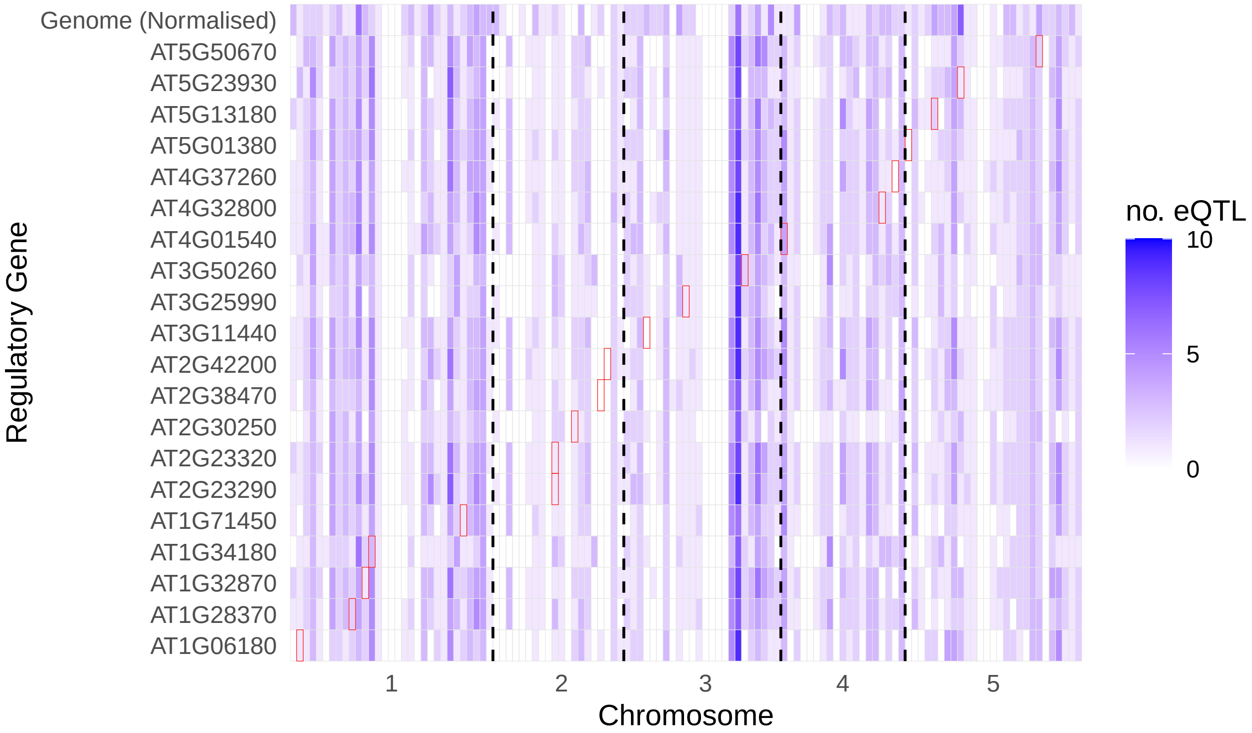


Figure 8: *Heatmap showing the density of eQTL in 1Mb blocks across the genome, for each bolting regulator. The top row represents a normalized eQTL density across the genome for comparison.*

To better understand spatial distribution, we decided to consider the density of associated eQTL loci in each 1 megabase block spanning the genome (Fig. 8). We noted a distinct 4 Mb region of low eQTL density around the centromere on chromosome 3, followed by a high-density region between 17-19 Mb. This pattern is likely an artefact of our data since centromeres are regions of very low gene density (Kendal and Suomela 2005). Based on this, a scarce number of bolting-related protein coding variants is naturally expected. This void can also be seen in Fig. cis-trans and at other centromeres, though smaller.

2.4 Early or late bolting is characterised by different variant clusters

Following the identification of bolting-related variants and the complex multi-layered eQTL interaction network, we wanted to uncover any patterns in interaction by considering the effect size of the expression change. Clustering the eQTL by their effect size revealed different sets of variants are associated with the increased or decreased expression of bolting regulators respectively (Fig. 9). The genes AT1G06180, AT2G42200 and AT5G13180 were exceptions in which the expression effect size/fold change reverses. After referring to the bolting gene regulatory network (Redmond et al. 2023), we found that the large cluster of regulators had decreased during bolting (decreasing TFs) whilst genes AT1G06180, AT2G42200 and AT5G13180 increased expression during bolting (increasing TFs).

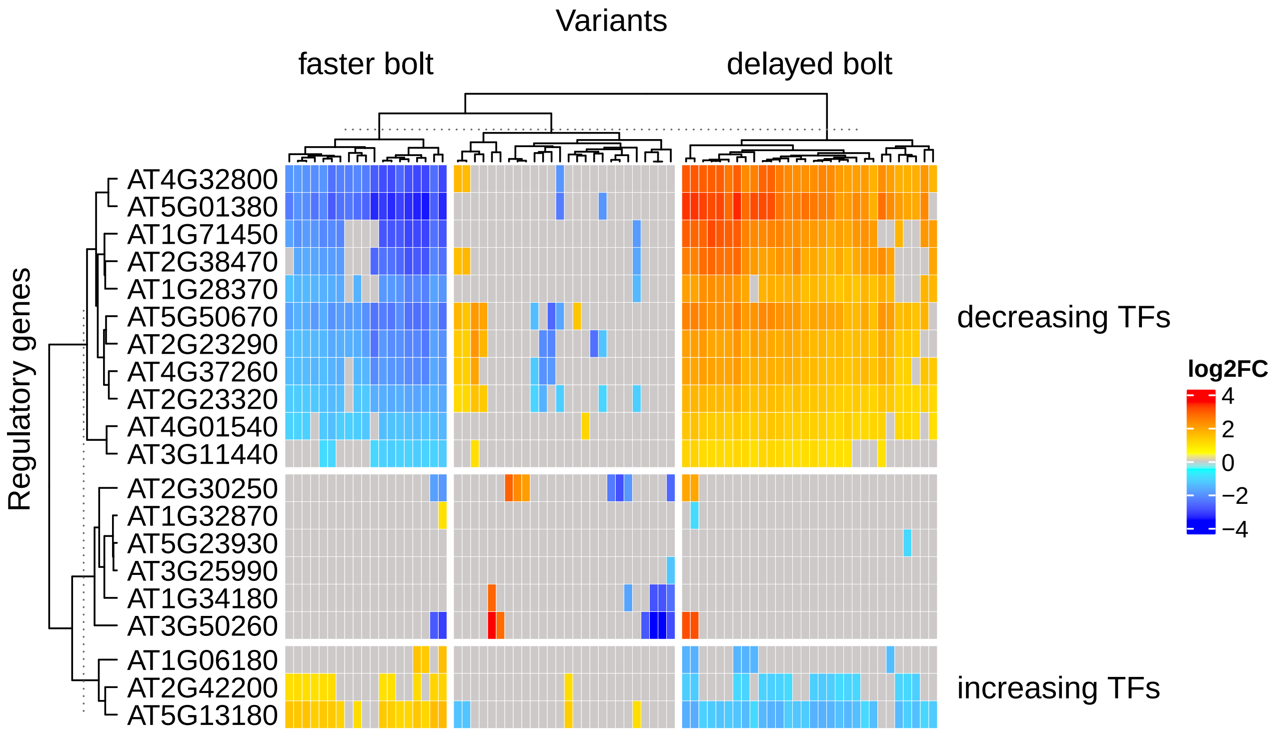


Figure 9: *Heatmap of eQTL interactions, with variants along the columns, and bolting regulators along the rows. Cells are then coloured based on the log2 fold-change between the variant and target gene. Clustering algorithms were used to order genes and variants in the heatmap accordingly [see 'Methods and data'].*

These results from Fig. 9 suggest the two variant clusters would oppositely impact bolting time. The variants on the right correlate with increased expression of decreasing TFs and decreased expression of increasing TFs. Since this would act against the natural flow of the bolting process, the variants on the right-hand side would likely cause a later bolting time. The cluster of variants on the left-hand side would have the opposite effect, causing an earlier bolting time. Even without any understanding of the biological mechanisms behind bolting, these results hint at potential pathways with opposite effects, through which specific genetic variants may expedite or delay bolting. This information can be directly tested in breeding programs aimed at optimizing bolting times.

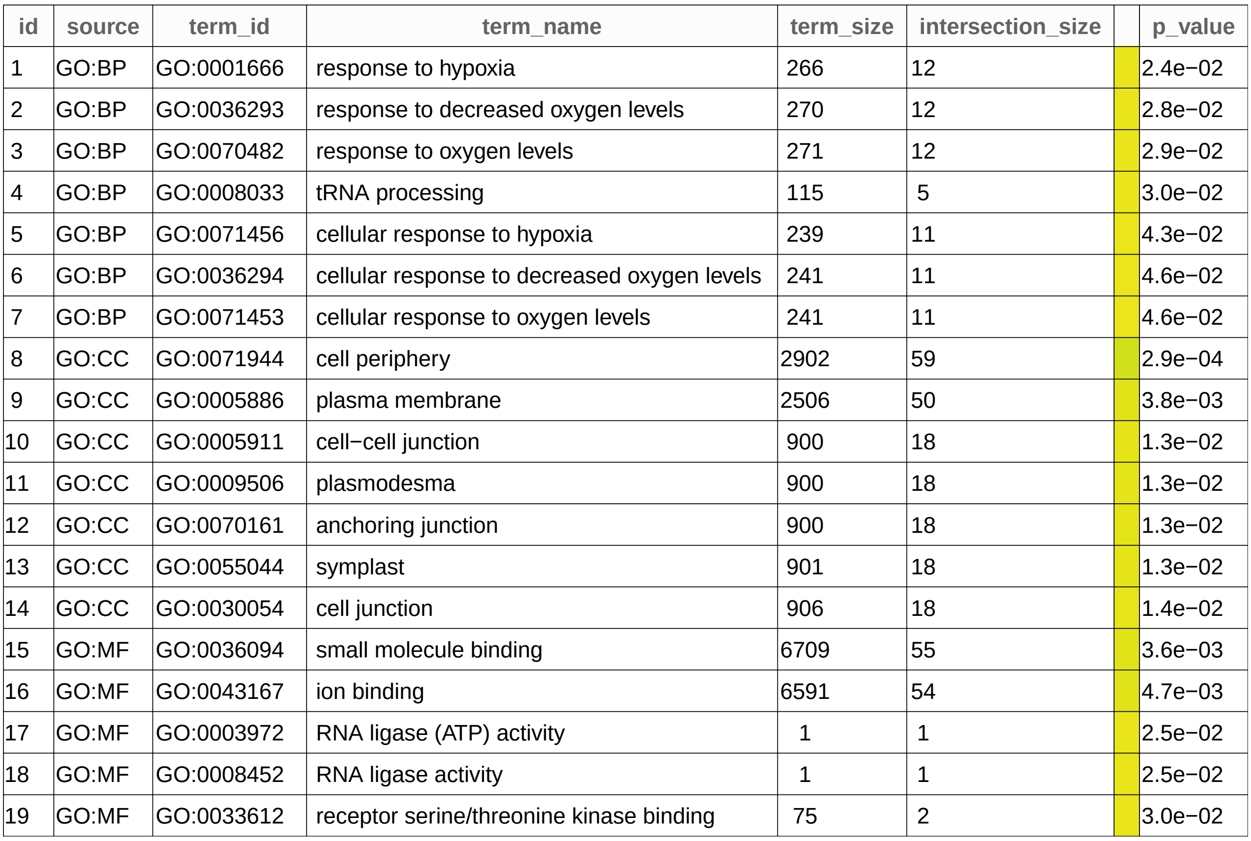
2.5 Epistasis potentially induces linkage-like behaviour

Interestingly, we also found several cases where different variants with the same target regulator had the exact same FDR and equal/opposite effect sizes (Fig. 5, snp_42 - snp_1841). We inspected their genotypes and found that a specific allele combination would be repeated across samples, a typical pattern of linkage disequilibrium observed when loci are within 10 kilobases apart in Arabidopsis (Brachi et al. 2010). This was quite surprising since these variants were located distally or even on different chromosomes which would indicate against a theory of linkage disequilibrium. Our first thought was that these cases were simply due to chance. With 1000 variants and a sample of only 65 plants, it is likely that some variants will exhibit a linked pattern.

Alternatively, we also considered the possibility that this behaviour emerged due to epistasis. An epistatic interaction could describe a fatal allele combination between the genotypes of different variants, which would mean that we can only observe certain combinations and not others, in the population. Regardless of whether these genotypes were due to chance or epistasis, the linkage-like phenomena add further complexity to the eQTL interaction and makes it difficult to identify which variants are truly causal to bolting.

2.6 Variants represent nonsynonymous changes in exons

To validate causality, whether an eQTL pair represented an actual interaction, we decided to examine specific variants and check if they represented nonsynonymous changes at their specific loci. In the example of variant: snp_1208 (Fig 5), we found that this variant was located in the gene body of JmjC DOMAIN-CONTAINING PROTEIN 27 (JMJ27) on chromosome 4, position 429435 and described an SNP from G (ref) to A (alt) [see 'Methods and data']. Then, we referred to the Arabidopsis genome annotation (TAIR10) and confirmed that this SNP was a nonsynonymous change in exon 6 from ATA (methionine/Met) to ATG (isoleucine/Ile), though no instance of our SNP was recorded in the database. JMJ27 encodes a transcription factor shown to regulate drought response (Wang et al. 2021), pathogen defence and most importantly flowering time, with mutants being shown to have an earlier flowering time (Dutta et al. 2017). We hypothesise that this missense change Met to Ile within the polypeptide would result in an altered tertiary protein structure, which prevents the transcription factor DNA-binding domain from merging with the TF binding site. Applying the same method, we also manually confirmed nonsynonymous changes for snp_42 and snp_1841. The latter was located in the gene body of AT5G4380, a.k.a. BIR1 that encodes a kinase. Loss-of-function of BIR1 has been found to activate defense responses and cell death (Yanan Liu et al. 2016), both of which are involved in development and bolting as will be detailed in the next section.



id	source	term_id	term_name	term_size	intersection_size		p_value
1	GO:BP	GO:0001666	response to hypoxia	266	12	🟨	2.4e−02
2	GO:BP	GO:0036293	response to decreased oxygen levels	270	12	🟨	2.8e−02
3	GO:BP	GO:0070482	response to oxygen levels	271	12	🟨	2.9e−02
4	GO:BP	GO:0008033	tRNA processing	115	5	🟨	3.0e−02
5	GO:BP	GO:0071456	cellular response to hypoxia	239	11	🟨	4.3e−02
6	GO:BP	GO:0036294	cellular response to decreased oxygen levels	241	11	🟨	4.6e−02
7	GO:BP	GO:0071453	cellular response to oxygen levels	241	11	🟨	4.6e−02
8	GO:CC	GO:0071944	cell periphery	2902	59	🟩	2.9e−04
9	GO:CC	GO:0005886	plasma membrane	2506	50	🟨	3.8e−03
10	GO:CC	GO:0005911	cell−cell junction	900	18	🟨	1.3e−02
11	GO:CC	GO:0009506	plasmodesma	900	18	🟨	1.3e−02
12	GO:CC	GO:0070161	anchoring junction	900	18	🟨	1.3e−02
13	GO:CC	GO:0055044	symplast	901	18	🟨	1.3e−02
14	GO:CC	GO:0030054	cell junction	906	18	🟨	1.4e−02
15	GO:MF	GO:0036094	small molecule binding	6709	55	🟨	3.6e−03
16	GO:MF	GO:0043167	ion binding	6591	54	🟨	4.7e−03
17	GO:MF	GO:0003972	RNA ligase (ATP) activity	1	1	🟨	2.5e−02
18	GO:MF	GO:0008452	RNA ligase activity	1	1	🟨	2.5e−02
19	GO:MF	GO:0033612	receptor serine/threonine kinase binding	75	2	🟨	3.0e−02

g:Profiler (biit.cs.ut.ee/gprofiler)

Figure 10: *Table of ORA results via gprofiler2.*

2.7 Altered stress-response genes cause asynchrony in bolting

Since we could not automate checking for nonsynonymous change, we could not validate every significant variant. We decided to use another approach, called over-representation analysis (ORA), to show which functional information across different sources (GO terms) are overrepresented in the list of genes which are variants are located in. Analysis revealed that our variant's genes are linked to several hypoxia response processes (GO:0001666; GO:0036293; GO:0070482), which act in the cell periphery (GO:0071944; GO:0005886; GO:0005911) and are likely function by binding to other molecules such as kinases (GO:0036094; GO:0043167; GO:0033612)(Fig. 10).

Though these terms are not the major players more directly involved in bolting, cell death (GO:0012501) for example, comparison with results from Redmond et al. 2023 similarly included nearly all of our significantly overrepresented terms. The only terms not part of their results were some of our molecular function terms (GO:0003972; GO:0008452; GO:0033612) linked to RNA ligase/kinase activity which suggest the way in which these bolting-related variants can interfere with certain pathways. Additionally, it is well known that growth and development coordinate closely with stress responses with a consensus that plants place effort on inhibiting their development when responding to stresses (Zhang, Zhao, and Zhu 2020).

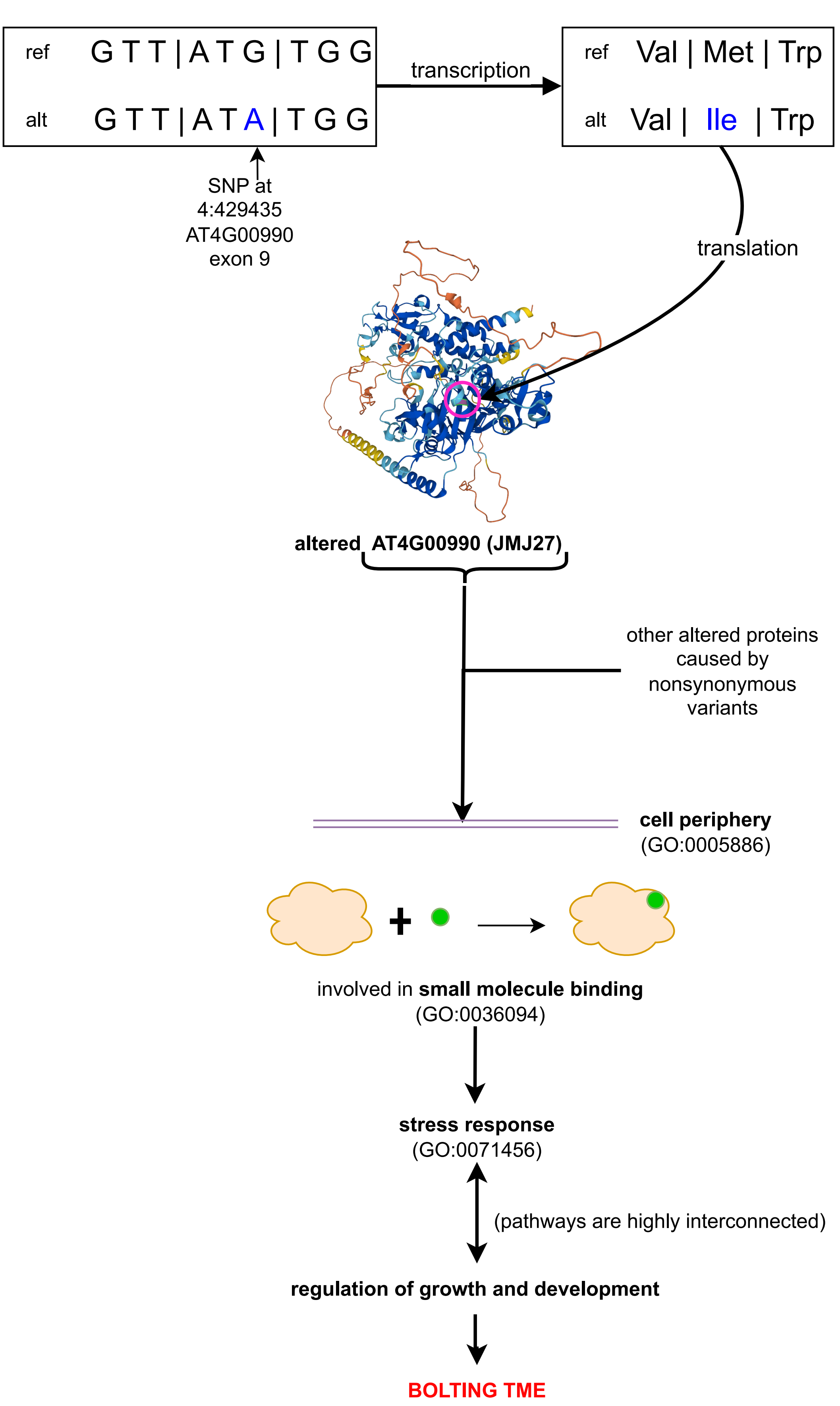

ref G T T | A T G | T G G
alt G T T | A T A | T G G
SNP at
4:429435
AT4G00990
exon 9
transcription
ref Val | Met | Trp
alt Val | Ile | Trp
translation
altered AT4G00990 (JMJ27)
other altered proteins caused by nonsynonymous variants
cell periphery
(GO:0005886)
involved in small molecule binding
(GO:0036094)
stress response
(GO:0071456)
(pathways are highly interconnected)
regulation of growth and development
BOLTING TME

Figure 11: *Diagram giving an example biological mechanism in which variants influence bolting time. JMJ27 was visualized via the AlphaFold protein structure database (Varadi et al. 2024) with the purple circle pinpointing the location of the changed amino acid*

3 Discussion

Through our comprehensive eQTL analysis, we have achieved a better understanding of the complex relationship between genetic variants and bolting in Arabidopsis. By first identifying variant loci correlated to the expression of bolting regulators, we made efforts to characterize the biological mechanisms in which causal variants act. After obtaining the gene that the variants were within, we manually checked for nonsynonymous changes at these loci which confirmed that variant loci caused an altered gene product. We then performed ORA on these genes which further confirmed the function of these genes in the same bolting-related biological processes found in Redmond et al. 2023. Specifically, we found overrepresented terms relating to hypoxic stress response in the cell periphery in which our variants directly alter small molecule binding.

Combining our findings, variants cause altered gene products involved with the stress response. Literature has often described how stress response tightly coordinates with growth and development such that responding to stress inhibits development. This suggests that our variants are in control of more fundamental developmental rates and as a side effect, the timing of bolting changes. We also hypothesise that our enrichment of the cell periphery term could indicate that variants interfere with hormone signalling pathways that occur intercellularly.

Additionally, our findings directly lead to future research and application. In our cluster analysis, we segmented the variants into groups that either accelerate or delay bolting. This information is crucial for agricultural strategies aimed at optimizing planting and harvesting schedules, potentially enhancing crop yields and quality. Finally, we showed how our variants exhibited interesting linkage-like behaviour, which we concluded may have been produced from epistasis.

While we succeeded in our initial objectives in this report, completing the analysis revealed both the strengths and challenges of using advanced genomic tools to dissect plant developmental processes. As such, our results led to many new hypotheses but also realized many limitations. For example, a big limitation was that our data only contained protein-coding variants. For this reason, we basically ignored cis-acting variants from our study, since they often occur in regulatory regions outside of gene bodies, which means we can only identify a subset of variable loci which impact bolting.

We also made the assumption that all eQTL interactions are fundamentally due to nonsynonymous changes, however, studies have shown that synonymous mutations have a chance to cause effect in systems. For example, a synonymous mutation could create an RNA polymerase binding site which would affect transcription (Bailey, Alonso Morales, and Kassen 2021).

However, our main limitations were due to amount of time we had in this project, which prevented us from completing additional analysis. Batch correction could have been performed in order to minimize the downstream effects that produced so many correlations. QTL could have been additionally performed to verify existing or discover other significant variants. We could have also included the effect of covariates in our analyses to improve accuracy. Finally, we could have compared our data with other accessions available on 1001 genomes (Weigel and Mott 2009) and eQTL results on AraQTL (Nijveen et al. 2017) to show if loci are consistent across the literature..

We propose that these limitations can act as platforms for further bolting study, which we would advise using the single-plant-omics approach. In our project, we demonstrated the potential of single-plant-omics to investigate developmental asynchrony through a single highly inbred population, which generates genotype data with few variants. Through single-plant-omics, performing an eQTL analysis is easy, and results in discoveries with higher statistical power.

4 Methods and data

4.1 Data sources and collection information

Data for the analysis was taken from Redmond et al. 2023 via. 'Supplementary Materials'. This included the raw genotype data for filtered variants (Table S11), raw expression data for filtered genes (Table S12), and the edge list for the bolting gene regulatory network which we used to obtain the set of bolting regulators (Table S10).

Using the paper as a reference, 75 plants from the Wassilewskija Ws-2 ecotype were grown in uniform conditions. At day 21, the 3rd and 4th leaves were harvested and RNA-sequenced leading to our raw genotype data. After quality control and discarding of outliers, they obtained the n=65 samples used throughout their dataset.

4.2 Genotype data preparation

The raw genotype data file contained the haplotype of each sample, however, was a matrix of character elements. We processed the data such that "0—1" and "1—0" mapped to 1 (int), "1—1" mapped to 2 (int) and "0—0" mapped to 0 (int). The raw genotype data also contained information about variant type, i.e. SNP/indel, and position, which was used later to map to genes and check for nonsynonymous change. Note that we label the variants in our data as 'snp_xxx' for code simplicity.

4.3 eQTL analysis and filtering criteria

eQTL analysis was performed using the MatrixEQTL R package (Shabalin 2012). Cis/trans classification was ignored based on the assumption that all eQTLs were trans-acting. A p-value of 0.01 was set as a minimum for eQTL discovery. Covariates were not included.

To further select for significant bolting-related eQTL, gene targets were filtered to include only the 44 bolting regulating transcription factors (Redmond et al. 2023), before applying a 1e-5 FDR-adjusted p-value and +-1 log2 fold-change threshold.

4.4 TAIR data and tools

The Arabidopsis Information Resource (Berardini et al. 2015) was used repeatedly throughout the report for a variety of reasons, mainly to search for gene functions.

When deriving the gene in which a variant was located in, we used a gene annotation (TAIR10_GFF3_genes.gff) of a previous Arabidopsis assembly (TAIR10). This file contained the start and end position of genes, along with their TAIR ID. We also used the Gene Description Search and Download tool to obtain gene descriptions and aliases of specific gene sets. When checking for nonsynonymous change, we used a genome browser (JBrowse 2) on the latest Arabidopsis genome assembly (Araport11) and SNPs were mapped to the cDNA sequence. This revealed whether SNPs were in 1st/2nd/3rd place at their codon, and we referred to a codon to amino acid table to determine whether there was a change in amino acid.

Note that as of 6th May 2024, The Arabidopsis Information Resource has updated its website and some tools that were used for this project have been reformatted or removed.

4.5 Clustering by the expression effect size

When performing clustering we used the ComplexHeatmap R package (Gu, Eils, and Schlesner 2016), which performed hierarchical clustering based on the log2 fold-change in expression for each variant-gene (eQTL) regression. After constructing the initial heatmap, we subjectively decided to split the heatmap into 3 clusters of variants and genes using k-means clustering via. the kmeans() R function with 3 centres, though repeated iterations were required to obtain the desired clusters.

4.6 Over-representation analysis

ORA was performed using the gprofiler2 R package (Kolberg et al. 2023). For every bolting-related variant, we obtained the gene that they were located within and used this list of genes as the query for the gost() function. We set a p-value threshold of 0.05.

References

Nica, Alexandra C. and Emmanouil T. Dermitzakis (June 19, 2013). “Expression quantitative trait loci: present and future”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 368.1620, p. 20120362. ISSN: 0962-8436. DOI: 10.1098/rstb.2012.0362.

Michaelson, Jacob J., Salvatore Loguercio, and Andreas Beyer (July 1, 2009). “Detection and interpretation of expression quantitative trait loci (eQTL)”. In: *Methods*. Global approaches to study gene regulation 48.3, pp. 265–276. ISSN: 1046-2023. DOI: 10.1016/j.ymeth.2009.03.004.

Bryois, Julien et al. (Oct. 14, 2021). *Cell-type specific cis-eQTLs in eight brain cell-types identifies novel risk genes for human brain disorders*. DOI: 10.1101/2021.10.09.21264604.

Liu, Yan et al. (Oct. 9, 2020). “Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits”. In: *Genetics, Selection, Evolution : GSE* 52, p. 59. ISSN: 0999-193X. DOI: 10.1186/s12711-020-00579-x.

Slatko, Barton E., Andrew F. Gardner, and Frederick M. Ausubel (Apr. 2018). “Overview of Next Generation Sequencing Technologies”. In: *Current protocols in molecular biology* 122.1, e59. ISSN: 1934-3639. DOI: 10.1002/cpmb.59.

GTEx Consortium (Sept. 11, 2020). “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science (New York, N.Y.)* 369.6509, pp. 1318–1330. ISSN: 1095-9203. DOI: 10.1126/science.aaz1776.

Chen, Yanqing et al. (Mar. 2008). “Variations in DNA elucidate molecular networks that cause disease”. In: *Nature* 452.7186, pp. 429–435. ISSN: 1476-4687. DOI: 10.1038/nature06757.

Galpaz, Navot et al. (2018). “Deciphering genetic factors that determine melon fruit-quality traits using RNA-Seq-based high-resolution QTL and eQTL mapping”. In: *The Plant Journal* 94.1, pp. 169–191. ISSN: 1365-313X. DOI: 10.1111/tpj.13838.

Keurentjes, Joost J. B. et al. (Jan. 30, 2007). “Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci”. In: *Proceedings of the National Academy of Sciences* 104.5, pp. 1708–1713. DOI: 10.1073/pnas.0610429104.

Sonah, Humira et al. (2015). “Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean”. In: *Plant Biotechnology Journal* 13.2, pp. 211–221. ISSN: 1467-7652. DOI: 10.1111/pbi.12249.

Redmond, Ethan J. et al. (Sept. 12, 2023). *Single-plant-omics reveals the cascade of transcriptional changes during the vegetative-to-reproductive transition*. DOI: 10.1101/2023.09.11.557157.

Cortijo, Sandra et al. (Jan. 24, 2019). “Widespread inter-individual gene expression variability in Arabidopsis thaliana”. In: *Molecular Systems Biology* 15.1, e8591. ISSN: 1744-4292. DOI: 10.15252/msb.20188591.

Cruz, Daniel Felipe et al. (Dec. 21, 2020). “Using single-plant-omics in the field to link maize genes to functions and phenotypes”. In: *Molecular Systems Biology* 16.12, e9667. ISSN: 1744-4292. DOI: 10.15252/msb.20209667.

Chen, Chen et al. (June 1, 2019). “Bolting, an Important Process in Plant Development, Two Types in Plants”. In: *Journal of Plant Biology* 62.3, pp. 161–169. ISSN: 1867-0725. DOI: 10.1007/s12374-018-0408-9.

Klingenberg, Christian Peter (Mar. 7, 2019). “Phenotypic Plasticity, Developmental Instability, and Robustness: The Concepts and How They Are Connected”. In: *Frontiers in Ecology and Evolution* 7. ISSN: 2296-701X. DOI: 10.3389/fevo.2019.00056.

Zohner, Constantin M, Lidong Mo, and Susanne S Renner (Nov. 12, 2018). “Global warming reduces leaf-out and flowering synchrony among individuals”. In: *eLife* 7. Ed. by Bernhard Schmid and Ian T Baldwin, e40214. ISSN: 2050-084X. DOI: 10.7554/eLife.40214.

Azodi, Christina B. et al. (Jan. 2020). “Transcriptome-Based Prediction of Complex Traits in Maize”. In: *The Plant Cell* 32.1, pp. 139–151. ISSN: 1532-298X. DOI: 10.1105/tpc.19.00332.

Chien, Pei-Shan et al. (Sept. 29, 2023). “Transcriptome-wide association study coupled with eQTL analysis reveals the genetic connection between gene expression and flowering time in Arabidopsis”. In: *Journal of Experimental Botany* 74.18, pp. 5653–5666. ISSN: 0022-0957. DOI: 10.1093/jxb/erad262.

Li, Zhonghua et al. (2020). “Combined GWAS and eQTL analysis uncovers a genetic regulatory network orchestrating the initiation of secondary cell wall development in cotton”. In: *New Phytologist* 226.6, pp. 1738–1752. ISSN: 1469-8137. DOI: 10.1111/nph.16468.

Kendal, Wayne S. and Brian P. Suomela (June 2, 2005). “Large-scale genomic correlations in Arabidopsis thaliana relate to chromosomal structure”. In: *BMC Genomics* 6.1, p. 82. ISSN: 1471-2164. DOI: 10.1186/1471-2164-6-82.

Brachi, Benjamin et al. (May 6, 2010). “Linkage and Association Mapping of Arabidopsis thaliana Flowering Time in Nature”. In: *PLoS Genetics* 6.5, e1000940. ISSN: 1553-7390. DOI: 10.1371/journal.pgen.1000940.

Wang, Qiongli et al. (2021). “JMJ27-mediated histone H3K9 demethylation positively regulates drought-stress responses in Arabidopsis”. In: *New Phytologist* 232.1, pp. 221–236. ISSN: 1469-8137. DOI: 10.1111/nph.17593.

Dutta, Aditya et al. (2017). “JMJ27, an Arabidopsis H3K9 histone demethylase, modulates defense against Pseudomonas syringae and flowering time”. In: *The Plant Journal* 91.6, pp. 1015–1028. ISSN: 1365-313X. DOI: 10.1111/tpj.13623.

Liu, Yanan et al. (2016). “Loss-of-function of Arabidopsis receptor-like kinase BIR1 activates cell death and defense responses mediated by BAK1 and SOBIR1”. In: *New Phytologist* 212.3, pp. 637–645. ISSN: 1469-8137. DOI: 10.1111/nph.14072.

Zhang, Heng, Yang Zhao, and Jian-Kang Zhu (Dec. 2020). “Thriving under Stress: How Plants Balance Growth and the Stress Response”. In: *Developmental Cell* 55.5, pp. 529–543. ISSN: 15345807. DOI: 10.1016/j.devcel.2020.10.012.

Varadi, Mihaly et al. (Jan. 5, 2024). “AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences”. In: *Nucleic Acids Research* 52 (D1), pp. D368–D375. ISSN: 0305-1048. DOI: 10.1093/nar/gkad1011.

Bailey, Susan F, Luz Angela Alonso Morales, and Rees Kassen (June 16, 2021). “Effects of Synonymous Mutations beyond Codon Bias: The Evidence for Adaptive Synonymous Substitutions from Microbial Evolution Experiments”. In: *Genome Biology and Evolution* 13.9, evab141. ISSN: 1759-6653. DOI: 10.1093/gbe/evab141.

Weigel, Detlef and Richard Mott (May 27, 2009). “The 1001 Genomes Project for Arabidopsis thaliana”. In: *Genome Biology* 10.5, p. 107. ISSN: 1474-760X. DOI: 10.1186/gb-2009-10-5-107.

Nijveen, Harm et al. (2017). “AraQTL – workbench and archive for systems genetics in Arabidopsis thaliana”. In: *The Plant Journal* 89.6, pp. 1225–1235. ISSN: 1365-313X. DOI: 10.1111/tpj.13457.

Shabalin, Andrey A. (May 15, 2012). “Matrix eQTL: ultra fast eQTL analysis via large matrix operations”. In: *Bioinformatics* 28.10, pp. 1353–1358. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts163.

Berardini, Tanya Z. et al. (Aug. 2015). “The Arabidopsis Information Resource: Making and Mining the ‘Gold Standard’ Annotated Reference Plant Genome”. In: *Genesis (New York, N.Y. : 2000)* 53.8, p. 474. DOI: 10.1002/dvg.22877.

Gu, Zuguang, Roland Eils, and Matthias Schlesner (Sept. 15, 2016). “Complex heatmaps reveal patterns and correlations in multidimensional genomic data”. In: *Bioinformatics* 32.18, pp. 2847–2849. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btw313.

Kolberg, Liis et al. (July 5, 2023). “g:Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)”. In: *Nucleic Acids Research* 51 (W1), W207–W212. ISSN: 0305-1048. DOI: 10.1093/nar/gkad347.