Gene expression variation and buffering mechanisms in *Arabidopsis thaliana*

Jakub Zastąpiło

A thesis submitted for the degree of Master of Science (by Dissertation) in Biological Sciences

School of Life Sciences

University of Essex

October 2021

Abstract

Non-genetic variability in gene expression is an inevitable consequence of the stochastic nature of processes driving transcription and translation, as well as the epigenetic modifications of the genome. This phenomenon has been observed in both unicellular and multicellular organisms. Largely thought to be deleterious to cell fitness, it is not uniform across the transcriptome. This implies the existence of mechanisms regulating expression variability, although they, and the role played by inter-individual expression variability, remain poorly researched in multicellular systems. I utilised multiple *Arabidopsis thaliana* time series expression datasets to identify variable genes and analyse their cellular functions.

I show that variable genes are enriched for Gene Ontology terms related to biotic and abiotic stress response and that, inversely, low variability genes are enriched for housekeeping terms. Moreover, I also investigated DNA methylation as a potential mechanism buffering expression variability by analysing methylation of *Arabidopsis* genes and promoters, and by comparing wild type plants with CG methylation reduction and CG methylation loss methyltransferase mutant specimens. I found that variable genes are less methylated in the CG context in wild type *Arabidopsis*. Loss of CG methylation alters expression variability of some genes. Of those, significantly greater portion of genes gained variability, compared to those that lost it. These results are an important step towards greater understanding of these processes in multicellular organisms, and their role.

Acknowledgements

I would like to thank both my supervisors, Dr. Ulrike Bechtold and Dr. Radu Zabet, for their support and patience. Their guidance and advice was crucial to completing this work, especially in this COVID 19-affected year.

I would also like to thank the University of Essex Plant Group for their thoughtful comments.

Lastly, I wish to thank my family for supporting me.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

Biorep, bioreplicate – biological replicate

bp – base pair

BP – Biological Process

BS – bisulfite sequencing

CATMA - Complete *Arabidopsis* Transcriptome Micro Array

cDNA - complementary DNA

CMT2 – CHROMOMETHYLASE 2

CMT3 – CHROMOMETHYLASE 3

Col-0 – Columbia

CV – coefficient of variation

DAVID - Database for Annotation, Visualization and Integrated Discovery

DMR – differently methylated region

DNA - Deoxyribonucleic acid

dsRNA – double-stranded RNA

FDR – false discovery rate

gbM – gene body methylation

GEO - Gene Expression Omnibus

GFP – green fluorescent protein

GO – Gene Ontology

h - hours

Met1 – METHYLTRANSFERASE-1

mRNA – messenger RNA

NCBI - National Center for Biotechnology Information

ncRNA – non-coding RNA

PANTHER - Protein Analysis Through Evolutionary Relationships

PCA – principal component analysis

RNA - Ribonucleic acid

scRNA-seq – single-cell RNA sequencing

SUVH4 – KRYPTONITE

TAIR - The *Arabidopsis* Information Resource

WT – Wild type

1. Introduction

1.1. Background

By 2050, the human population is expected to reach 9.7 billion (United Nations, 2019). However, close to 750 million people were exposed to severe food insecurity in 2019 - a number that, should we remain on our current course, is expected to grow (FAO et al., 2020). As such, tackling global hunger remains one of humanity's top priorities. It is, however, vital, that it be done in a sustainable manner – already one-third of Earth's land not covered by ice is occupied by either settlements or agricultural areas (Ellis and Ramankutty, 2008). The additional danger posed by worsening climate caused by global climate change (Rahmstorf and Coumou, 2011) suggests that understanding of plant stress tolerance will be vital in addressing these issues. Indeed, an analysis by Boyer (1982) indicates that majority of theoretical yield for plants grown in the United States is lost due to imperfect adaptation to the environment they are grown in.

Essential to improving our understanding of stress tolerance is improving our knowledge of mechanisms involved in plant stress response. For many species, numerous genes involved in stress response have been identified. Almost 5000 *Arabidopsis* proteins have been assigned "response to stress" Gene Ontology (Ashburner *et al.*, 2000) (Gene Ontology Consortium, 2019) term based on experimental evidence alone. This does not mean that our understanding of stress response is complete, even just for *Arabidopsis*. The traits responsible for response to stress, especially abiotic stress, are multigenic (Wang et al., 2003) and, as such, response to stressors involves multiple transcription factors and signalling molecules. These gene-networks are complex – some gene products are conducting

response to only specific stresses, whereas others are involved in multiple signalling pathways (Chinnusamy et al., 2004).

One phenomenon with potential role in management of plant stress response is inter-individual gene expression variation, here defined as variability in gene expression between organisms. The effects of variation in genome upon transcription are well understood, particularly as they contribute to diseases like cancer (Budinska et al., 2013), yet non-genetic expression variation has also been observed, even in models like cancer (Inde and Dixon, 2018).

While non-genetic expression variability appears to be exploited by immune response of complex organisms like animals (Hagai et al., 2018), and plays a role in survival under stress in *Eukarya* like yeast (Bishop et al., 2007), the role of gene expression variability in *Arabidopsis* is largely unknown, with the exception of seed germination times (Johnston and Bassel, 2018) (Abley et al., 2021).

1.2. Gene expression variation

1.2.1. Genome-based variation

Of the three main sources of gene expression variation, the one discovered first, before the advent of genetics itself, is genetic variation, which encompasses differences in expression between two organisms of the same species caused by differences in their genome (Fay et al., 2004). Among the forms it may take are single nucleotide polymorphisms (Ranjith-Kumar et al., 2007), copy-number variations (Stranger et al., 2007) and polyploidy (Wang et al., 2006). Importantly, the inheritance of genome state of germline cells is heritable (Rahbari et al., 2016), meaning variation caused by de-novo mutations is perpetuating across generations. The influence of genome variation on expression is an axiom of genetics, with

extensive genotype-phenotype linkages (Lehner, 2013). For instance, plants of varying genotypes may differ in their stress response genes, leading to different transcriptomes and phenotypes under stress (Carlson et al., 2017). This type of variation may even appear in cell populations descendant from a single cell (Liu et al., 2019).

The sources of genetic variation are as varied as the forms it takes and range from errors in DNA replication (Kunkel, 2000), to flawed repair of DNA damage (Cooke et al., 2003), to transposon insertion (Tsugeki et al., 1996). These alterations to the genome may then go on to alter transcriptome in a number of ways. The most significant of them are changes to structure of the gene itself. An insertion or a deletion may shift the reading frame of the gene in a so-called frameshift mutation (Ripley, 1990), which drastically changes the content of RNA produced downstream of the mutation. Another example is a mutation which introduces a stop codon, which shortens the gene product (Vidal et al., 1999). Mutations not affecting the sequence of the gene itself may still alter its expression, however, by altering the way gene expression machinery interacts with its promoters, altering the frequency of transcription (Donald and Cashmore, 1990).

1.2.2. Non-genetic variation

Genome is not the only source of expression variation. That two cell lines derived from the same genotype may grow to possess distinct phenotypes could be explained by mutation (Liu et al., 2019), yet same phenomenon also appears within isogenic populations (Inde and Dixon, 2018). Within these populations, by definition, the influence of the genetic variation is next to null. Therefore, it must be assumed that an alternative source of gene expression variability exists.

The phenomenon of non-genetic variation has been known since before the DNA was modelled (Green, 1941). For the purposes of this work, non-genetic variation in gene expression is defined as any variability in gene expression patterns between genetically identical cells or organisms that do not result from alterations to their genome. Broadly, sources non-genetic variability can be categorised to belong to one of two categories.

1.2.2.1. Cell state variation

A major source of variation is epigenetics which, in this work, is defined as inheritable modifications of chromatin that do not change the underlying sequence of nucleotides, yet alter gene expression (Bird, 2007). The influence of epigenetics is most clearly defined in determination of cell fate, where it plays a crucial role in development of multicellular organisms (Tollervey and Lunyak, 2012). Differentiation is driven by two factors – noise, which in biological context is defined as random fluctuations in biological activities such as transcription rates, and dedicated biological processes such as hormonal signalling (Serra et al., 2018). Epigenetic changes are not only responsible for maintenance of cell fate, but also other types of cellular memory, such as stress memory (Liu et al., 2014), although they share this role with various proteins (Thirumalaikumar et al., 2020).

Epigenetic regulation can be separated into four mechanisms: DNA modification, chromatin modification, non-coding RNA-based regulation, and RNA modification (Aristizabal et al., 2020). DNA modification takes the form of DNA methylation. In *Arabidopsis thaliana*, cytosine methylation is the most common type of DNA methylation, although adenine methylation also plays a biological role (Liang et al., 2018).

In *Arabidopsis*, Cytosine methylation occurs in CG, CHG, and CHH contexts, where H stands for C, T, or A (Stroud et al., 2014). CG-context methylation is more frequent than CHG and CHH (Lister et al., 2008), and is maintained through different mechanisms (Stroud et al., 2014) (Zabet et al., 2017). The first, and the one most relevant to this work, relies on Methyltransferase-1 protein (Finnegan and Dennis, 1993) (Kankel et al., 2003), which is primarily responsible for the inheritance of CG-context methylation. The second mechanism, responsible for methylation of cytosine in CHG and CHH contexts, relies on action of SUVH4 in conjunction with CMT2 or CMT3 proteins (Du et al., 2014). Lastly, CHH methylation is asymmetric, and cannot be maintained by CMT2 alone. As such, RNA-directed DNA methylation processes play a crucial role in re-establishing it de-novo (Law and Jacobsen, 2010). RNA-directed DNA methylation is not limited to CHH context (Mathieu et al., 2007). However, while this mechanism is capable of immobilising some transposable elements in mutants with deficient Methyltransferase-1 (Marí-Ordóñez et al., 2013), it cannot compensate for the loss of methyltransferase function, as suggested by significant drop in methylation in these mutants (Catoni et al., 2017).

One role played by cytosine methylation in all contexts in *Arabidopsis* is immobilisation of transposable elements, by preventing their transcription (Kato et al., 2003). Importantly, loss of CG methylation maintenance is not lethal in *Arabidopsis*, as it leads to activation of a number of alternative epigenetic expression control mechanisms (Mathieu et al., 2007), although it still results in development of abnormalities in the phenotype ranging from altered rosette shape to lowered fertility, which over several generations may eventually result in sterility (Finnegan et al., 1996).

Chromatin modification covers modification of histones, and inclusion of histone variants. Histones are multi-unit proteins, which play a role in regulation of gene expression by regulating the form of chromatin (Littau et al., 1965). As histones are composed of multiple sub-units, different variants may be expressed by the cell – for instance, in response to stress (Ascenzi and Gantt, 1997). Histones may be modified by methylation, acetylation, phosphorylation or ubiquitination of amino-acid residues that make them up (Zhang et al., 2007), among other means. Histone modification is not exclusive with DNA methylation, and the two may co-occur (Cedar and Bergman, 2009).

The roles played by ncRNAs in gene expression regulation are as varied as ncRNAs themselves. They may silence genes, preventing transcription, or interfere with mRNAs, preventing translation (Matsui et al., 2013). RNA molecules may even generate de-novo DNA methylation (Aufsatz et al., 2002). It is thought that this is the mechanism that allows CG methylation maintenance-impaired *Arabidopsis* mutants to survive (Mathieu et al., 2007). Moreover, the RNA-mediated silencing is also implicated in post-transcriptional gene silencing, which is a mechanism used by plant cells to provide virus resistance (Mourrain et al., 2000).

Much like DNA, mRNA molecules too can be subject to modification, forming the epitranscriptome (Fray and Simpson, 2015). In *Arabidopsis*, methylation of adenosine in mRNA influences gene expression by regulating transcript abundance (Parker et al., 2020) by, among other means, inhibiting cleavage of mRNA molecules (Anderson et al., 2018). This type of mRNA methylation functions in regulation of differentiation as well (Shen et al., 2016), and mutants without the protein necessary for its application are lethal in embryo (Luo et al., 2014).

Epigenetics is not the only factor in determining the cell state. Chemicals secreted by other cells, such as transcription factors (Santuari et al., 2016), or other signalling molecules (Rentel et al., 2004), may alter cellular behaviour. Moreover, genetically identical cells or organisms may, when exposed to different environmental conditions, express different phenotypes, in a phenomenon known as phenotypic plasticity (Schlichting, 1986). Phenotypic plasticity is well-documented in both plants (Schmitt et al., 2003) and animals (Miura, 2005), and it is indeed thought to play a role in some plant response to stress (Campbell et al., 2019). Various environmental stresses may alter gene expression of an organism in a diverse array of pathways, ranging from activation of transcription factors (Gao et al., 2008) or RNA-binding proteins (Marondedze et al., 2019), to interaction with epigenetic markers like histone modification (Yan et al., 2019) and DNA methylation (Chen et al., 2018).

How do these mechanisms contribute to variation? Difference in environment, even when minor, can generate variation (Trontin et al., 2011). But even in absence of environmental differences, a genetically homogeneous population may develop phenotypic heterogeneity. Much like mutation leads to changes in the genome, spontaneous epimutation gives rise to changes in the epigenome (van der Graaf et al., 2015). Non-genetic differences, such as variation in methylation (Shahryary et al., 2020), may lead to differences in phenotype (Denkena et al., 2021).

1.2.2.2. Stochastic variation

Variation in expression can, however, be attributed to one additional source – stochastic nature of gene expression itself (Roberfroid et al., 2016). To understand the source of this effect, one must first examine the processes responsible for gene

expression, the first of which is transcription. In eukaryotes, gene expression is a complicated process which involves regulatory DNA regions both close and distant to the gene, and promoter sequences, and requires formation of a transcription preinitiation complex (Roeder, 1996) by various types of transcription factors and cofactors, RNA polymerase, as well as other molecules (Li et al., 2016).

Because the concentrations of enzymes involved in transcription are low and the process very often takes place on the scale of single molecules, transcription itself is to some degree random, as a result of random behaviour of individual molecules (McAdams and Arkin, 1997). Therefore, transcription occurs infrequently and sporadically, in form of transcriptional bursts (Tunnacliffe and Chubb, 2020). This randomness can be expressed in terms of “noisiness” of gene expression – variability inherent to the system (Raj and van Oudenaarden, 2008). Evidence of bursting has been found in bacteria (Golding et al., 2005), eukaryotes (Quintero-Cadena et al., 2020), and even viruses integrated into human genome (Skupsky et al., 2010).

Transcriptional bursting is not entirely random, however – by changing factors related to genetic environment of the gene, such as local epigenetic state, identity of regulatory elements, and availability of transcription factors, the cell can alter the frequency and magnitude of expression bursts between genes (Nicolas et al., 2017).

Another potential source of variation is the rate at which produced mRNA degrades (Cao and Grima, 2020) – the process of mRNA decay, like mRNA synthesis, is stochastic (Elgart et al., 2010). As with transcriptional bursting, mRNA half-lives are thought to vary depending on their susceptibility to decay pathways (Beelman and Parker, 1995) and potentially epitranscriptomic modifications

(Anderson et al., 2018). While this doesn't impact noisiness of transcription, it does affect the other stochastic process involved in gene expression – translation. While not all gene products are made through translation (Mattick, 2003), those that are it represents another stochastic process on the way to creation of functional product. Lastly, degradation of gene products too is stochastic (Komorowski et al., 2013). All the stochastic processes compound onto each other, which leads to an increase in noise.

Gene expression noise has a deleterious effect on the cell, with increases in noise leading to decreased fitness (Schmiedel et al., 2019). A too large departure from the optimal expression value of a gene can be deleterious to the cell's fitness (Dykhuizen et al., 1987), though tolerance varies between genes (Keren et al., 2016). Not all genes are subjected to noise equally. Cells possess mechanisms in place that allow them to modulate the effects of noisy expression on a gene-by-gene, and network-by-network basis (Barroso et al., 2018). Noise propagates through gene networks, meaning high-noise genes pass on some of their noisiness downstream. Moreover, the ability to suppress noise is both limited (Lestas et al., 2010), and comes with a cost that increases as intended degree of precision does (Voliotis and Bowsher, 2012).

Nonetheless, noise need not lower the cell fitness. In situations where expression level of a gene is not optimized for fitness of the cell, more noise can be beneficial, as it allows a subset of the population to express the optimal amount of the gene (Duveau et al., 2018). An optimal noise range exists for expression of each gene, for which fitness of the organism is the highest. Housekeeping genes, for instance, are enriched within low-noise gene set (Barroso et al., 2018), as variance in their concentrations has deleterious effects on cell variability.

One important aspect of stochastic variation is that it impacts on smaller and simpler organisms to a significantly greater degree than larger and more complex ones. In populations of single cell organisms, noise resulting from stochastic nature of biological processes is significant enough to have measurable impact on the phenotype. Bacteria (Elowitz et al., 2002) and *S. cerevisiae* (Blake et al., 2003) are commonly used models for study of these phenomena. Non-genetic heterogeneity in *E. coli* in particular has been well studied. Choi et al. (2008) found that the expression of the lac operon, the model of an inducible operon, is stochastic in nature when levels of inducing molecule are medium, leading to a heterogeneity in a population. What's more, Kotte et al. (2014) discovered that isogenic *E. coli* cell lines split into distinct phenotypes, and that in the event of the environmental change the subset of the population well suited to the new environment begins replicating, while the subset poorly adapted to new conditions ceases growing, though remains viable.

This proves that, while noise is not optimal for individual organisms, it is beneficial to the population as a whole. Moreover, it has been reported that, when an alternate, inferior respiratory substrate is present, a subset of the population will stochastically differentiate to make use of it, thus allowing it to remain viable should oxygen levels be depleted (Carey et al., 2018). Stochastic switching between states is not limited only to stresses, however – it is also present in altruistic production of colicin by a subset of a population, within which cells stochastically switch between energetically intensive process of production and rest (Bayramoglu et al., 2017).

The above examples represent a bacterial risk mitigation strategy known as "bet hedging". Cells that are ill-suited to the environment are unable to grow, as they enter persistence state (Balaban et al., 2004), yet, should the conditions change, they ensure part of the population survives. While bet hedging is most apparent in

bacterial populations, it is present in more complex organisms as well. Freshwater fish, for instance, faced with harsh conditions, opt to produce a larger number of smaller eggs (Morrongiello et al., 2012). While this negatively impacts chances of survival of any individual offspring, it increases the chances that one of the group survives. The form this strategy takes can vary between kingdoms, but the principle – lowering risk to the population while increasing individual risk – is maintained between bacteria and eukaryotes. In plants this phenomenon is present in the form of varying seed germination times (Johnston and Bassel, 2018).

In more complex, multicellular organisms, the impact of noise of individual transcription events on the organism as a whole is lessened. Stochastic events still play an important role, particularly in early development (Dietrich and Hiiragi, 2007). Gene expression noise is responsible for some of the decision-making in regard to cell speciation and differentiation (Serra et al., 2018), meaning it is fundamental in development of complex organisms. It even plays a role in “life” cycles of viruses, as it determines whether or not latency should be engaged (Weinberger et al., 2005). The effects of stochastic nature of gene expression, however, become less relevant at greater scales – when averaged out across thousands of cells, the impact on the organism is reduced. While processes that drive determination and formation of organs are inherently stochastic, they result in development of morphologically distinctive organs whose shape is largely conserved between individuals of the same species (Hong et al., 2018).

The changes caused by stochastic processes may persist well into plant development, however. Both mutation (Uphoff et al., 2016) and epimutation (Johannes and Schmitz, 2019) are stochastic processes, and both genome and epigenome (Hofmeister et al., 2017) are heritable, whereas gene expression noise

itself is not. Seed germination times and seed dormancy are both traits that vary within populations (Simons and Johnston, 2006), affected by the genome and the epigenome (Han et al., 2019).

1.2.3. Measuring gene expression variation

Two main categories of methods exist for measurement of gene expression, which are measurement of mRNA contents of the cell, and measurement of gene product concentration. The first method relies on existing DNA sequencing technologies – mRNA is extracted, exploiting poly-A tails (Wang et al., 2007), and converted to cDNA using reverse transcriptase (Nagalakshmi et al., 2010). As a result, this method allows for study of expression of all protein-coding genes in the cell, with only limitation on genes studied being that of DNA sequencing used. There exist two main avenues for analysis of transcriptome. The first avenue requires sequencing of a larger population of cells, which are typically flash-frozen to capture the expression profile (Rayirath et al., 2009). This method, while unable to capture differences between cells in the sequenced population, has seen much use in analysis of expression profiles under stress (Matsui et al., 2008), or as a part of development-focused time series (Klepikova et al., 2015). The alternative is single cell transcriptomics. Unlike the previous avenue, this allows for measurement of stochastic noise within cell populations (Alemany et al., 2018).

These techniques do, however, suffer from a number of caveats. Both multi-cell and single-cell approach, depending on starting amount of RNA, require amplification, in order to provide enough material for sequencing tools. Unfortunately, amplification (Parekh et al., 2016), reverse transcription (Conn and Conn, 2019), and fragmentation (Roberts et al., 2011) may all introduce biases into

the data – either as a consequence of the enzymes used, or because of the stochastic nature of the processes involved. Because single cell RNAseq involves less starting material, the impact of the amplification bias is much greater, thus lowering reliability of resulting data (Parekh et al., 2016). A number of methods exist that can help quantify, or reduce, the amount of uncertainty (Sun et al., 2019), such as use of unique identifiers. New sequencing methods are currently in development. Long-read sequencing platforms, such as Nanopore and SMRT sequencing (Cui et al., 2020), too can potentially be used for scRNA-seq (Lebrigand et al., 2020). Moreover, both of these techniques only consider mRNA, which means that even on scale of a single cell noise and variability associated with translation is lost.

The second category of methods instead relies on measuring concentration of gene product. In case of proteins, this can be accomplished by a variety of assays should the cell be lysed, such as colorimetric assay (Sapan et al., 1999). One method that is particularly advantageous relies on tagging relevant proteins with GFP (Tsien, 1998), and then subjecting the cells to flow cytometry, as demonstrated by Blake et al. (2003). Fluorescence-based analysis creates a system sensitive enough to pick up gene expression noise (Elowitz et al., 2002), which does not require cells to be lysed, allowing for continuous monitoring, and which can be scaled up much more robustly than scRNA-seq. This does, however, come with caveats. To employ the tagging method, genome must be modified to tag relevant proteins with GFP, and the number of potential tagged proteins is limited.

For both methods, analysis of high variability genes within expression dataset can be carried out in a number of ways. One approach, used for analysis of cell-to-cell variability in expression, relies on calculation of distance to median – a metric where coefficient of variation for each gene is compared to running median of

coefficient of variation values (Newman et al., 2006). This method can be used to analyse both protein (Newman et al., 2006) and transcript abundance (Hagai et al., 2018). Another solution instead relies on squared CV, but otherwise follows the steps above (Kolodziejczyk et al., 2015).

Measuring underlying causes of variation, meanwhile, heavily depends on the exact phenomenon analysed. Differences in the genome can be identified by DNA sequencing (Li et al., 2008). Differences between states of analysed organisms too can be analysed, yet such analysis is more difficult than that of nucleotide sequence, owing to the large number of epigenetic markers, which often must be examined individually. DNA methylation of cytosines can be quantified down to single nucleotide resolution by use of bisulfite sequencing (Smith et al., 2009). Analysis of adenosine methylation of DNA, however, requires wide assortment of methods ranging from single-molecule real-time sequencing to immunoprecipitation of fragmented reads (Liang et al., 2018). Histone modifications, or histone variants, can be analysed using methods based on chromatin immunoprecipitation as well (Kimura, 2013), provided a specific antibody is identified. ncRNA sequencing is possible, though identifying individual types of ncRNA requires a wide array of computational tools (Veneziano et al., 2015). Analysis of mRNA modification, meanwhile, cannot currently be performed on the scale of the entire transcriptome, and relies on immunoprecipitation (Fray and Simpson, 2015).

Gene expression noise is difficult to analyse. It is typical for studies to analyse a single GFP-tagged protein at a time (Elowitz et al., 2002), which permits measurement of changes in the protein’s concentration. While the effects of gene expression noise may persist well past single-cell scale, and can thus be studied

using the methods above, the contribution of noise to variability within multi-cellular organisms decreases with the amount of cells.

1.3. Plant stress response

Plants are sessile organisms, meaning they are unable to change habitats on their own. This has forced them to develop various ways of coping with alterations to the environment. These are typically separated into biotic stress, which is caused by other organisms, and abiotic stress, which is caused by inorganic environmental changes. Examples of originators of biotic stress include infections by pathogens, such as viruses, fungi, bacteria, and oomycetes, which can be separated into necrotrophs and biotrophs (Oliver and Ipcho, 2004), but also damage caused by insects, and presence of other, competing plants (Widdicombe and Thelen, 2002). Abiotic stresses, meanwhile, cover a wide range of changes such as increases or decreases in temperature, presence, abundance or absence of specific chemicals in the soil, overabundance or scarcity of light, and/or physical damage to the plant (Kilian et al., 2007) (Yan et al., 2019).

Two specimens of the same species subjected to the same stress can vary in their response. The most obvious cause here is genetic variation, as differences found inside loci associated with response to a given stress naturally result in difference in response (Vallejo et al., 2010). It is also vital to note the role of acclimation - plants, including *A. thaliana*, possess a degree of phenotypic plasticity that allows them to, after an exposure to certain stresses, alter their metabolism in ways that allow them to alleviate the effects of that stress (Hannah et al., 2005). Age of the specimen too is a factor that influences not just stress resistance response, but also acclimation process itself (Leuendorf et al., 2020).

There exist a number of methods to ascertain whether or not an organism has initiated a stress response pathway. The one most relevant to this study is analysis of its transcriptome – much of the response to stress is directed by transcription factors (Gao et al., 2008), and the alterations of the expression profile of cells caused by their activity can be quantified (Golem and Culver, 2003). Therefore, it stands to reason that by analysis of upregulated and downregulated groups of genes the activity of stress response pathways can be analysed. The same applies to acclimation, although only to some degree – it has been discovered that effects of acclimation, such as increased stress resistance, may persist longer than expression of the genes responsible for causing these changes (Leuendorf et al., 2020), albeit the extent of the effect is not fully known.

1.4 Conclusion, and rationale

Gene expression variability leads to variability in the phenotype. Moreover, as discussed before, the phenomenon is partially a result of the scale at which biological processes take place, rendering it unavoidable. Phenotypic variability may, however, have its benefits. In animals, gene expression variability is thought to play a role in immune response (Hagai et al., 2018). Within plants, variability is present in the form of seed germination times (Abley et al., 2021), where it is a part of a bet-hedging strategy. Within yeast, it was found that higher heterogeneity was linked to improved survival of individual specimens under stress (Bishop et al., 2007). What role it plays in *Arabidopsis thaliana*, beyond seed germination, is largely unknown. Existing research indicates that it has a connection to stress response (Cortijo et al., 2019) – but, importantly, our understanding of mechanisms controlling it is limited.

What are the possible causes for variation in gene expression? Naturally, first and foremost, variation in the genome itself (Li et al., 2017), as changes in structure of regulatory proteins alters expression. Research indicates, however, that even isogenic populations exhibit variation in gene expression (Cortijo et al., 2019). Within these populations, there exist two main mechanisms of note to explain variation in gene expression – cell state-based variation, and variation arising from stochasticity of gene expression itself.

Different influences on gene expression can be reduced by experimental design. Genetic variation, naturally, can be greatly reduced by measuring expression variability within an isogenic population. While mutations within the organism continue to occur, spontaneous mutation rate is low enough (Ossowski et al., 2010) to all but ensure genetic homogeneity. This work relies on expression data from ground whole leaves, resulting in resolution at which stochastic differences between individual cells become subsumed into the mean expression value of the whole organ. That is not to say that stochastic processes are not potentially responsible for variation observed between samples, however – epimutation is a stochastic process (Johannes and Schmitz, 2019).

What remains is cell state variation. Of this category, the last source of variation that has been accounted for is large-scale environmental variation. While microvariations, which too can impact expression (Trontin et al., 2011), would be difficult to identify and remove, more significant environmental impacts can be removed by both experimental design and principal component analysis, to discard plant samples that depart too significantly from the majority.

Curiously, previous research indicates that chromatin environment of highly variable genes in *Arabidopsis* is not supportive of expression (Cortijo et al., 2019), which may support the hypothesis it is epigenetic changes inherited from earlier development that are responsible for higher expression in individual specimens.

The main hypotheses behind this work are that inter-individual variation in gene expression does indeed play a biological role in biological processes, including stress response, and that gene methylation is one of the mechanisms used to buffer gene expression variability. As such, it is predicted that loss of methylation should result in an increase in gene expression variability.

1.5 Objectives

In order to conduct a thorough analysis of inter-individual gene expression variability, two expression time series of *Arabidopsis thaliana* Col-0 ecotype specimens grown in identical conditions were selected – long term (Bechtold et al., 2016), and short term (Alvarez-Fernandez et al., 2021). The aim of this analysis is to identify genes with high expression variability within the two expression series, and to determine the biological processes enriched within this gene-set through Gene Ontology. Additionally, a similar analysis is intended to be carried out with regards to low expression variability gene-set.

The second objective is to analyse gene expression variability in the context of gene and promoter methylation to identify if an association exists between the two. This analysis will use bisulfite sequencing series of *Arabidopsis thaliana* Col-0 ecotype (Stroud et al., 2012) (Stroud et al., 2013), and compare expression to methylation in three contexts. An additional Gene Ontology analysis of genes categorised by expression and methylation shall be carried out.

The third objective is to further explore the findings of previous analyses in regards to linkage between gene expression variability and methylation. Separate paired gene expression and methylation series of WT *Arabidopsis* and hypomethylated methyltransferase mutants (Catoni et al., 2017) shall be used to identify if changes in methylation of genes and promoters are associated with changes in gene expression variability.

2. Methods

All scripts for the analysis are available on Github, at github.com/ZastaJak/athaGE_variation_and_buffering.

2.1 Input and normalisation of microarray data

Publicly available microarray mock expression data was imported into R (R Core Team, 2020). Two series were utilized. The first was mock drought series (Bechtold et al., 2016), accessible at GSE65046 as control and zero measurements, composed of 4 bio-replicates at each time point, with measurements performed every day for 14 days, resulting in 56 measurements. The second was mock data from high light series (Alvarez-Fernandez et al., 2021), accessible at GSE78251 as control and zero measurements, composed of 4 bio-replicates at each time point, with measurements performed every 30 minutes for 6 hours, resulting in 52 measurements. In both series tissue from *Arabidopsis thaliana* leaf 7 of a separate plant for each bio-replicate was harvested, and sequenced using CATMA microarrays (Sclep et al., 2007). The plants in both series were 5 weeks old at the start of measurement series.

The two control time series differed in CATMA microarray version and their set of probes. A consensus probe-set was constructed, featuring only probes present in both series. The joined datasets were normalized together, using cyclic loess normalization method, as proposed by Bolstad et al. (Bolstad et al., 2003), carried out using *normalizeBetweenArrays* function of the *limma* (Ritchie et al., 2015) package.

2.2 Microarray probe processing and gene assignment

Each CATMA probe was referenced to a pre-generated assignment table, in order to match them with genes, in the form of TAIR IDs. Probes that did not match with any known genes, or that matched to multiple genes, were discarded**.**

To assign chromosomal location to each probe-gene pair, gene information was extracted from a TxDb (Lawrence et al., 2013) object, supplied by the *TxDb.Athaliana.BioMart.plantsmart22* package (Carlson, 2015), generating a Genomic Ranges (Lawrence et al., 2013) object containing chromosomal location and strand of all *Arabidopsis* genes. Probes-gene pairs were matched to this object and assigned chromosomal locations**.** Genes absent in TxDb object were compared to TAIR changelog, to identify replacements. Probes matched to that could be replaced were assigned new IDs and chromosomal locations, whereas those that could not were discarded.

Genes that were assigned multiple probes were coerced into single entry. For each bio-replicate of such a gene, a mean value was calculated from the expression values of assigned probes.

A Genomics Ranges object was formed from the single-probe and corrected multi-probe data. The gene information was used as the main body of the object, with strand, chromosome, name and location, while expression values were assigned to metadata columns.

2.3 Data exploration - Principal Component Analysis

Principal Component Analysis was performed separately for mock drought and mock high light samples. PCA was carried out using *prcomp* function of the *stats* (R Core Team, 2020) package. PC1 with PC2 were plotted and depicted as a scatterplot generated using *ggplot2* (Wickham, 2016) package, using custom colour palette

(Martin Krzywinski, 2020), to allow for visual inspection of the data. Bio-replicates that differed significantly from others within their time point were removed.

2.4 Computation of mean and variability measures

A number of variables was calculated for each gene between bio-replicates within each measurement. Variance, standard deviation and mean were computed, using *var*, *sd* and *mean* functions – first two of *stats* package, the second of *base*. These were used to derive coefficient of variation (standard deviation divided by mean), squared coefficient of variation, and Fano factor (variance divided by mean). Lastly, Distance to Median was computed using adopted version of the *Distance-to-median* function from the *scran* (Lun et al., 2016) package. All these variables were also computed between all bio-replicates for both drought and high light mock data sets, disregarding time-points, resulting in a single value for drought and high light mock datasets respectively, henceforth referred to as **drought mock** (**DM**) and **high light mock** (**HLM**). Four potential measures of gene expression variability were selected for further analysis, based on their capacity to measure variability between samples of differing magnitude.

2.5 Selection of gene expression variability measures, and cut-off value

Coefficient of variation (CV), squared coefficient of variation (CV^2), Fano factor and Distance to Median were compared in order to select the most suitable metric of variation and a cut-off value by which to filter variable genes. The four were selected because each is corrected by mean value directly (CV, CV^2, Fano factor) or indirectly (Distance to Median), which ensures that highly expressed genes are not assigned higher variability. Histograms of the distribution of these variables were generated for DM and HLM using the *ggplot2* package, which were annotated with p-values

generated by use of the *wilcox.test* function of the *stats* package that compared values of DM and HLM. A threshold value was assigned to each measure of variability, determined in each of the four individually by comparing their HLM and DM distributions, to ensure a sufficient amount of genes would overlap between the distributions to allow for further analysis, yet no greater than 5% of the analysed geneset in size.

2.6 Analysis of distribution of coefficient of variation across mock data

Histograms comparing coefficient of variation values between filtered and unfiltered genes were generated following the previous method. Afterwards, heatmaps of the genes passing the threshold value in one or both of the series were made, using *ggplot2* package. The columns, corresponding to measurement time points, as well as DM and HLM, were clustered together based on similarity between their values. The clustering was generated using *hclust* function of the *stats* package, using “complete linkage” method to identify similar time points. This information was also used to generate a dendrogram, which has been appended to the heatmap. Similar heatmaps were generated for other variables, in addition to the ones that featured CV of all genes, including those not passing the threshold value. An additional plot was generated using *ggplot2* package depicting the relationship between the mean value of gene expression and coefficient of variability in DM and HLM.

2.7 Gene Ontology analysis of high variability geneset

In order to identify Gene Ontology Biological Process terms enriched within high-variability dataset, three methods were selected. The first two make use of external tools – Panther Overrepresentation Test (Released 20210224) (Thomas et al., 2003) (Mi et al., 2010), and DAVID version 6.8 (Huang et al., 2009a) (Huang et al., 2009b).

The third is carried out by *go_enrich* function of the GOfuncR (Grote, 2020) package, using the Wilcoxon rank-sum test method, which ranks the genes based on the submitted values – in this implementation, coefficient of variation – and analyses for enrichment the top and bottom-ranked genes. GOfuncR does not by itself contain GO term annotation for *Arabidopsis* genes, so it was supplied separately by TAIR-generated gene association (Berardini et al., 2004).

Panther and DAVID were supplied with three gene lists: the genes passing the CV threshold value in DM, the genes passing CV threshold in HLM, and those that passed the CV threshold in both DM and HLM, termed the “consensus list”. For Panther, the output was specified to feature False Discovery Rate for correction, and Fisher’s exact test type.

Panther and DAVID analyses of the consensus geneset were conducted twice, using two backgrounds – first, the “all *Arabidopsis* genes” background, built into both tools, and, second, the custom background, generated from all genes that were made into the Genomic Ranges object. All other Panther and DAVID analyses made use of the custom background, and no background was required for Wilcoxon analysis.

The results of enrichment analysis were compared using p-value and q-value metrics. *Go_enrich* results used a different method of p-value correction, and were corrected to match the output of Panther and DAVID manually, using *p.adjust* function of the *stats* package. P-value and q-value thresholds of 0.05 were used to signify relevant terms, in accordance with convention.

2.8 Generation of visual representation of Gene Ontology data

The results were plotted in two ways. First, they were used to determine what terms were enriched in high coefficient of variation subset of genes. Panther and DAVID

results for the analysis of the consensus genelist using custom background were compared to results of Wilcoxon rank-sum tests. To identify enriched GO terms, relevant ($q < 0.05$) results of these four tests were plotted, using *ggplot2*. A second plot was generated, made of terms that featured in output of both Panther and DAVID, and Wilcoxon rank-sum tests analysis of either DM or HLM. GO IDs were translated to term names using *GO.db* package (Carlson, 2021). GO terms absent from GO.db annotation were removed.

The results were used to establish difference between HLM and DM datasets. The output was filtered as before, to select relevant terms, and then manually inspected to select differing terms relevant to differences between the two.

2.9 Gene Ontology analysis of low variability geneset

A histogram of distribution of coefficient of variation in DM and HLM was generated, as described before. A second threshold value was assigned to signify genes with low expression variability, determined by the analysis of DM and HLM distributions, not to exceed 10% of the analysed geneset in size.

As in analysis of high variability geneset, three lists were used – genes below the threshold in HLM, those below the threshold in DM, and those below the threshold in both.

An analysis was performed using a variation of the previous protocol – the only background used was the custom background. Results of the same Wilcoxon analysis as conducted earlier were used, reversed. As before, first, a plot of all terms with $q < 0.05$ was generated using *ggplot2*, but instead depicting terms that appeared in output of at least 3 GO methods.

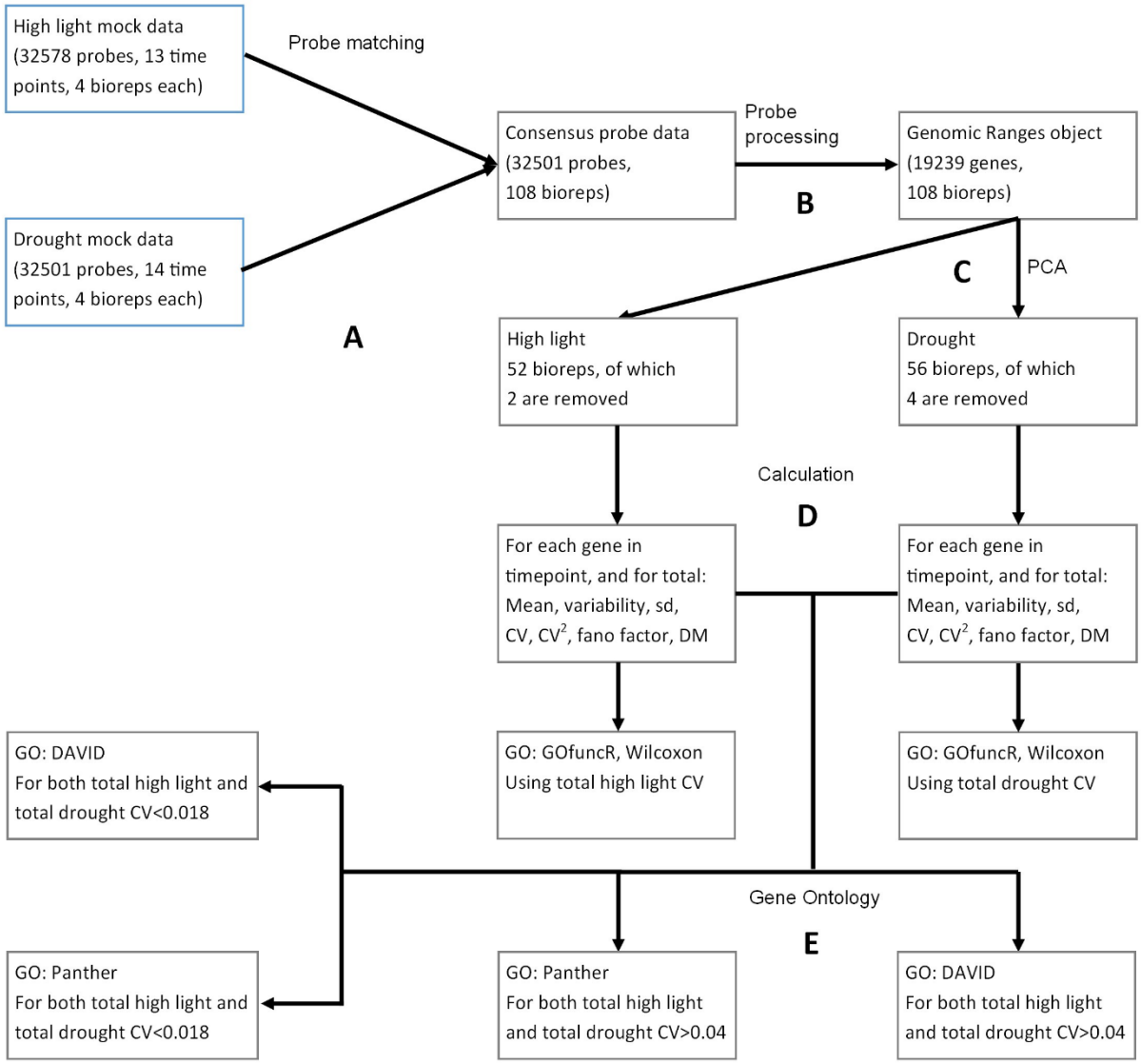

Figure 2.1: *Workflow cartoon representing steps taken in analysis of the two mock microarray datasets.*

A: Matching data by probes. Due to differences between CATMA microarray versions, a difference existed between mock high light and drought probesets. In this step, probes not present in all experiments are removed.

B: Probe processing. This step represents a variety of processes necessary to create a Genomic ranges object. These include normalisation, matching probes to genes, matching probe-gene pairs to genetic locations, calculating mean values for genes matched to multiple probes, and forming the GRanges object itself.

C: Principal component analysis. In this step, high light and drought mock datasets were separately subjected to PCA, in order to remove outlying biological replicates.

D: Calculations. Bioreps were grouped by time point. For each gene, variables were calculated within bioreps: mean, variability, standard deviation, coefficient of variation, Fano factor, and distance to mean. Moreover, same variables were calculated for all genes within each analysed mock dataset, generating DM and HLM.

E: Gene ontology. GOfuncR, using Wilcoxon rank-sum test, used HLM and DM lists separately, whereas Panther and DAVID were supplied genes with CV higher than 0.04 in both HLM and DM. Gene Ontology analysis was also performed on low-noise dataset, using threshold value of 0.018 for CV in HLM and DM.

2.10 WT bisulfite sequencing data pre-processing.

Publicly available bisulfite-converted sequencing data for 3-week-old leaves harvested from three wild type Columbia ecotype *Arabidopsis thaliana* bio-replicates was downloaded from NCBI Sequence Read Archive database. Bio-replicate 1, SRR501624 was generated as part of a separate study (Stroud et al., 2012) from bio-replicate 2, SRR534177, and bio-replicate 3, SRR534193 (Stroud et al., 2013). The data was processed following protocol outlined by (Catoni et al., 2018). The imported reads were trimmed, using Trimmomatic (Bolger et al., 2014) version 0.39. The tool was supplied with merged file containing all standard Trimmomatic adapters, as well as two specific Illumina ones. The trimming was performed in paired end mode, with suggested default values, aside from headcrop of 6, and conversion of quality scores to phred33.

2.11 Bisulfite alignment and methylation extraction

Bismark (Krueger and Andrews, 2011) script was used to perform alignment and methylation call, using trimmed reads. The reads were aligned to *Arabidopsis thaliana* genome sequence TAIR 9 release (Lamesch et al., 2012). Default settings were used, aside from number of mismatches set to 1 for greater sensitivity, and Bowtie 2 (Langmead and Salzberg, 2012) score_min parameter set to L,0,-0.6, to allow for less stringent alignment. The output was processed using *deduplicate_bismark* script of Bismark in paired end mode, in order to deduplicate the file and remove unnecessary reads. Methylation was extracted using *bismark_methylation_extractor* script.

2.12 Per-gene methylation proportion assignment

The generated CX_report files were imported into R separately using the readBismark function of the *DMRcaller* package (Catoni et al., 2018). The data was corrected by chloroplast methylation, based on the assumption that the chloroplast chromosome should not contain any methylation, as described by (Catoni and Zabet, 2021). Corrected data was analysed with *analyseReadsInsideRegionsForCondition* DMRcaller function using two separate Genomic Ranges objects. The first object contained locations of genes analysed previously (see 2.4), and the second was composed of coordinates of promoters for these genes, here defined between 1000 bp upstream and 50 bp downstream of gene start. The analysis was conducted separately for each of the three methylation contexts.

2.13 Low-resolution analysis

Resulting 18 objects were loaded into a single R session, alongside CX reports for each bio-replicate. Methylation profile of chromosome 1 was calculated using CX reports for each methylation context by *computeMethylationProfile DMRcaller* function with default parameters aside from window size of 500000 bp. Similarly, spatial correlation of methylation levels was calculated for each methylation context using *computeMethylationDataSpatialCorrelation* function, with distances of 1, 10, 100, 1000 and 10000. Lastly, coverage of methylation was calculated using *computeMethylationDataCoverage* function for CG context only, with minimum numbers of reads of 0, 5, 10, 15, 20, and 25. The profiles were plotted using *ggplot2* package and merged together using the *patchwork* (Pedersen, 2020) package.

2.14 Methylation proportion mean calculation

Methylation proportions of each gene body and promoter were extracted, and combined. Genes that were not annotated for methylation proportion were excluded from further analysis. For each gene, mean value of methylation proportions between the three bio-replicates was calculated for each context, for both the gene body and the promoter region. Additionally, gene expression coefficient of variation and mean expression values for high light mock (HLM) and drought mock (DM) calculated previously (see 2.4) were imported and attached to methylation proportion data.

2.15 Mean expression analysis, and gene splitting

HLM and DM mean expression data was analysed to determine threshold values for further analysis. A histogram was plotted, using *ggplot2*. Based on the distribution of expression values, 8.4 and 12.05 were selected as cut-off values for, respectively, low expression and high expression genes. Both values correspond to approximately top or bottom 10% expressed genes in both DM and HLM. Data for gene methylation was split into two categories based on their methylation – gene-body methylation (gbM), and transposable-like methylation. The latter was composed of genes with CHG or CHH context methylation proportion higher than 0.05, while the former contained all other genes. Promoter data was not split.

2.16 Analysis of gene expression magnitude and gene and promoter methylation

To analyse relationship between expression magnitude and methylation, all analysed genes were categorised based on their expression and methylation. For CG methylation, the threshold between methylated and un-methylated genes was set to 0.1 methylation proportion for the whole gene body or promoter. For CHG and CHH

methylation, the threshold was set to 0.05, to accommodate for lesser frequency of methylation in these contexts. For each context and for each form of methylation (gbM, transposable element-like, and promoter), genes were split into methylated and unmethylated, and then into lowly-expressed, medium-expressed, and highly-expressed. Relationship between gene expression magnitude and methylation proportion was plotted using *ggplot2*, with separate expression values for HLM and DM.

To analyse statistical significance of the relationship between gene expression magnitude and methylation of genes and promoters, methylation values of genes split into lowly expressed, medium expressed and highly expressed categories were plotted on a boxplot. Only genes which fit into the same category in both HLM and DM were analysed. p-values between the groups were calculated using the Wilcoxon rank-sum test method, using *wilcox.test* function of the *stats* (R Core Team, 2021) package.

2.17 Analysis of gene expression variability and gene and promoter methylation

In order to determine connection between gene methylation and gene expression variability, analysed genes were split into a separate set of categories. The thresholds used for methylation proportion were the same as for the previous analysis, while coefficient of variation threshold was set to 0.04, as used in the previously (see 2.7). The data was plotted separately for HLM and DM for each form of methylation and for each context using *ggplot2*. As before, to determine significance of the relationship the methylation proportion values of genes which matched variability category between HLM and DM were plotted on a boxplot, with p-values calculated using the Wilcoxon rank-sum test method.

Gene Ontology Biological Process analysis of the four categories was conducted using PANTHER (Released 20210224) (Thomas et al., 2003) (Mi et al., 2010). Only genes with gbM methylation were analysed, and only CG methylation proportions were used. Genes that matched category between DM and HLM were split into 4 lists. Background list was created from all genes with assigned methylation values.

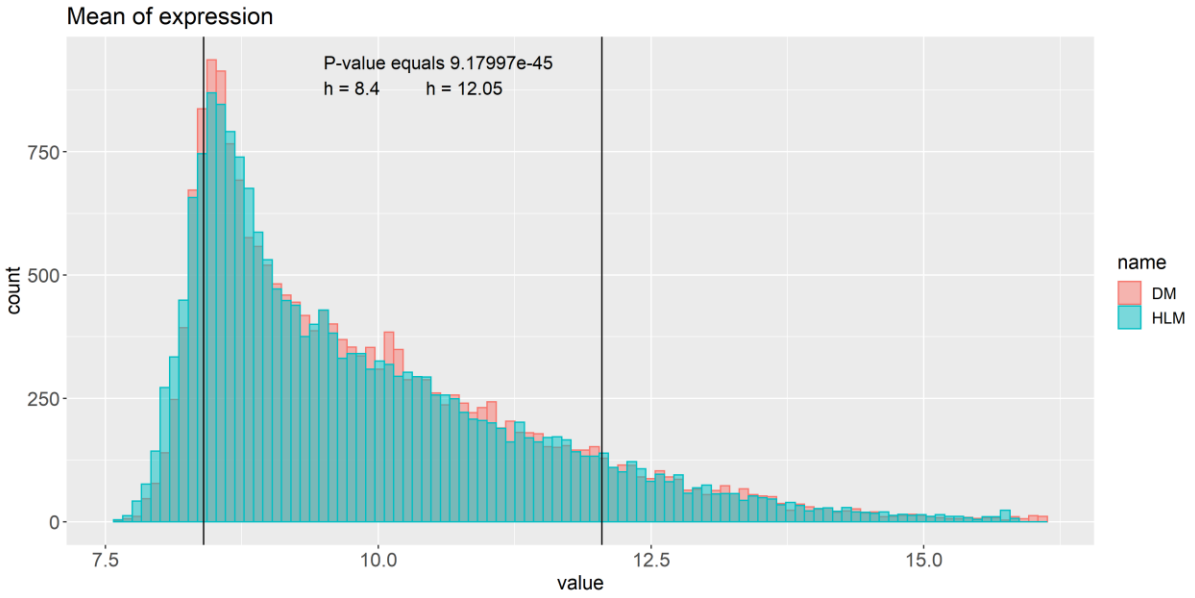

Figure 2.2: *Gene expression mean comparison.*

The histogram depicts HLM and DM distribution of mean expression values, with x axis representing gene expression mean. Vertical lines represent selected cut-off values, at 8.4 and 12.05, which correspond to approximately the lower 10% of genes sorted by expression value and the upper 10% of genes sorted by expression value. p-value was calculated using Wilcoxon rank sum test.

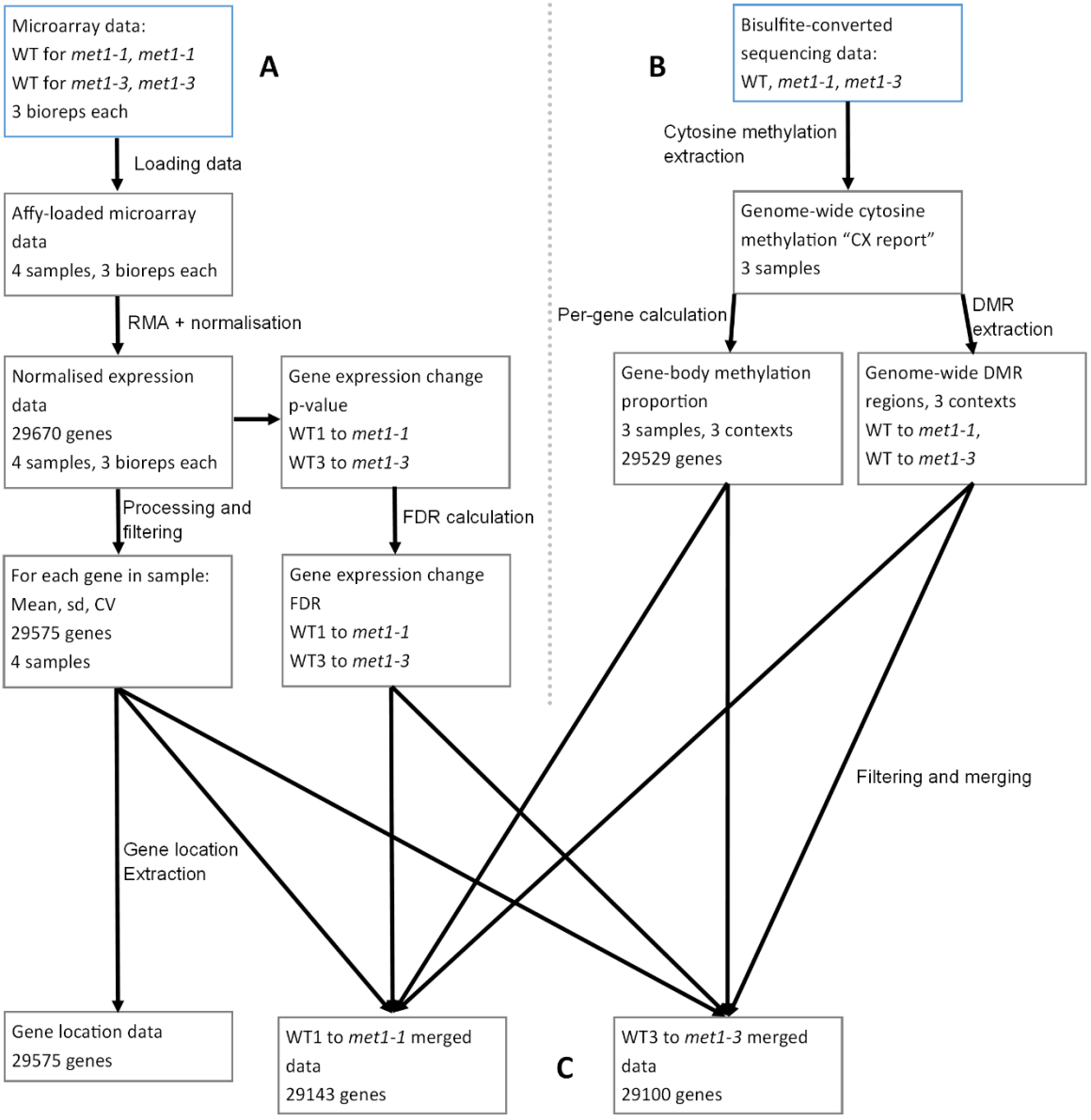

Figure 2.3: *Workflow cartoon representing steps taken in analysis of mutant microarray and bisulfite sequencing data.*

A: Analysis of microarray expression data (see 2.18-2.19, 2.24). The four expression series were loaded into R and transformed and normalised together. Genes which were obsolete, or which were duplicates of other genes, were removed. Per-gene expression data was used to calculate mean, sd and CV between bio-replicates for each series, and also to determine whether a gene was differentially expressed between WT and respective *met1* mutant, represented by p-values, which were FDR-corrected. Additionally, genomic coordinates of all analysed genes were extracted.

B: Analysis of bisulfite-sequencing data (see 2.20-2.23). BS-seq data for WT, *met1-1*, and *met1-3,* was used to calculate per-cytosine methylation proportion for the entire genome. Cytosine data was used to calculate proportion for entire gene bodies, using previously extracted gene location data, and to identify differently methylated regions genome-wide.

C: Dataset filtering and merger (see 2.24). Expression data and methylation data were merged into WT1 to *met1-1* dataset and WT3 to *met1-3* dataset. The two gene-sets were filtered to eliminate genes with low sequencing coverage and those overlapping methylation gain DMRs. The merged datasets contain mean expression value, CV of expression, expression fold change, p-value and q-value measuring differential expression, overlaps with DMRs for both gene bodies and promoters and methylation proportion data for gene bodies and promoters in both WT and *met1* mutant for all genes.

2.18 WT and mutant gene expression data pre-processing and statistical analysis

Raw gene expression data for *met1-1* mutant, wild type for *met1-1* mutant, *met1-3* mutant and wild type for *met1-3* mutant, were used, composed of three bio-replicates each, generated by (Catoni et al. 2017). The data was entered into R using *ReadAffy* function of the *affy* (Gautier et al., 2004) package, using *athtiling1.0rcdf* probe annotation (Naouar et al., 2009). The data was further processed using Robust Multi-Array Average expression algorithm, without normalisation component, which was carried out separately, using *normalize.ExpressionSet* function of the *affyPLM* (Bolstad, 2004) (Bolstad et al., 2005) (Brettschneider et al., 2007) package, with loess method and antilog transformation. Expression values for all bio-replicates were extracted into a separate object. A design matrix was constructed with *stats* package and used to fit a linear model for expression data using *lmFit* function of *limma* (Law et al., 2014) (Ritchie et al., 2015) package. A contrasts matrix was constructed with *makeContrasts* function and used to compute contrasts for linear model fit between wild type for *met1-1* (WT-1) and *met1-1* and wild type for *met1-3* (WT-3) and *met1-3* with *contrasts.fit*. The linear model fit was processed to compute Bayesian statistics for the comparison using *eBayes* function. p-values were FDR corrected using *p.adjust* function of *stats* package.

2.19 Obsolete gene removal, and gene location extraction

The list of analysed genes was extracted and compared to TAIR 10 changelog to identify obsolete and replaced genes. Genes that were obsolete without replacements were removed, and genes that were obsolete with replacements were updated with their new TAIR IDs. Entries that were found to be duplicates of other

genes were removed. Chromosomal coordinates of analysed genes were extracted from *TxDb.Athaliana.BioMart.plantsmart22* (Carlson, 2015) package, and were made into a genomic ranges object, together with coordinates of their promoters, defined between 1000 bp upstream and 50 bp downstream of gene start. Genes absent from the annotation file were discarded.

2.20 WT and met1 mutant Bisulfite sequencing data pre-processing

Bisulfite-converted sequencing data for 2-week old leaves harvested from wild type Columbia ecotype *Arabidopsis*, *met1-1*, and *met1-3* from (Catoni et al. 2017) was selected for analysis. The SRR files were processed as before, using newly written gene and promoter location lists, generating three CX report files, and 18 objects containing methylation proportion data. CX reports were corrected by chloroplast methylation and plotted for methylation coverage, spatial correlation, and low-resolution profiles, as described before.

2.21 Differently methylated region identification

An analysis was carried out to identify differently methylated regions (DMRs). Methylation data was corrected by chloroplast methylation, as described before, and processed using *computeDMRs* function of the *DMRcaller* package. The parameters used for analysis of CG context DMRs were the default, aside from method, where “bins” was selected due to its suitability to all three methylation contexts and statistical test, which was set to “score”. Moreover, for analysis of CHG and CHH context DMRs, minimum proportion difference was lowered to 0.1, in accordance with lower magnitude of methylation in those contexts. For all three, the p-value threshold was set to 0.01, to increase confidence in detected DMRs. The DMRs were calculated between WT and *met1-1* mutants, WT and *met1-3* mutants, and

between *met1-1* and *met1-3*, in three contexts separately, for the entire *Arabidopsis* genome.

2.22 Gene DMR assignment

In order to assign DMRs to genes, *findOverlaps* function of the *IRanges* (Lawrence et al., 2013) package was used to locate overlaps between DMR ranges object and gene locations, with minimum overlap size set to 50. Genes were split into three groups – those overlapping only loss DMRs, those overlapping only gain DMRs, and those overlapping loss and gain DMRs both. The overlaps were identified for all three contexts, and between WT and *met1-1*, WT and *met1-3*, and *met1-1* and *met1-3*. This step was repeated for promoter locations.

2.23 Categorising genes by methylation, and filtering genes with few reads

Gene methylation data was split into three categories, based on their methylation in WT – genes with transposable element-like methylation, here defined as CHG or CHH context methylation proportion higher than 0.05, genes with gene-body methylation, defined as CG context methylation proportion above 0.1 and CHG and CHH below 0.05, and non-methylated genes, that did not fit into one of the previous two groups. Genes without methylation proportion value in one or more contexts were excised. For each gene, WT methylation and *met1-1* methylation proportion data were compared. If the number of total reads for a cytosine for either measurement was less than 25, both measurements were discarded. This was done for each bio-replicate separately, and for both gene-bodies and promoters. This step was repeated for WT and *met1-3*, generating a separate list.

2.24 Expression value calculation, and unification of expression and methylation data

Mean, standard deviation, and coefficient of variation of expression were calculated between bio-replicates for each gene separately for each of the 4 expression series. Gene expression and gene methylation data were combined. Non-CG context methylation measurements were discarded. In CG context, genes overlapping with “gain” or “both” DMRs were discarded.

2.25 WT and mutant gene expression comparison, and differently expressed gene filtering

Mean WT expression values were plotted against methylation proportion, using *ggplot* package, with points categorised based on their DMR overlap. Additionally, WT-1 expression was plotted against *met1-1*, and WT-3 against *met1-3*, with genes categorised based on their expression change and overlap with DMRs. Genes with significant expression change, here defined as FDR higher than 0.05, and log2 of expression fold change higher than 1 or lower than -1, were excluded from further analysis, in order to avoid bias caused by expression difference.

2.26 Classification of genes by coefficient of variation, and comparison between WT and mutants, and mutants and mutants

Log2 of coefficient of variation of expression fold change between WT-1 and *met1-1* and WT-3 and *met1-3* of remaining genes was calculated. Genes were assigned into three categories – “decreased CV”, for genes with log2 of fold change lesser than -1, ”increased CV”, for genes with log2 of fold change greater than 1, and “no change”, for all other genes. Coefficient of variation values in WT against mutants were plotted for genes that overlapped CG methylation loss DMRs. A barplot was drawn, to show

relative sizes of the three groups for each category for these genes, as well as a venn diagram showing overlap between genes found in the same group in WT-1 to *met1-1* and WT-3 to *met1-3*. Significance between sizes of overlaps between groups within gene categories were calculated using Fisher's Exact Test, which plotted as a matrix using *ggplot2*.

A gene ontology analysis of genes with “increased CV” overlapping between *met1-1* and *met1-3* for gbM genes was conducted using PANTHER with FDR correction, with background list generated with all analysed genes, including non-methylated and differentially expressed ones, but without genes with unknown methylation value, less than 25 reads in BS-seq for WT and *met1-1*, or overlapping with DMRs other than methylation loss. As before, p-value and q-value threshold of 0.05 was used, following convention.

3. Results

3.1 Identification and analysis of variable genes.

This study was carried out using two publicly available microarray time series (see 2.1), one short-term and one long-term, which allowed for analysis of a large number of bio-replicates, with the goal of identifying variable genes. As a consequence, however, a series of steps was necessary to address both the discrepancies between the datasets, such as different microarray versions and lack of normalisation, and ensure only genes for which expression values are accurate were analysed.

As a result of consensus probeset creation, 77 probes present in mock high light dataset were removed, leaving 32501 probes for normalization. In the following steps, 11138 probes that could not be matched to a gene, matched to multiple genes, or were assigned to obsolete IDs, were discarded. After computation of mean values for genes matched to multiple probes, this resulted in the list of 19239 genes that were analysed (figure 3.1)

Of 52 analysed HLM bio-replicates, 2 were removed, and 4 were removed from the 56 analysed DM bio-replicates, as a result of Principal Component Analysis, based on their divergence from remaining bio-replicates (figure 3.2A-B). In HLM, bioreplicates B of 1.5 h measurement and C of 3.5 h measurement clearly diverge from the overall distribution of values. In DM, the bio-replicates are not distributed in a distinguishable pattern, and as such bio-replicates were selected for removal based on their divergence.

Coefficient of variation was selected as the metric of gene expression variability, due to its ease of implementation in the analysis of other datasets and a distribution that

allows for simple implementation of the cut-off. Based on the distribution of values across the two series (figure 3.2C), 0.04 was chosen as the cut-off value, in order to categorise enough genes as variable to allow for further analysis.

To ensure coefficient of variability was the suitable metric that was not affected by mean value to a significant degree, CV and mean values of genes in HLM and DM were compared (figure 3.2D-E). In both samples, CV of the majority of genes was not strongly affected by the mean value of expression, although those with very low or very high expression show some differences, however. Genes with the lowest mean values seem to have higher CV than the distribution of other genes would suggest, although not high enough to pass the 0.04 threshold. Inversely, genes with the highest expression values have much lower CV than other highly-expressed genes.

3.1.1 Expression variability patterns change over time both within the day, as well as on a developmental timescale.

In order to estimate how the coefficient of variation changes over time, the coefficient of variation was plotted for mock drought measurements with DM, mock high light measurements with HLM, and both together (figure 3.3). The similarity-based clustering of time points, represented by columns, shows that terms that were close temporally tended to be clustered as well. This is seen within both mock high light (figure 3.3B) dataset, and the mock drought dataset (figure 3.3A). For mock high light, the furthest amount of “steps” separating two time points clustered immediately together was 2, with each step representing a measurement step, equal to 0.5 h for mock high light. In mock drought, this amount was significantly higher, equal to 4 (representing 4 days between the two measurements clustered together). When the

two datasets were joined (figure 3.3C), this difference became more apparent, with mock drought 0 hour, 2 hour and 4 hour time points clustered together with mock high light. This shows that, as plants develop, even within the same environment, differences in expression grow between them. Indeed, it is also evident when looking at the results of principal component analysis (figure 3.2A-B). Within mock high light dataset, covering 6 hours total, only two bio-replicates were found to significantly differ from the others (figure 3.2A). In mock drought dataset, covering 13 days, there were four (figure 3.2B) – twice as many as in mock high light, despite having only one more time point. Additionally, three of these four belonged to measurements taken late into plant development. Importantly, while these were removed as invalid, on the plot representing all CV values (figure 3.3C), all mock drought timepoints past the 8-day mark, aside from day 11, are clustered away from the remaining timepoints. The effect persists in genes whose DM or HLM values passed the 0.04 CV threshold (figure 3.4, figure 3.5A-B).

3.1.2 Distribution of coefficient of variation differs between the samples.

Visual inspection of heatmap comparing variability between mock high light and mock drought samples (figure 3.3) reveals that mock drought dataset displayed higher variability within time points, represented by brighter colours. A statistically significant difference was observed between the two samples (figure 3.5A), with mean of mock drought CV values higher than that of high light.

The trend remained present after filtering for genes whose CV passes the 0.04 threshold in DM or HLM (figure 3.5.B), and its intensity, measured by difference in means, increased significantly. The difference between means was lesser, through

still statistically significant, in genes that pass the CV threshold in both DM and HLM (figure 3.5C), and less apparent on visual inspection (figure 3.4C).

A possible explanation for the difference between DM and HLM lies in the nature of the two datasets. Because HLM is calculated between mock bio-replicates measured at different times across 6 hours, an additional source of variation is present, in the form of the circadian rhythm. As such, significant amount of genes that appear variable in HLM are not variable in DM (figure 3.6), which was calculated over a much longer period of time, and is not influenced by the circadian rhythm. DM features a source of variation absent from HLM as well, in the form of developmental expression differences, yet it is lesser in effect, as evidenced by smaller fraction of DM genes that are not variable in HLM. Additionally, genes variable as a result of the circadian rhythm would not be variable between bio-replicates at each individual time-point. Therefore, the difference between CV distribution of genes which pass the threshold in either of the samples, and the CV distribution of genes which pass the threshold in both DM and HLM, is likely explained by genes variable in HLM only due to the circadian rhythm, as opposed to DM or individual high light mock series time-point measurements.

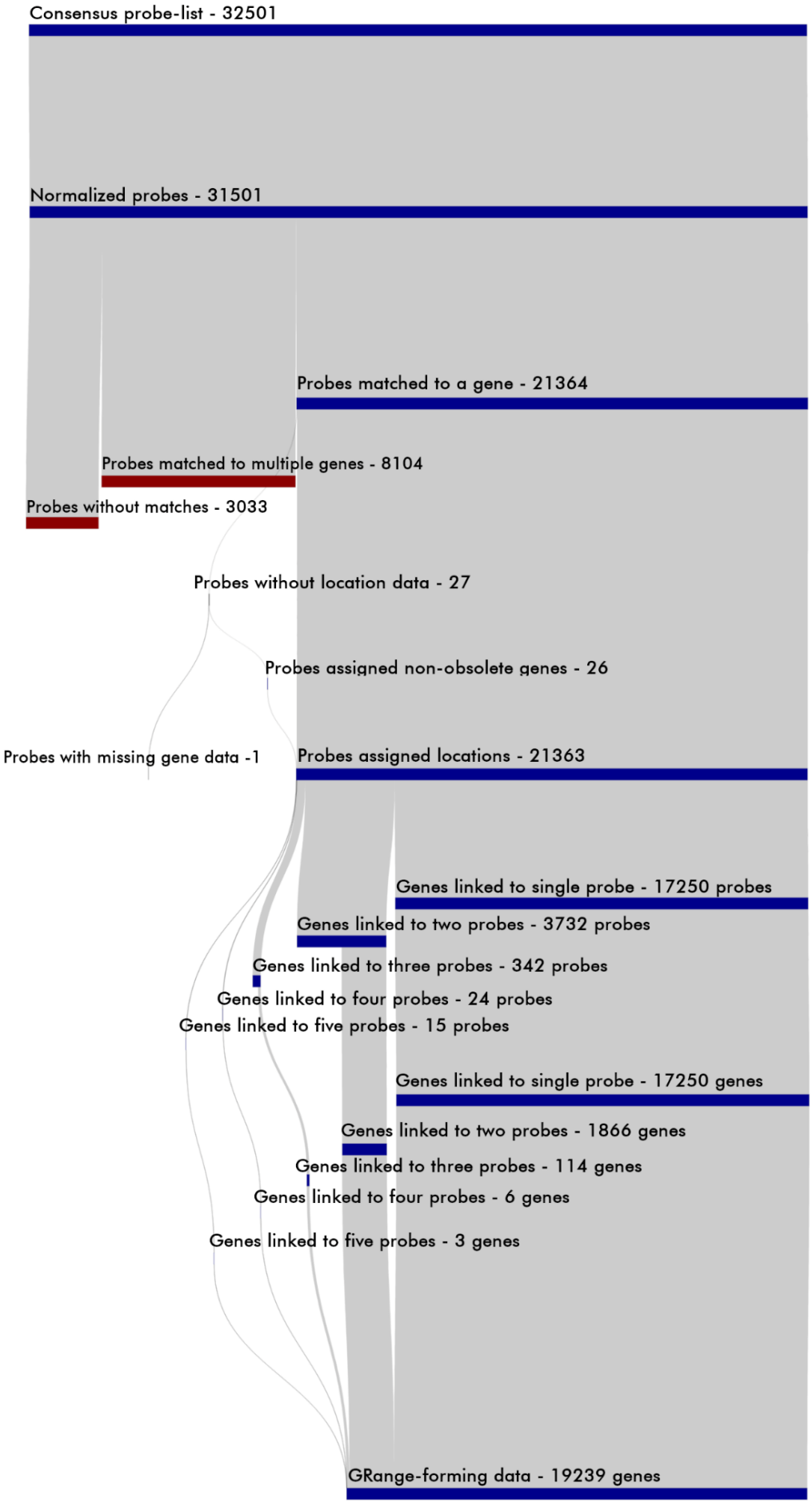
Consensus probe-list - 32501
Normalized probes - 31501
Probes matched to a gene - 21364
Probes matched to multiple genes - 8104
Probes without matches - 3033
Probes without location data - 27
Probes assigned non-obsolete genes - 26
Probes with missing gene data -1
Probes assigned locations - 21363
Genes linked to single probe - 17250 probes
Genes linked to two probes - 3732 probes
Genes linked to three probes - 342 probes
Genes linked to four probes - 24 probes
Genes linked to five probes - 15 probes
Genes linked to single probe - 17250 genes
Genes linked to two probes - 1866 genes
Genes linked to three probes - 114 genes
Genes linked to four probes - 6 genes
Genes linked to five probes - 3 genes
GRange-forming data - 19239 genes

Figure 3.1: *Microarray data processing.*

Sankey diagram representing the steps taken to process the merged high light mock and drought mock microarray data. Bars in red represent rejected probes, bars in blue represent probes or genes passed on to further analysis. Each “flow” represents a quantity of genes or probes passed on to the next node, the size of which is represented by the text above it.

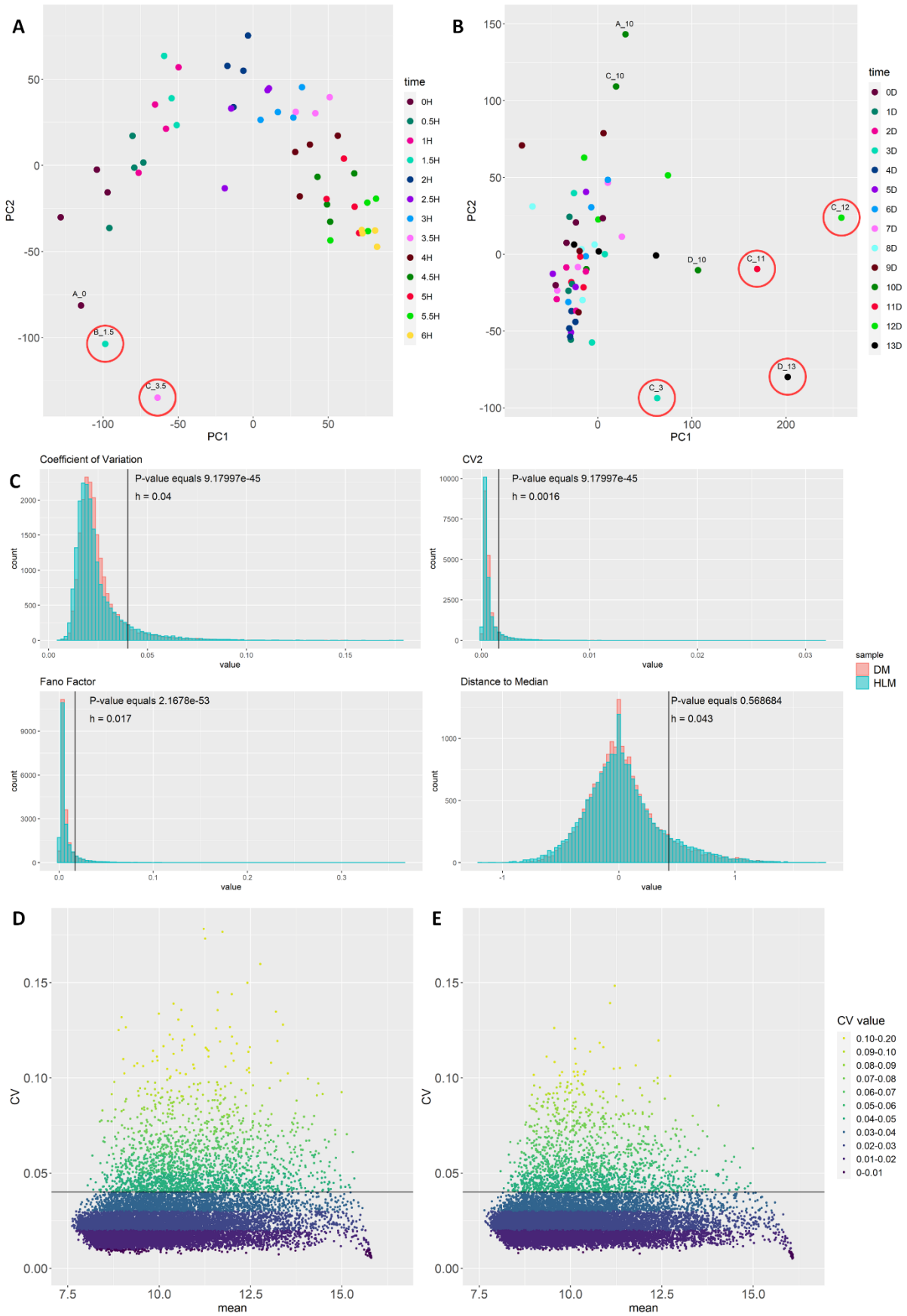
A
PC1
PC2
time
0H
0.5H
1H
1.5H
2H
2.5H
3H
3.5H
4H
4.5H
5H
5.5H
6H
A_0
B_1.5
C_3.5
B
A_10
C_10
C_12
C_11
D_10
D_13
C_3
0D
1D
2D
3D
4D
5D
6D
7D
8D
9D
10D
11D
12D
13D
C
Coefficient of Variation
P-value equals 9.17997e-45
h = 0.04
CV2
P-value equals 9.17997e-45
h = 0.0016
Fano Factor
P-value equals 2.1678e-53
h = 0.017
Distance to Median
P-value equals 0.568684
h = 0.043
count
value
sample
DM
HLM
D
E
CV
mean
CV value
0.10-0.20
0.09-0.10
0.08-0.09
0.07-0.08
0.06-0.07
0.05-0.06
0.04-0.05
0.03-0.04
0.02-0.03
0.01-0.02
0-0.01

Figure 3.2: *Principal Component Analysis and metric of variance comparison.*

A: High light mock dataset principal component analysis, carried out using *prcomp* function of the *stats* package. Datapoints highlighted in red circles were recognized to be too dissimilar from others in their group and removed. Each datapoint corresponds to a single bio-replicate. Bioreplicates from the same time points, in groups of four, are grouped by colour. Relevant bio-replicates are labelled.

B: Drought mock dataset principal component analysis, carried out using *prcomp* function of the *stats* package. Datapoints highlighted in red circles were recognized to be too dissimilar from others in their group, and removed. Each datapoint corresponds to a single bio-replicate. Bioreplicates from the same time points, in groups of four, are grouped by colour. Relevant bio-replicates are labelled.

C: Comparison between DM and HLM values for various metrics of variance calculation. Vertical line represents 0.04 cutoff point for CV, its derivative 0.0016 for CV2, or alternative cutoff values for Fano factor and Distance to Median.

D, E: Comparison between mean expression values and CV in HLM (**D**) and DM (**E**). Each datapoint corresponds to a gene. Datapoints are coloured by their CV values, with more variable genes assigned brighter colours. Horizontal line represents the 0.04 cutoff point for CV.

3.1.3 Gene ontology analysis indicates enrichment of stress response-associated Biological Processes within high variability geneset.

For GO analysis, 1542 genes passed the threshold value in DM, 1895 in HLM, and 680 in both DM and HLM, termed the “consensus list”. Gene ontology results of DAVID, Panther, and GOfuncR using Wilcoxon rank-sum test method, were plotted. Without filtering, the most terms by far were returned by GOfuncR (figure 3.7A), which is consistent with its method of operation in contrast to the other two methods, as it analysed a much larger number of genes. A more reliable overview was given by filtering the output. Figure 3.7B, which depicts the output filtered so that a term must feature in both DAVID and Panther, in addition to at least one GOfuncR analysis, depicts a number of gene ontology terms relevant to stress response. Majority of these terms covered genes responding directly to a stimulus – be it an abiotic change to the environment, like water deprivation or cold, a biotic stress, like bacterium or fungus, or a chemical involved in stress response, like karrikins, abscisic acid, and jasmonic acid. The remaining terms were either processes involved in stress response – “toxin catabolic process”, “defence response”, or in regulation of defence response, like “regulation of systemic acquired resistance” and “regulation of defence response”. The lone exception was “leaf senescence”, which is linked to stress response in *Arabidopsis*. The same trend of terms associated with stress response being enriched in high variability gene-set remains present in unfiltered GO results as well (appendix 1).

Interestingly, analysing HLM and DM results passing the threshold separately, instead of together in the consensus list, shows the differences between the two. While a number of terms differing between output for DM and HLM is not relevant

(appendix 2), and two terms present, GO:0009816 and GO:0055114, have been marked as obsolete, a few could clearly be linked to the circumstances behind the two datasets (figure 3.7C). Two of the terms unique to HLM, “circadian rhythm”, and “cellular manganese ion homeostasis”, are an example. The first is directly related to the circadian rhythm, while the second may be present because of the role of manganese in photosynthesis, the intensity of which varies over time. In DM, there were three differing terms relevant to experimental design, which are “leaf senescence”, “xyloglucan metabolic process” and “cell wall macromolecule catabolic process”. All three represent processes involved in plant growth and development – the first because senescence is an important fixture of these processes, while the latter two both concern processes altering the cell wall.

3.1.4 Cell housekeeping genes are enriched within the low-variability dataset.

In order to contrast results of high variability geneset gene ontology biological process analysis, a similar process was carried out to analyse low variability genes (figure 3.8)**.** 6151 genes were below the threshold for the HLM list, 4490 genes for the DM list, and 2625 genes for the negative consensus list. Due to a different distribution of low-variability genes against high-variability genes, a different proportion of the total geneset was analysed (figure 3.8A). As before, vast majority of results were obtained by use of the GOfuncR method (figure 3.8B). However, the previous method of filtering was insufficient, as DAVID results featured only one GO term in output, “covalent chromatin modification” (appendix 3). As a result, terms that appeared in output of at least three methods were considered for analysis (figure 3.8C). Biological process terms present within the low variability geneset overwhelmingly relate to the so-called “cellular housekeeping functions”, which are

necessary for survival of the cell, as well as other processes, such as “meiosis I” and “cell cycle”, that are consistently present within a significant portion of cells in the leaf. Additionally, the few terms that were identified as depleted within the low-variability geneset all are, themselves, involved in response to environmental factors.

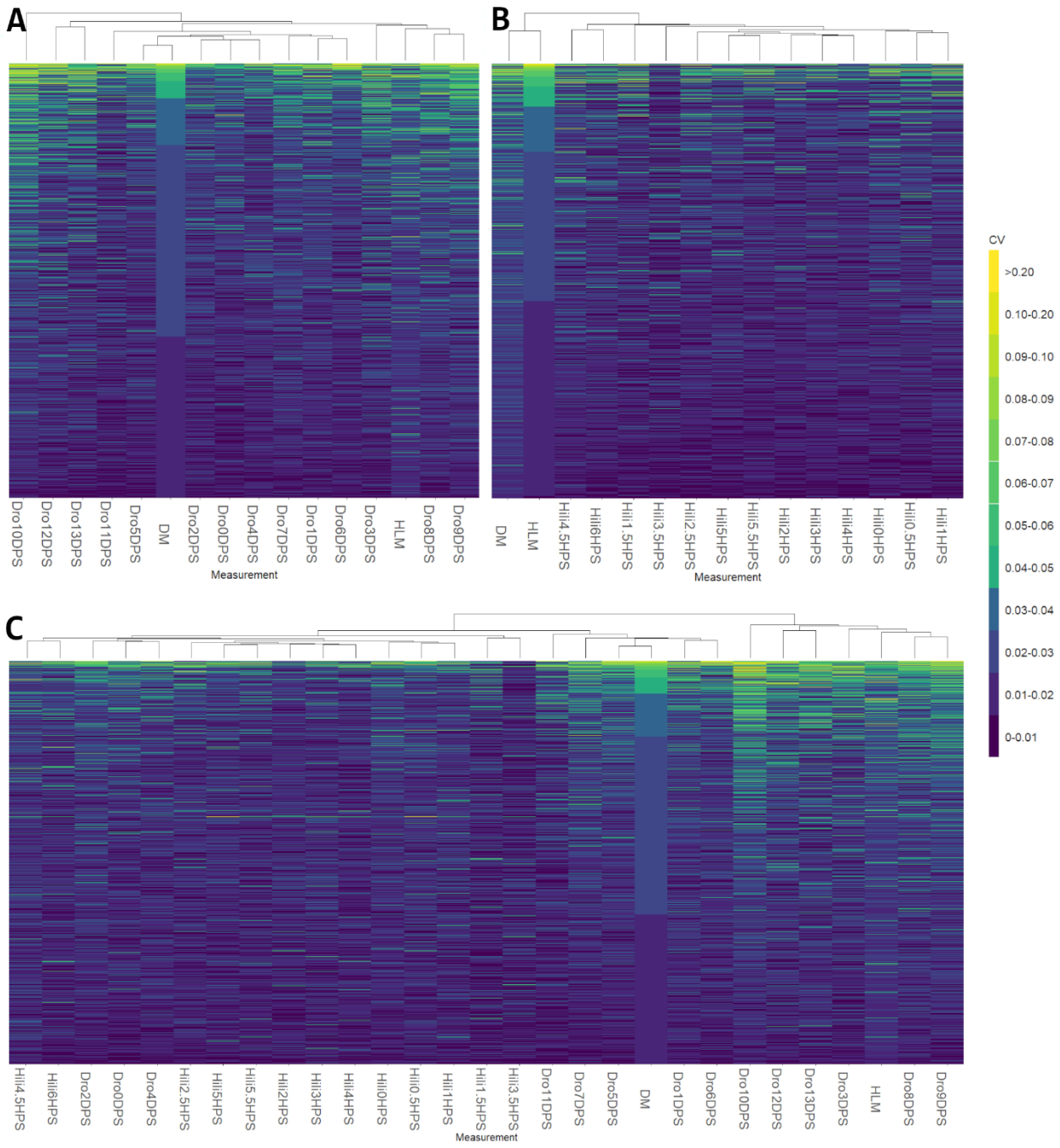

Figure 3.3: *Comparison of coefficient of variation across series.*

All heatmaps depict distribution of coefficient of variation within their samples. Each column represents a distinct time point, or CV calculated between all bio-replicates in a series (HLM and DM). Columns have been clustered by similarity. Mock drought has been abbreviated as “Dro”, whereas mock high light has been abbreviated as “Hili”. “DPS” stands for “days past start”, and “HPS” starts for “hours past start”. Samples coloured green or in brighter colours represent CV values passing the threshold of variability, those coloured blue and purple represent CV values below the threshold.

A: Clustered heatmap of all CV values for all genes in timepoints of mock drought dataset and DM and HLM, ordered by CV of genes in DM.

B: Clustered heatmap of all CV values for all genes in timepoints of mock high light dataset and DM and HLM, ordered by CV of genes in HLM.

C: Clustered heatmap of all CV values for all genes in timepoints of both datasets combined, and DM and HLM, ordered by CV of genes in DM.

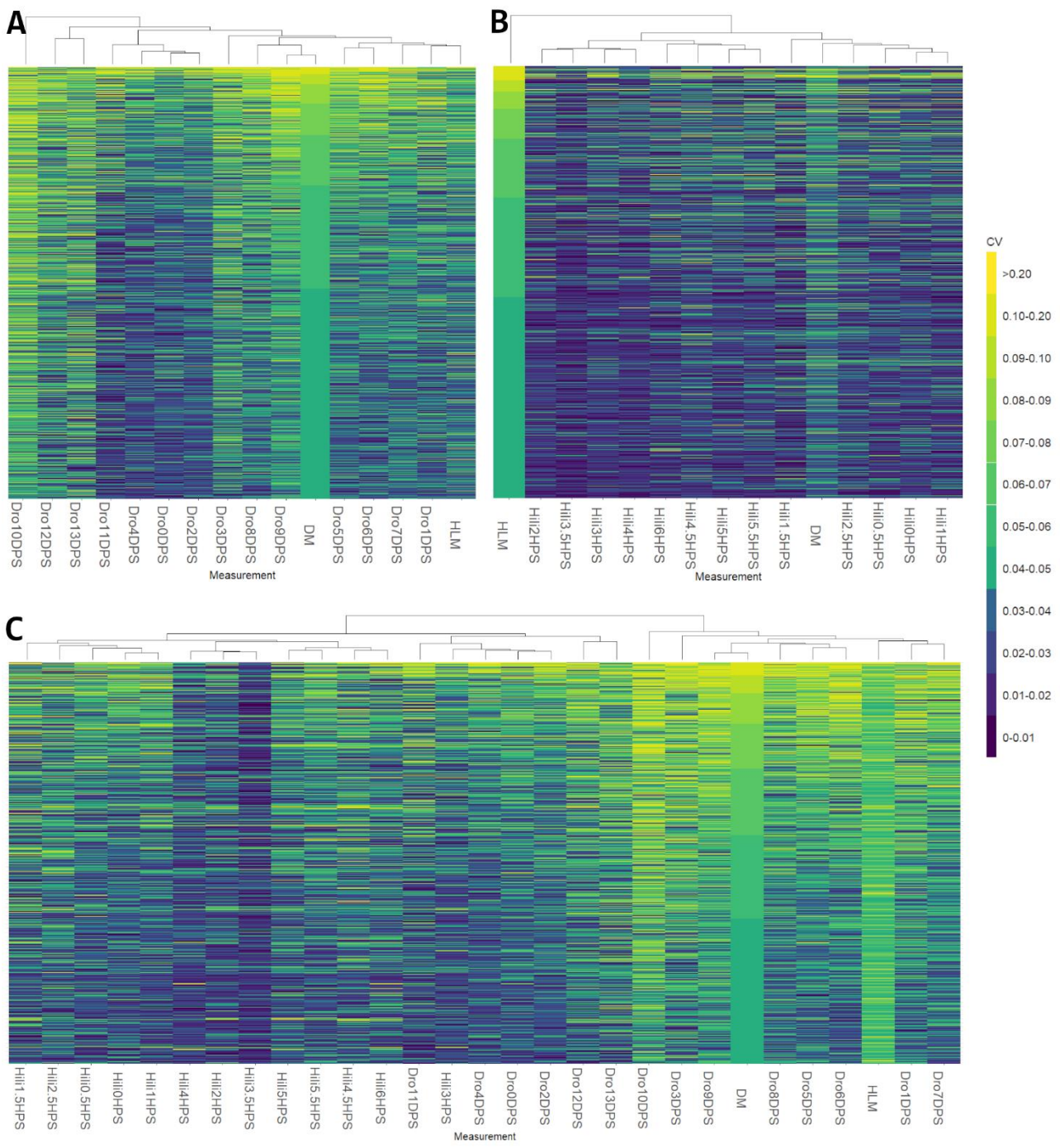

Figure 3.4: *Comparison of coefficient of variation across series after filtering.*

All heatmaps depict distribution of coefficient of variation within their samples. Each column represents a distinct time point, or CV calculated between all bio-replicates in a series (HLM and DM). Columns have been clustered by similarity. Mock drought has been abbreviated as “Dro”, whereas mock high light has been abbreviated as “Hili”. “DPS” stands for “days past start”, and “HPS” starts for “hours past start”. Samples coloured green or in brighter colours represent CV values passing the threshold of variability, those coloured blue and purple represent CV values below the threshold.

A: Clustered heatmap of CV values for genes in mock drought dataset, and DM and HLM,

for which CV > 0.04 in DM, ordered by CV of genes in DM.

B: Clustered heatmap of CV values for genes in mock high light dataset, and DM and HLM, for which CV > 0.04 in HLM, ordered by CV of genes in HLM.

C: Clustered heatmap of CV values for genes in both datasets, and DM and HLM, for which CV > 0.04 in HLM and CV > 0.04 in DM, ordered by CV of genes in DM.

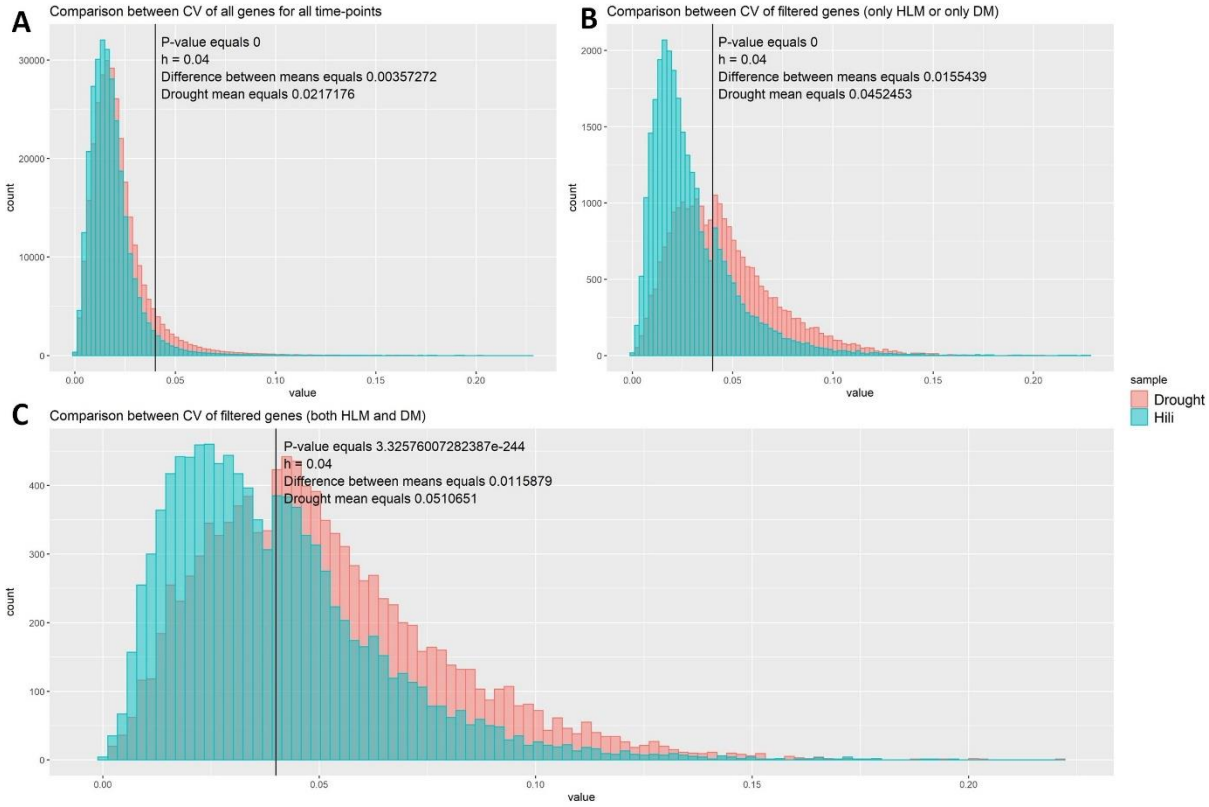

Figure 3.5: *Comparison between distribution of coefficient of variation values for mock drought and mock high light.*

The three histograms each compare distribution between CV values of selected genes for timepoints in mock drought dataset and of selected genes for timepoints in mock high light dataset. High light has been abbreviated as “Hili”. HLM and DM have been excluded. P-value is calculated using Wilcoxon test, carried out by *wilcox.test* function of the R stats package, with the null hypothesis that the distributions of mock drought and high light CVs differ by a location shift of 0.

A: Histogram comparing CV between mock drought and mock high light for all genes within all time-points. Mock drought sample represents data shown in figure 3.3A, whereas mock high light sample represents data shown in figure 3.3B.

B: Histogram comparing CV between drought and high light for genes whose CV passes the 0.04 threshold in DM or HLM respectively. Mock drought sample represents data shown in figure 3.4A, whereas mock high light sample represents data shown in figure 3.4B.

C: Histogram comparing CV within bio-replicates between mock drought and mock high light for genes whose CV passes the 0.04 threshold in both DM and HLM. Both samples represent data depicted in figure 3.4C.

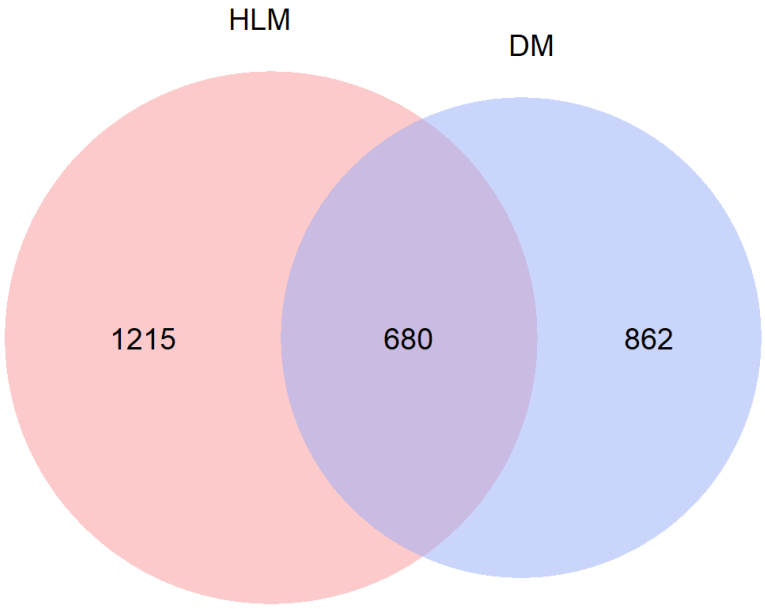

Figure 3.6: *Comparison between genes passing the CV threshold in both control datasets.*

The Venn diagram shows correlation between genes passing the CV threshold of 0.04 in HLM and DM. The numbers represent the amount of genes within each group—genes present in just HLM, just in DM, or in both.

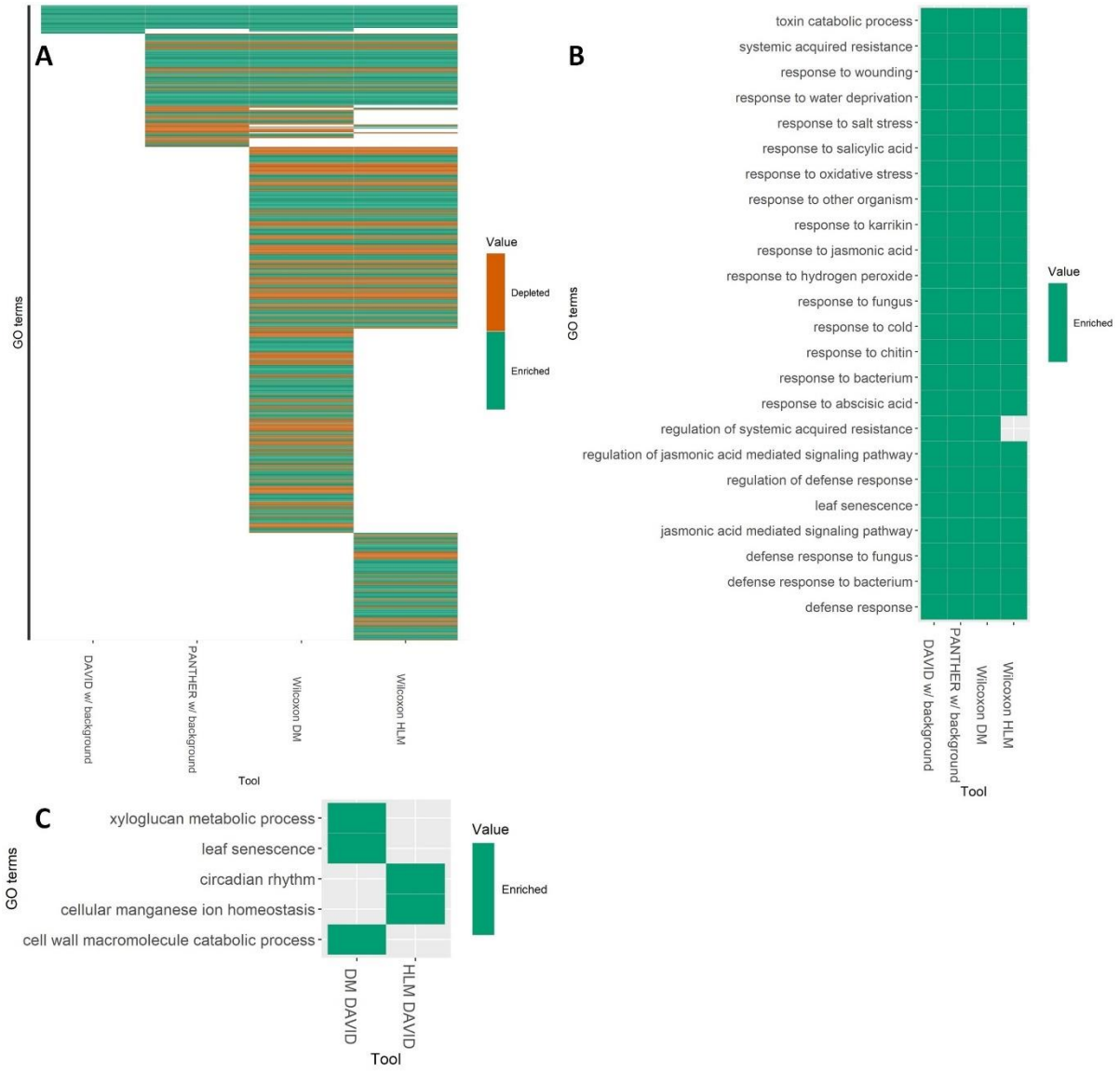

Figure 3.7: *High-variability gene ontology analysis.*

A, B: Comparison between Gene Ontology Biological Process (GO BP) results from DAVID, Panther, and GOfuncR utilizing Wilcoxon rank-sum test. Both DAVID and Panther were supplied with the consensus list (list of genes with CV higher than 0.04 for DM and HLM), and with list of all analysed genes as background. GOfuncR was supplied with CV values for DM and HLM separately. The q-value was used for DAVID and Panther results, and calculated from p-value for GOfuncR results using *p.adjust* function of *stats* package. Only results with q-value lower than 0.05 are represented.

A: An unfiltered heatmap of all GO terms returned by DAVID, Panther, and GOfuncR analysis. Terms shown in green are enriched within the high variability subset of genes, whereas terms in orange are depleted. DAVID results only feature enrichment information.

B: Heatmap of GO terms that are present in output of both DAVID and Panther, and in output of GOfuncR for at least one dataset, DM or HLM. As a result, only enriched terms are present.

C: Comparison between selected differing terms between DAVID GO BP results for HLM or DM. For both gene lists, genes with CV values higher than 0.04 in their respective measurement group were included, and analysis was carried out using background of all analysed genes. Only terms with q-value lower than 0.05 were analysed. Terms were selected based on their relevancy to the nature of the two experiments.

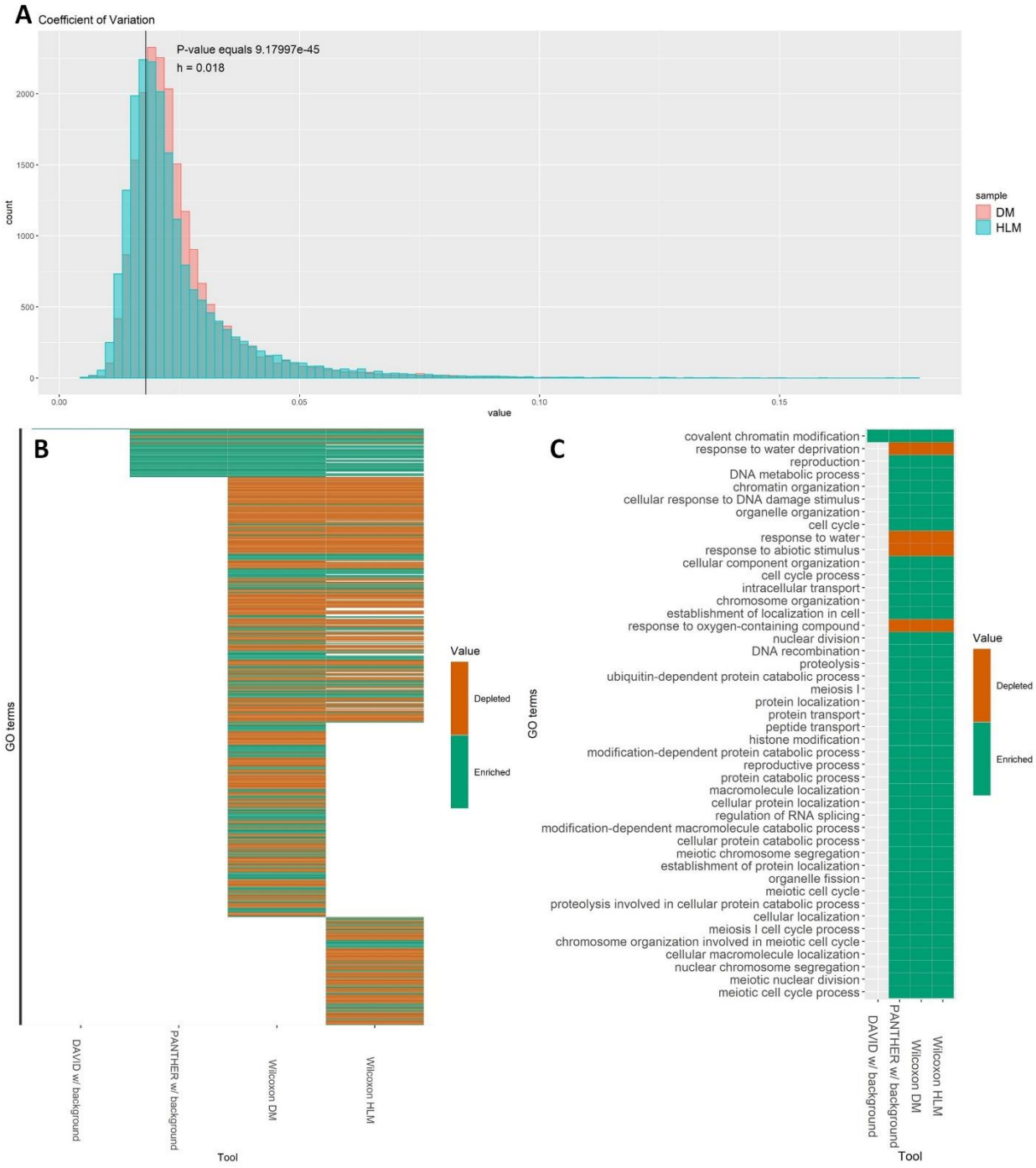

Figure 3.8: *Low-variability gene ontology analysis*

A: Histogram depicting coefficient of variation values present in DM and HLM. Vertical line represents the selected threshold, genes below which were selected. p-value was calculated using *wilcox.test*, with the null hypothesis that the distributions of mock drought and high light CVs differ by a location shift of 0.

B,C: Comparison between Gene Ontology Biological Process (GO BP) results from DAVID, Panther, and GOfuncR utilizing Wilcoxon rank-sum test. Both DAVID and Panther were supplied with the negative consensus list (list of genes with CV lower than 0.018 in DM and HLM), and with list of all analysed genes as background. GOfuncR was supplied with CV

values for DM and HLM separately. The q-value was used for DAVID and Panther results, and calculated from p-value for GOfuncR results using *p.adjust* function of *stats* package. Only results with q-value lower than 0.05 are represented.

B: Unfiltered gene ontology Biological Process results for low-variability geneset. Terms in green are enriched, terms in orange are depleted.

C: Filtered GO BP results for low-variability geneset. Only terms featuring in output of at least three analyses are shown.

3.2. Methylation analysis of variable genes.

The 19239 genes analysed in Chapter 3.1 were subjected to methylation analysis, using publicly available bisulfite sequencing data of *A. thaliana* Col-0 ecotype generated by (Stroud et al., 2012) and (Stroud et al., 2013), accessible through the Gene Ontology Omnibus (SRR501624, SRR534177 and SRR534193). Of those, 20 could not be assigned methylation proportions, leaving 19219 successfully annotated. Additionally, gene expression mean and coefficient of variation values for each gene as calculated between all mock drought bio-replicates over 14 days (DM) and all mock high light bio-replicates over 6 hours (HLM) were imported (see 2.1-2.5).

The low resolution profiles and base analysis of wild type BS-seq data showed that the three bio-replicates were largely similar (figure 3.9A-C). For all three methylation profiles, the area of the centromere was highly methylated compared to the rest of the chromosome. Additionally, while the samples diverge in methylation magnitude, the pattern of methylation they follow is largely similar. The spatial correlation of methylation in samples in all three contexts is similar as well (figure 3.9D-F), with negligible differences. In sequencing coverage, the differences are greater. Bioreplicate 3 displayed consistently lower coverage for all minimum numbers of reads compared to bioreplicates 1 and 2 (figure 3.9G).

3.2.1 Genes with high expression variability in both long and short-term samples are less methylated compared to non-variable genes.

The comparison of expression coefficient of variation and mean of methylation proportions (figure 3.10, figure 3.11) revealed significant differences between

variable and non-variable genes. The magnitude of difference was different between gbM genes, genes with transposable element-like methylation, and promoters. For the purposes of this analysis, genes were classified as having transposable element-like methylation if their methylation percentage in CHG or CHH contexts was higher than 0.05. Remaining genes were classified as gene body methylation (gbM) genes. Promoters were analysed based on their methylation and expression of their downstream genes.

For CG context, more methylated genes appeared variable in HLM than in DM (figure 3.10). In spite of this, the methylation of genes that were variable in the two datasets was significantly lesser than that of genes that were non-variable in both samples (figure 3.11). The difference was greatest in magnitude in genes with transposable element-like methylation – however, there were only 34 transposable element-like methylated genes in DM, and 36 in HLM, of which 15 featured in both. Between promoter methylation and gbM genes, the effect was more significant for gene body methylation. The difference between methylation proportions of genes non-variable and variable in both samples was 0.081 for gbM genes, and 0.042 for promoters, and both were statistically significant (figure 3.11).

For CHG and CHH contexts, the difference between methylation proportions remained statistically significant for both the methylation of promoters, as well as genes with transposable element-like methylation. The magnitude of the difference was much smaller for transposable elements, however, with the difference of means at 0.01 in CHG context and 0.005 in CHH context. Due to previous filtering step, measurements of gbM genes in CHG and CHH contexts intentionally did not include any genes with high methylation proportion in these contexts.

Because the coefficient of variation can be sensitive to minor changes in expression, methylation proportions were compared to mean of expression. The results showed that genes with transposable element-like methylation had predominantly low or medium expression (figure 3.12). The statistical analysis (figure 3.13) revealed that the relationship between expression intensity and methylation proportions varied significantly between the two gene methylation types, and promoter methylation. For gbM genes, those with medium expression were statistically the most methylated. Genes with high expression are second, and genes with low expression have the lowest methylation. For promoters, genes with low expression were more strongly methylated, while there was no statistical difference between those with high and medium expression (figure 3.13). For genes with transposable element-like methylation, genes with low expression had highest methylation compared to medium-expression and the high-expression genes (figure 3.13).

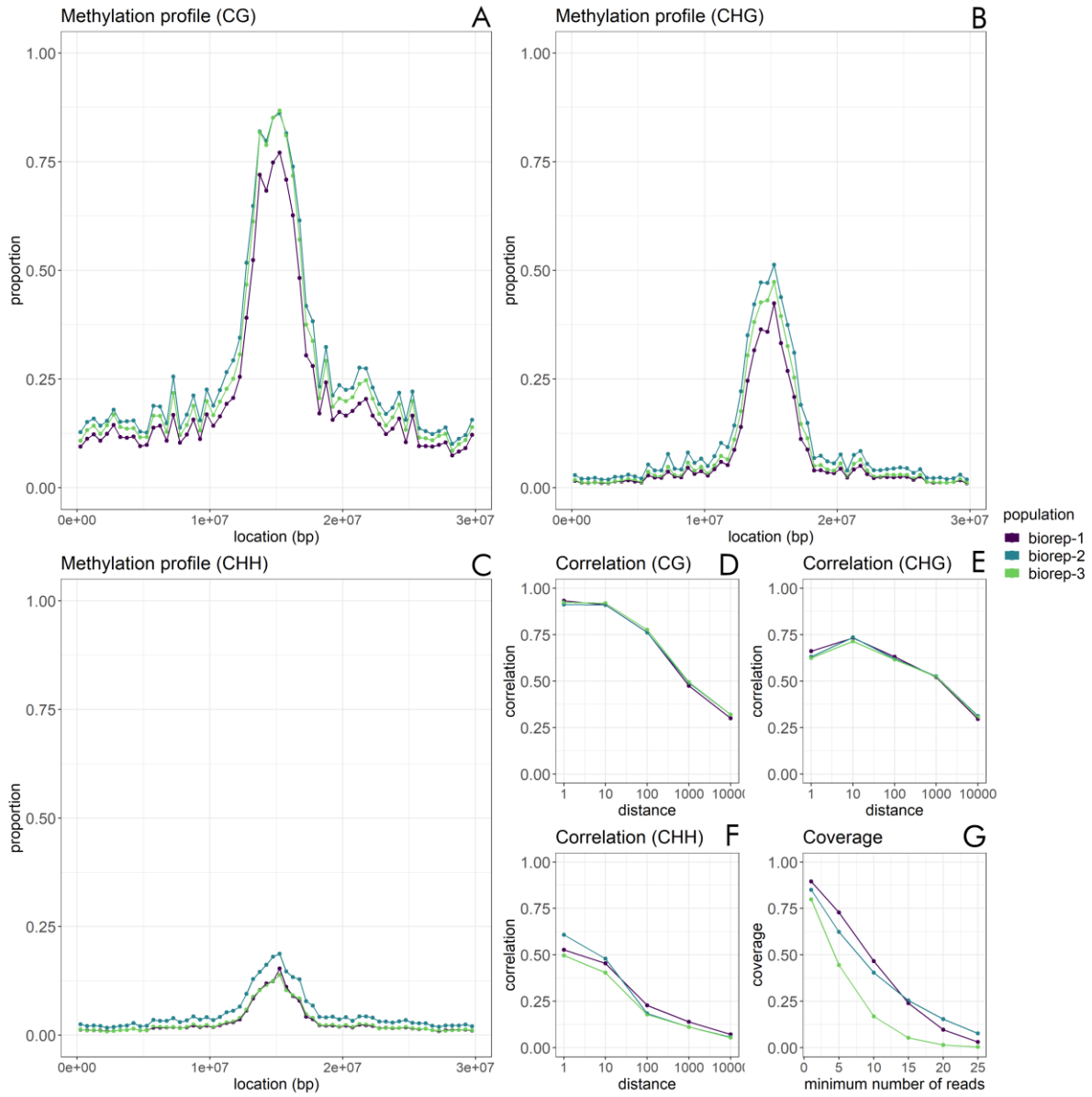

Figure 3.9: *Genome-wide methylation trends and base analysis of three wild type bio-replicates.*

A, B, C: Line chart depicting low resolution methylation profiles of *Arabidopsis* chromosome 1, in 500000 bp resolution, in CG (**A**), CHG (**B**) and CHH (**C**) contexts.

D, E, F: Line chart showing the spatial correlation of methylation of cytosines calculated for the entire *Arabidopsis* genome, in CG (**D**), CHG (**E**) and CHH (**F**) contexts.

G: Line chart showing the sequencing coverage per cytosine methylated in the CG context for selected minimum numbers of reads for the entire *Arabidopsis* genome.

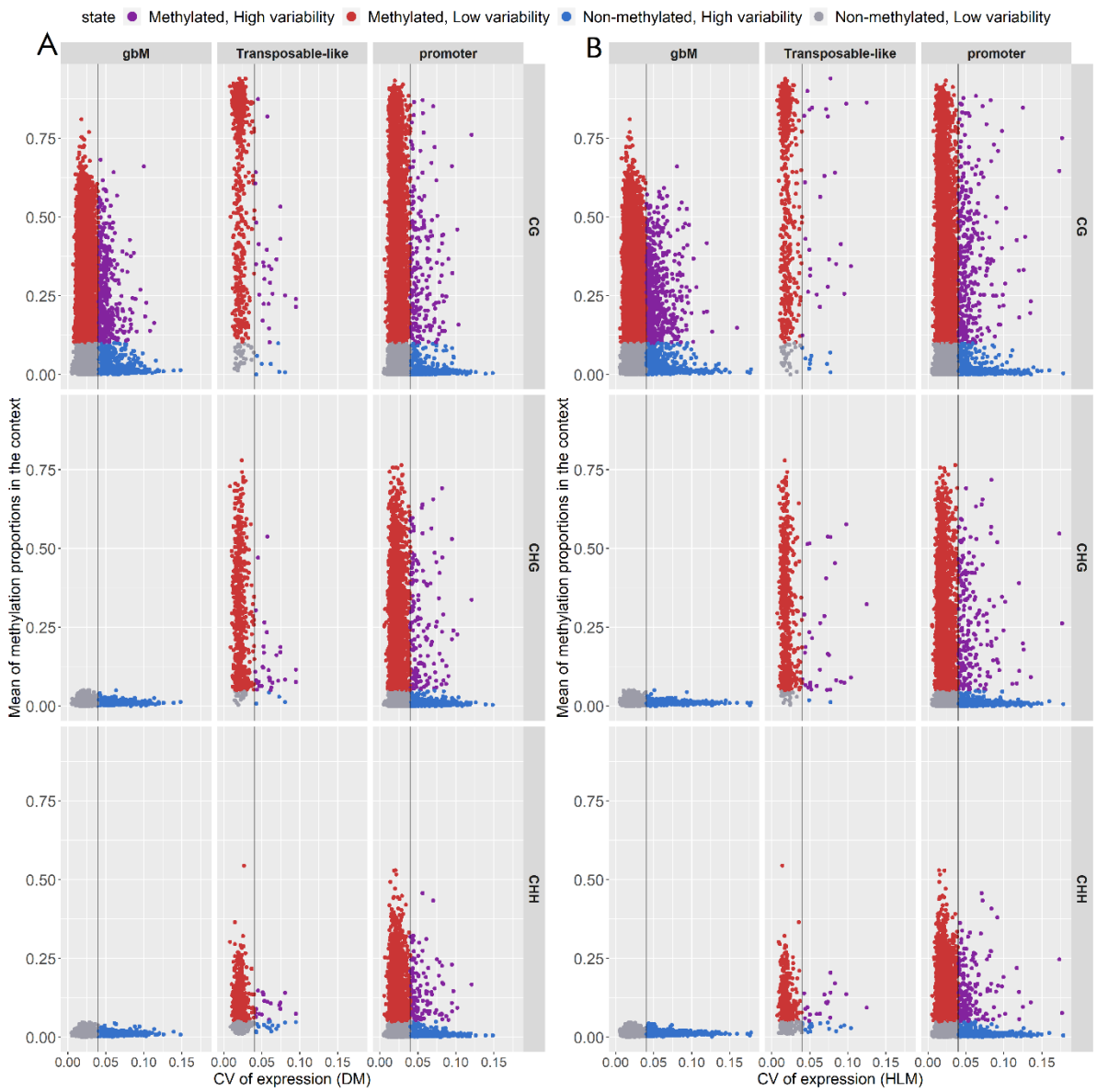

Figure 3.10: *Relationship between coefficient of variation for gene expression and proportion of methylation.*

For each scatterplot, x-axis represents the coefficient of variation (CV) of gene expression for a gene in the expression sample, and y-axis represents the mean of methylation proportion values in the selected context. Genes are categorised based on proportion of methylation and CV of expression.
Each dot represents a gene or a promoter. Red dots represent non-variable and methylated genes, purple represent variable and methylated genes, blue represent variable and non-methylated genes, and grey represent non-variable and non-methylated genes.
The scatterplots are split into columns by methylation types – gbM genes, genes with transposable element-like methylation, and promoter methylation. Each row shows methylation proportion in a different methylation context. These are, sequentially, CG, CHG, and CHH.

A: Methylation proportion compared to CV of expression values across all mock drought bio-

replicates (DM)
B: Methylation proportion compared to CV of expression values across all mock high light bio-replicates (HLM)

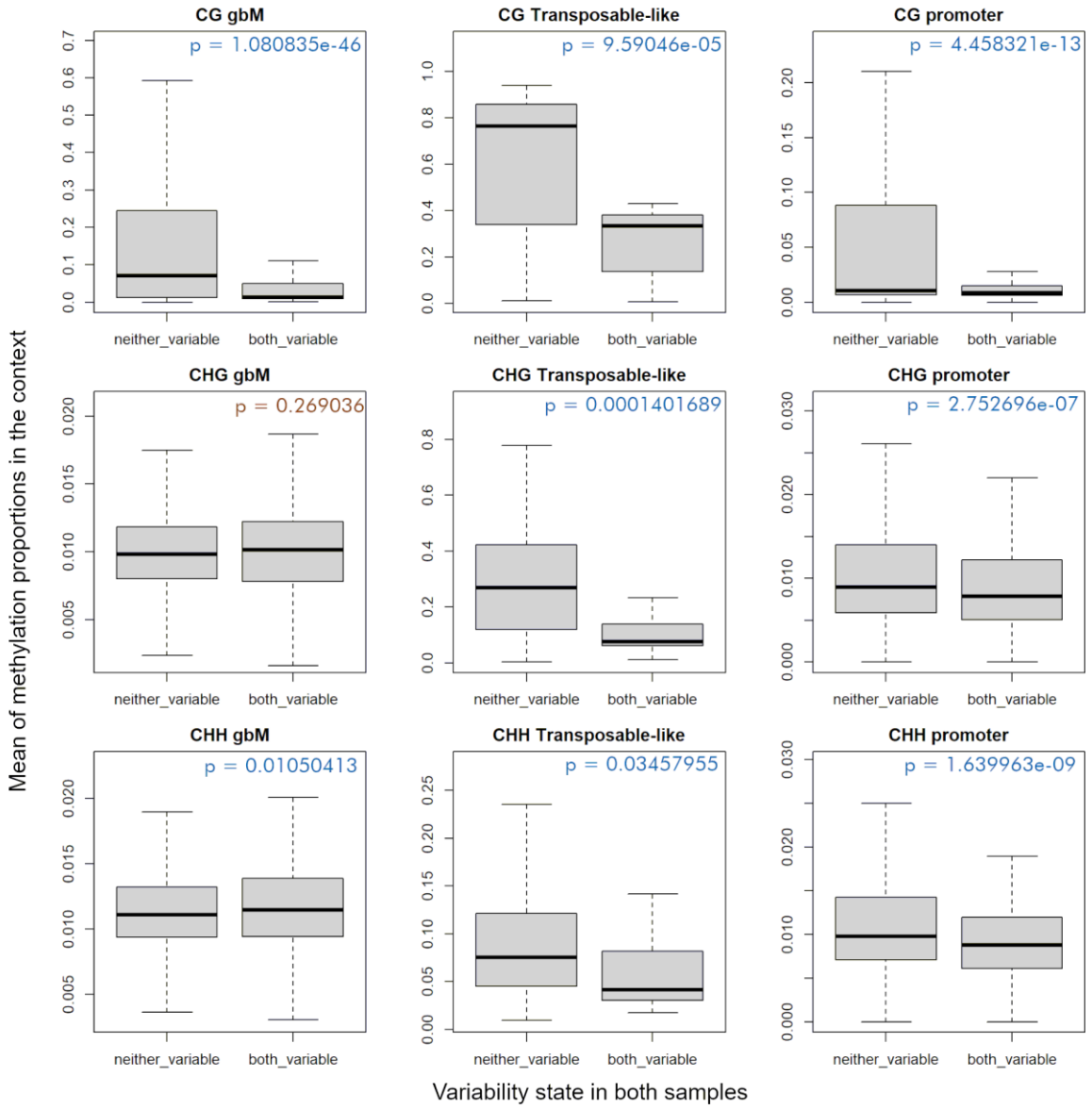

Figure 3.11: *Statistical analysis of the relationship between gene expression coefficient of variation and methylation proportion.*

For each boxplot, the x-axis represents CV of expression, while the y-axis represents methylation proportion in the selected context. The plots are split into columns by methylation types – gbM genes, genes with transposable element-like methylation, and promoter methylation, and into rows by methylation contexts – CG, CHG, and CHH.

“Neither variable” contains genes where CV <= 0.04 in both HLM and DM, “both variable” contains genes where CV > 0.04 in both HLM and DM. P-values were calculated using Wilcoxon rank-sum test, with null hypothesis that the distributions differ by location shift of 0.

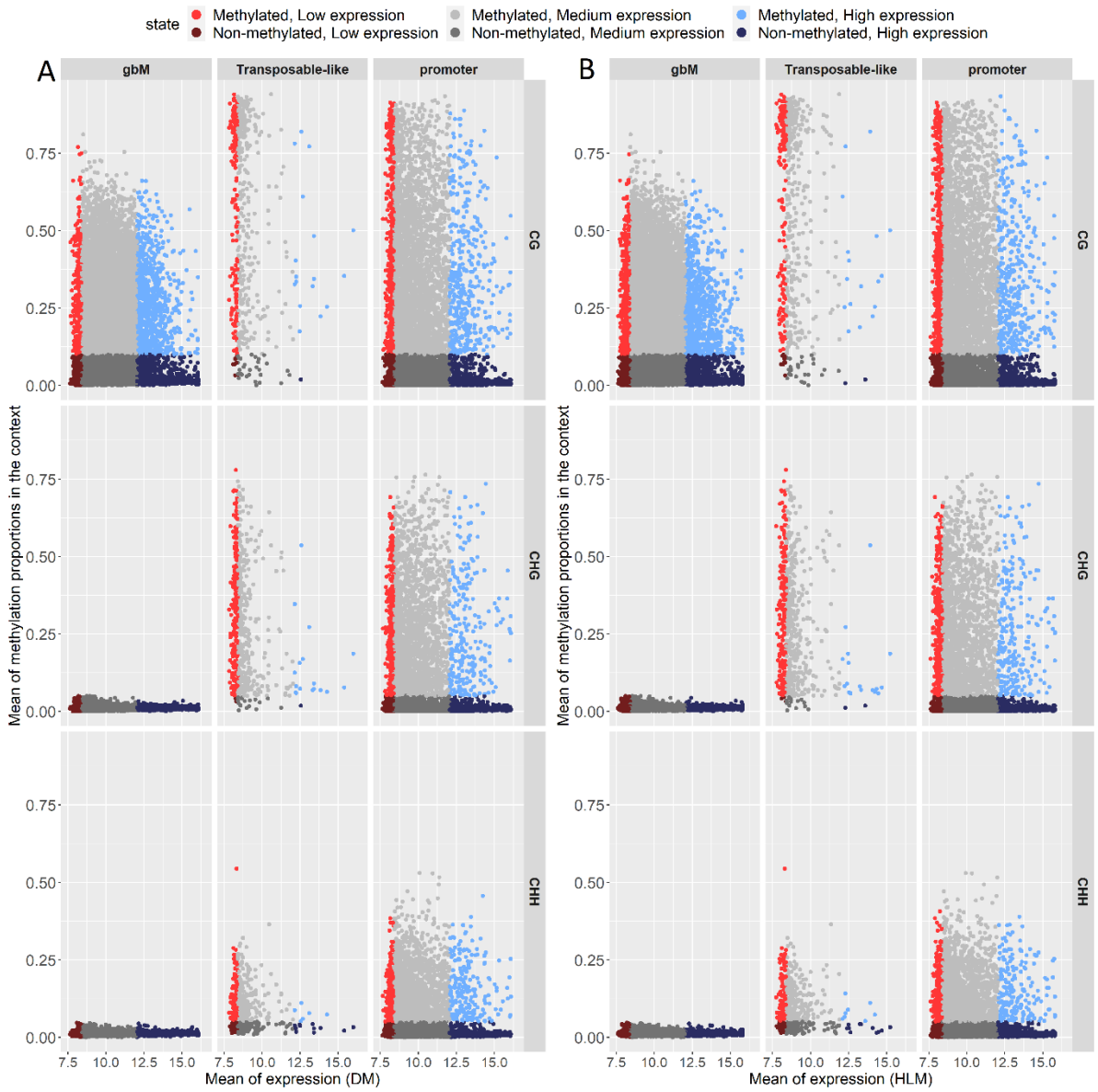

Figure 3.12: *Gene expression mean and methylation proportion comparison between methylation contexts, methylation types, and two expression samples.*

For each scatterplot, x-axis represents the mean of gene expression for a gene in the expression sample, and the y-axis represents the mean of methylation proportion values in the selected context. Genes are categorised based on methylation proportion and mean of expression.
The scatterplots are split into columns by methylation types – gbM genes, genes with transposable element-like methylation, and promoter methylation.
Each dot represents a gene or a promoter. Red dots represent the approximately 10% least expressed genes, blue dots represent the approximately 10% most expressed genes, and grey dots represent genes with medium expression. Lighter colours represent methylated genes, darker colours represent non-methylated genes.
Each row shows methylation proportion in a different methylation context. These are, sequentially, CG, CHG, and CHH.

A: Methylation proportion compared to mean of expression values across all mock drought bio-replicates (DM)
B: Methylation proportion compared to mean of expression values across all mock high light bio-replicates (HLM)

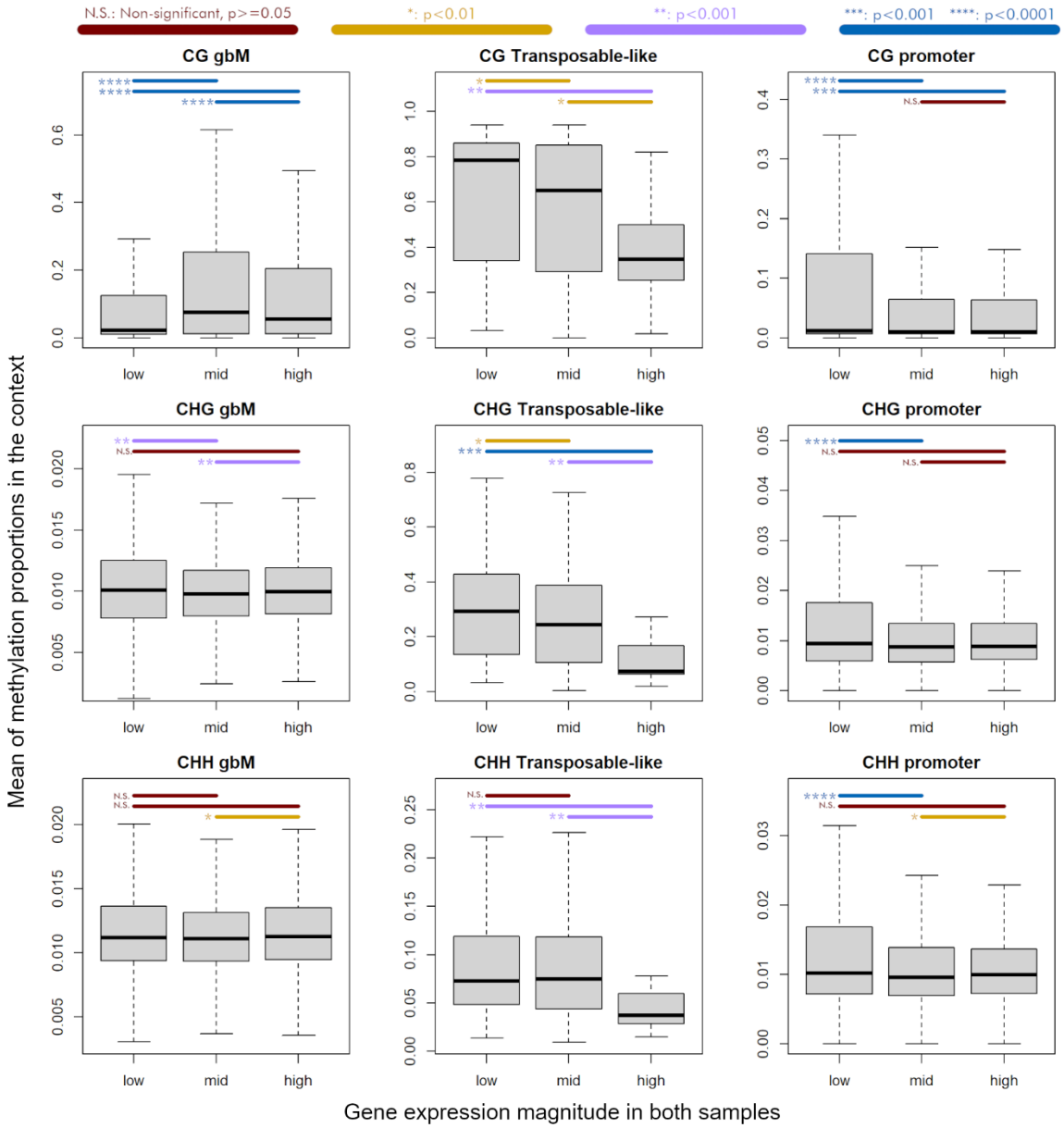

Figure 3.13: *Statistical analysis of the relationship between gene expression magnitude and methylation proportion in different methylation contexts and methylation types.*

For each boxplot, the x-axis represents expression value, while the y-axis represents methylation proportion in the selected context. The plots are split into columns by methylation types – gbM genes, genes with transposable element-like methylation, and promoter methylation, and into rows by methylation contexts – CG, CHG, and CHH.

"Low" – expression values < 8.4; "medium" – expression values between 8.4 and 12.05; "high" – expression values > 12.05. Only genes falling into the same category in HLM and DM were analysed. P-values were calculated using Wilcoxon rank-sum test, with null hypothesis that the distributions differ by location shift of 0.
The colour of the line between two boxes indicates the statistical relationship between the two distributions as calculated by Wilcoxon rank-sum test. Blue is p<0.001; purple is p

between 0.001 and 0.01; orange is p between 0.01 and 0.05, and red is non-significant. In descending order, the lines represent the relationships between low-to-medium, low-to-high, and medium-to-high distributions.

Table 3.1: *Comparison of selected Gene Ontology Biological Process terms enriched in CG methylation gene groups.*

Each column represents a different set of genes, based on expression variability and methylation proportion.

Variable genes – CV > 0.04; non-variable genes – CV <=0.04
Methylated genes – proportion of methylation > 0.1; non-methylated genes: proportion of methylation <= 0.1
Only genes with gbM methylation are analysed. Lists are composed of genes that are in the same category in both DM and HLM.
Lists were analysed for Gene Ontology BP terms using PANTHER. Only selected output terms are present on the table. All terms have FDR-adjusted p value less than 0.05.

	Non-variable, methylated in both HLM and DM	**Variable, methylated**	**Variable, non-methylated**	**Non-variable, non-methylated**
Enriched	Endosomal transport; Protein acylation; Gene silencing; Cell cycle	Response to organonitrogen compound; Regulation of defence response	L-pipecolic acid biosynthetic process; Regulation of defense response to insect; Response to salicylic acid	Regulation of transcription, DNA-templated; Regulation of RNA biosynthetic process;
Depleted	Cellular response to hypoxia; Detoxification; Cell wall modification; Response to water	N/A	Chromosome organization; ncRNA metabolic process; RNA processing; translation	Regulation of RNA splicing; DsRNA processing; Protein acylation; Histone modification; Regulation of response to stress

3.2.2 Gene categories derived from methylation and variability vary in enriched and depleted Gene Ontology Biological Process terms.

Genes were split into 4 categories and analysed through PANTHER Gene Ontology Biological Process analysis (table 3.1). The genes were split based on their methylation proportion, and on their variability in HLM and DM, generating four categories.

For CG methylation of gbM genes, 7223 genes were non-variable and methylated, 118 genes were variable and methylated, 547 genes were variable and non-methylated, and 8670 genes were non-variable and non-methylated (appendices 4-7).

Similarities between the GO terms enriched between groups have emerged, which help identify the role played by both the expression variability and methylation. The two non-variable groups are more similar to each-other than the two variable groups. Both variable groups, methylated and unmethylated, were enriched for GO terms involved in stress response, which is consistent with the previous analysis (3.1.3-3.1.4), and both non-variable groups were depleted for some terms related to stress response.

One similarity present between variable and non-variable non-methylated genes is that both were depleted for terms related to RNA processing. While the non-variable non-methylated gene group was enriched for “regulation of RNA biosynthetic process”, both non-methylated groups were depleted for “RNA metabolic process” and its child term, “mRNA metabolic process”. This was reversed in non-variable methylated gene-set, where both of these terms were enriched, and regulatory genes were depleted. The difference between the methylation of genes regulating

RNA metabolism and those responsible for that process is consistent with past research (Zhang et al., 2006). The significantly higher degree similarity between groups sharing variability status compared to groups sharing methylation status confirms that, while methylation and gene expression variability may be correlated, methylation plays varied roles in regulation of transcription beyond just determining variability, which too is in line with existing knowledge (Yang et al., 2015).

In summation, these analyses demonstrate that a relationship exists between methylation and gene expression variability where highly methylated genes are less variable, which is particularly significant for gbM genes. The results of gene ontology analysis indicate that both play functions within the organism, although those of gene expression variability are more limited in scope.

3.3. Comparison of variation in wild type and methylation loss mutants.

To analyse the relationship between gene expression variability and CG methylation, separate gene expression and methylation datasets were used, containing data for wild type (WT), methyltransferase mutant-1 (*met1-1*) and methyltransferase mutant-3 (*met1-3*), generated by (Catoni et al., 2017), accessible through the GEO (GSE89592). In both mutants, CG methylation is reduced – partially in *met1-1* (Kankel et al., 2003), and near totally in *met1-3* (Saze et al., 2003) (Baek et al., 2011). 29670 genes in total were identified in the expression arrays used.

527 genes were removed for *met1-1* measurement because of lack of sufficient methylation data, incorrect gene assignment, or low number of reads, and 570 genes were removed for *met1-3* measurement.

The chromosome-wide profiles and base analysis of WT and methyltransferase mutant BS-seq data show that only CG methylation levels were significantly

decreased between the WT and both *met1* mutants (figure 3.14A-C). Likewise, in CHG and CHH contexts, there was little difference between the spatial correlation of methylation (figure 3.14D-F) of WT and *met1* mutants. Methylation coverage in CG context for the three samples was similar, with *met1-3* displaying slightly higher coverage than *met1-1* and WT (figure 3.14G). Spatial correlation in the CG context differed significantly, with highest degree of correlation for WT, and lowest for *met1-3* (figure 3.14D). This implies that, in *met1* mutants, existing methylation patterns have been disturbed.

3.3.1 Methyltransferase-1 mutants differ from wild type in gene expression and CG methylation.

In order to establish differences between methyltransferase mutants and WT, an analysis was carried out to identify differently methylated regions (DMRs), and to identify any genes overlapping these DMRs. For the purposes of this analysis, genes were classified as having transposable element-like methylation if their methylation percentage in CHG or CHH contexts was higher than 0.05. Remaining genes were classified as either gbM methylated, if their CG methylation proportion was higher than 0.1, or non-methylated, if it was equal to or below 0.1. Promoters were analysed based on their methylation and expression of their downstream genes. Non-methylated genes were analysed separately from gbM genes to aid in GO analysis.

For gbM genes in WT and *met1-1* comparison, 8663 genes overlapped with CG methylation loss DMRs, meaning they lost methylation, and 358 did not. (figure 3.15A). For gbM genes in WT and *met1-3* comparison, 8730 genes overlapped with CG methylation loss DMRs, and 294 did not.

The differences between WT and mutants are not limited to methylation, but also affect gene expression. Some genes, both among those that lost methylation, and among those that maintained methylation, changed their expression significantly between WT and mutant (figure 3.15B). In order to identify the impact of change in methylation on gene expression variability, only genes that both lost methylation and did not significantly differ in expression between WT and mutants were analysed.

Table 3.2: *Methylation and expression data processing outcomes.*

This table contains information about the genes that were removed from the analysed geneset, as described in methods (see 2.19 to 2.24). It contains two types of rows – "removal" rows, and "information" rows.

Removal rows have 3 cells. The leftmost column describes the amount of genes before removal, and the rightmost column describes the amount of genes after removal. The details column in the middle describes how many genes were removed, and the reason for removal. After the second information row, the removal is conducted separately for WT-*met1-1* and WT-*met1-3* lists.
Information rows contain information about the genes at their current analysis step.

Starting gene number	Details	Post-filtering gene number
29670 genes extracted from sequencing data. 29575 with correct gene IDs, 93 with updated gene IDs, and 2 with obsolete gene IDs.		
29670	Removal of 2 obsolete gene IDs	29668
29668	Removal of 93 genes that had their IDs updated, 92 of which were duplicates of other genes	29575
29575	Removal of 46 genes for which methylation proportion information could not be extracted in one or more contexts	29529
Of 29529 genes analysed, 4693 have transposable element-like methylation, 9038 have gbM, and 15798 are non-methylated		
29529	**For *met1-1***: 420 genes were removed because of very low bisulfite sequencing coverage of less than 25 reads over the entire gene	29109
29109	**For *met1-1***: 9 genes were removed because they overlapped gain DMRs	**29143**
29529	**For *met1-3*:** 386 genes were removed because of very low bisulfite sequencing coverage of less than 25 reads over the entire gene	29100
29100	**For *met1-3*:** 0 genes were removed because they overlapped gain DMRs	**29100**

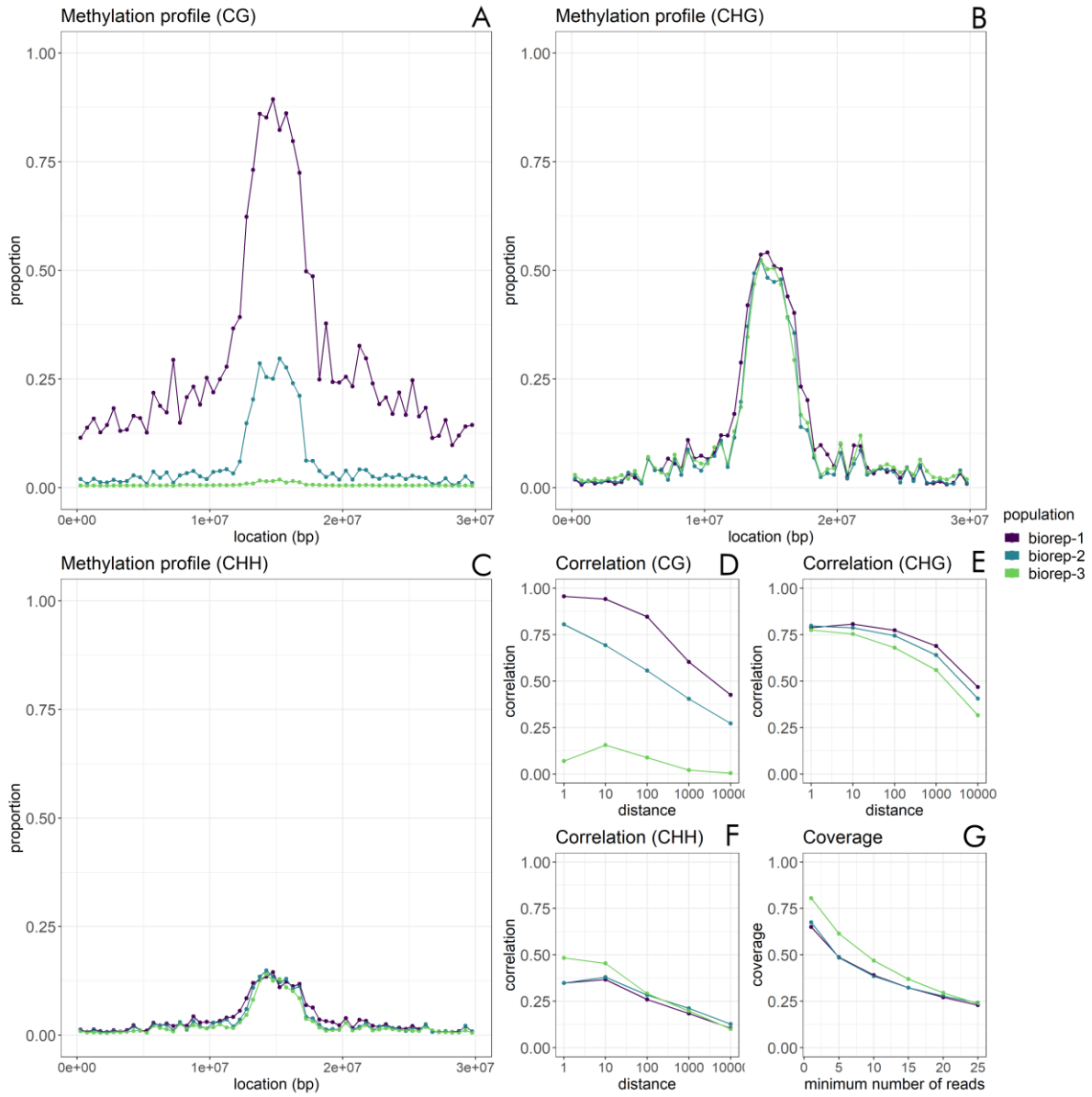

Figure 3.14: *Genome-wide methylation trends and base analysis of wild type and met1 mutants.*

A, B, C: Line chart depicting low resolution methylation profiles of *Arabidopsis* chromosome 1, in 500000 bp resolution, in CG (**A**), CHG (**B**) and CHH (**C**) contexts.

D, E, F: Line chart showing the spatial correlation of methylation of cytosines calculated for the entire *Arabidopsis* genome, in CG (**D**), CHG (**E**) and CHH (**F**) contexts.

G: Line chart showing the sequencing coverage per cytosine methylated in the CG context for selected minimum numbers of reads for the entire *Arabidopsis* genome.

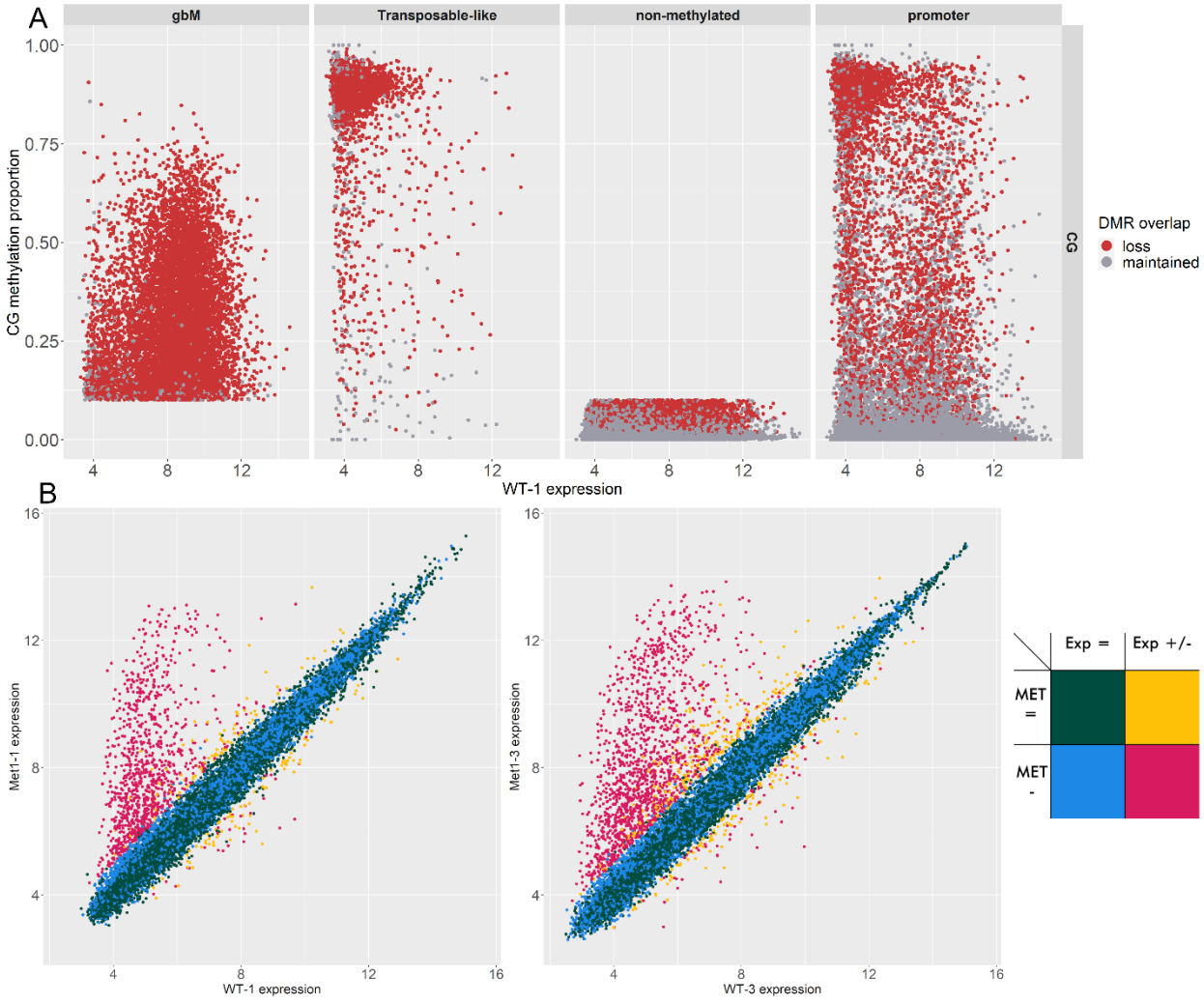

Figure 3.15: *Comparison of changes between met1 mutants and wild type.*

A: The comparison between expression in wild type for *met1-1* (WT-1) and CG methylation proportion in WT, with points coloured based on their overlap with DMRs between WT-1 and *met1-1*.
Each dot represents a gene or a promoter. Genes were split into 3 groups— genes with transposable element-like methylation, which have either CHG or CHH methylation proportion higher than 0.05, gbM genes, which have CG methylation proportion higher than 0.1 but CHG and CHH methylation proportion equal to or below 0.05, and non-methylated genes, which have CG methylation proportion below or equal to 0.1, and CHG and CHH below or equal to 0.05. x-axis represents WT expression, y-axis represents proportion of methylated cytosines in CG context. Each point represents a gene.

B: The comparison between WT-1 and *met1-1* gene expression (left), and WT-3 and *met1-3* gene expression (right). Genes are separated into four groups: No DMR overlaps and maintained expression (green), no DMR overlaps and significantly altered expression (yellow), overlap with loss DMRs and maintained expression (blue), and overlap with loss DMRs and significantly altered expression (red). x-axis represents expression in WT, y-axis represents expression in met-1 mutant. Each point represents a gene.

3.3.2 Expression variability of genes overlapping with loss DMRs changes in CG methylation loss mutants.

In this analysis, only genes and promoters that overlapped DMRs where CG methylation was lost and with unchanged expression between WT-1 and *met1-1* and WT-3 and *met1-3* were analysed, in order to analyse the effect of methylation loss on variability. Differentially expressed genes were excluded because their expression changes render coefficient of variation calculation unreliable.

In total, 14136 genes were analysed between WT-1 and *met1-1*. The change in expression coefficient of variation varies significantly between genes (figure 3.16A). The coefficient of variation of most genes was not significantly altered (figure 3.16B). For all three categories of genes, more genes increased in variability rather than decreased as a result of methylation loss. Likewise, more hypomethylated promoters were associated with increased variability. The ratio of genes with increased variability to decreased variability was highest in gbM genes, and lowest in genes with transposable element-like methylation.

For WT-3 and *met1-3*, the trend is reversed (figure 3.16B), with greater amount of hypomethylated genes and promoters decreased in CV than increased. A smaller amount of genes was analysed here than for *met1-1*, 13262. As before, the variability of most genes was not significantly altered. The ratio of genes with decreased variability to increased variability was highest in genes with transposable element-like methylation, and lowest in gbM genes.

The changes in variability observed in methylation loss mutants support the theory that a connection between gene expression variability and methylation exists. However, the difference between the direction of the trend in *met1-1* and *met1-3*

mutants shows that the changes are highly varied, and implies that a closer examination is necessary.

3.3.3 Increase in expression coefficient of variation of hypomethylated gbM genes is statistically significant.

The gene contents of groups identified in previous analysis were compared, to determine overlaps between *met1-1* and *met1-3* mutants (figure 3.17). For no group was the amount of genes overlapping between groups greater than 50% of the sum of the groups (figure 3.17A). The fraction of genes shared between the two samples is lowest for gbM genes with decreased variability, and highest for gbM genes with no significant change in variability.

Statistical analysis (figure 3.17B) shows that the overlap between genes without significant change in variability in *met1-1* and *met1-3* was significantly greater than that of genes with increased CV or decreased CV in all four groups. For gbM and non-methylated genes, and for promoters, the size of overlaps of increased CV genes was statistically significantly greater than that of decreased CV genes in these groups, with greatest difference for gbM genes. Within transposable element-like methylated gene group, there was no significant difference between the sizes of increased CV and decreased CV overlaps.

The results of this analysis show that the increase in variability as a result of methylation loss is more consistent than decrease in variability across the two mutants. This suggests that gene expression variability buffering is indeed one of the functions of methylation of gbM genes and, to a lesser extent, in genes with low methylation and promoters. The significant overlap in genes with unchanged

variability implies that this mechanism is not universal, however. Because of the greatest size of the gbM overlap, it was selected for a Gene Ontology analysis.

3.3.4 RNA-related Gene Ontology BP terms are enriched in analysis of gbM genes with increased expression variability overlapping between *met1-1* and *met1-3.*

The analysis of 605 genes with increased expression coefficient of variation in both the *met1-1* and *met1-3* mutants was carried out using PANTHER. 169 IDs were unmapped, and 1 gene was mapped to another, leaving 436 uniquely mapped IDs. Of the significant (FDR lesser than 0.05) terms, the most highly enriched was “ribosomal small subunit export from nucleus”, at 30.07 fold enrichment. “Ribosome localization”, “RNA metabolic process”, “nucleic acid metabolic process” and “cellular macromolecule metabolic process” to were among enriched terms. No terms were significantly depleted within the geneset (appendix 8). The results of the gene ontology analysis partially match the terms enriched in non-variable methylated genes (see 3.2.2).

Of the two families of GO terms enriched in gbM genes that gained variability in both hypomethylated mutants, the RNA metabolic process is the more interesting one. As the measurements of gene expression used throughout this work rely on mRNA sequencing, it is feasible that the significant disparity between variability gain and loss gene-sets is the result of variance in processes that play a role in production and degradation of RNA molecules.

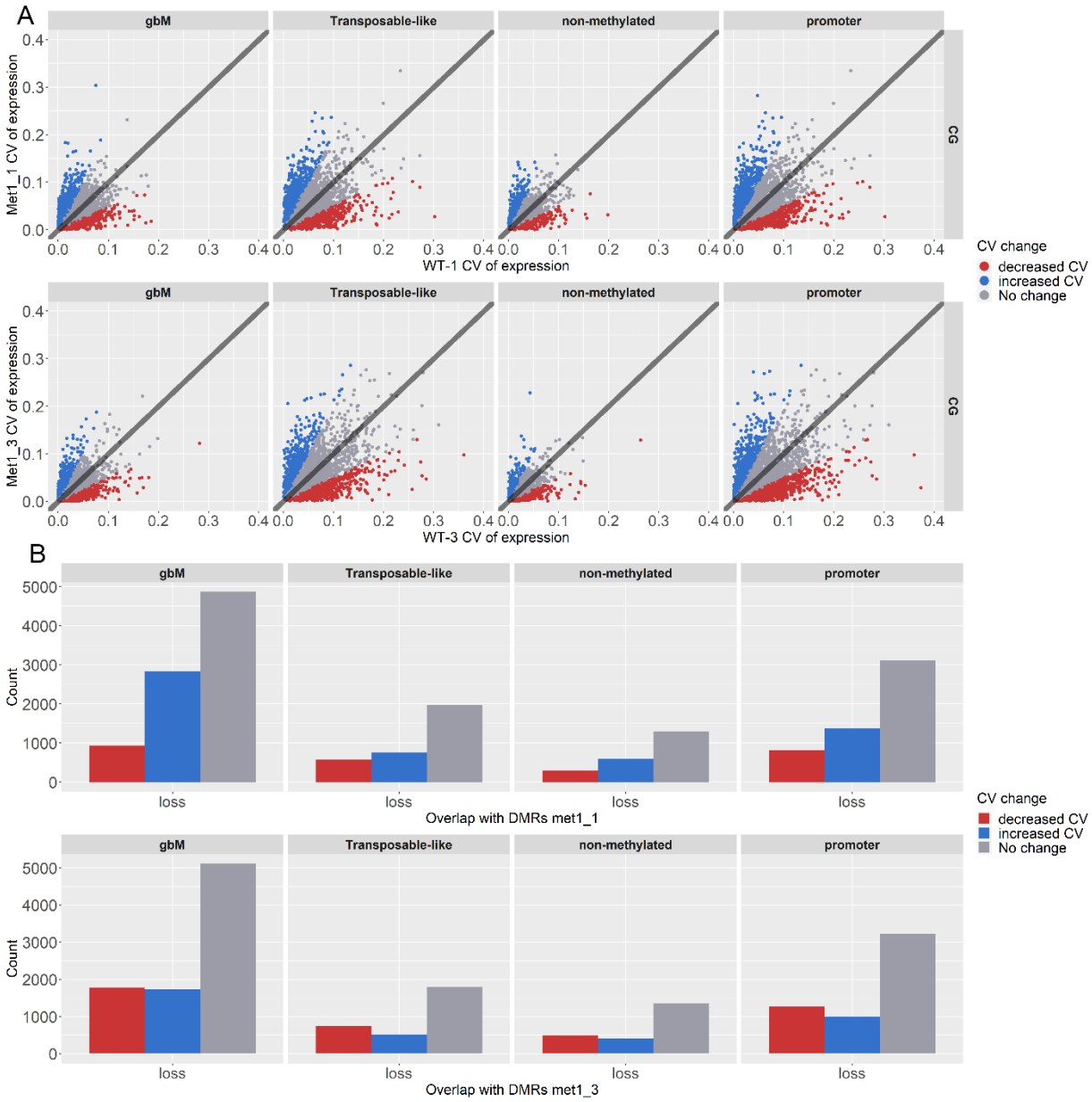

Figure 3.16: *Comparison between coefficient of variation of expression in WT and met1 mutants.*

A: Comparison of WT-1 coefficient of variation against *met1-1*, and WT-3 coefficient of variation against *met1-3*. Only genes without significant expression change between WT and mutant which overlap with loss DMRs are analysed. Each dot represents a gene or a promoter. The x-axis represents CV of expression in WT, and the y-axis represents CV of expression in met1 mutant. The diagonal line represents 1:1 ratio. Genes are categorised based on their coefficient of variation change: increased for genes with CV fold change greater than 1, decreased for genes with CV fold change less than –1, and no change for those in-between.

B: Relative sizes of each gene category shown in panel A. The top barplot depicts the sizes of categories between WT1 and *met1-1*, and bottom one between WT-3 and *met1-3*.

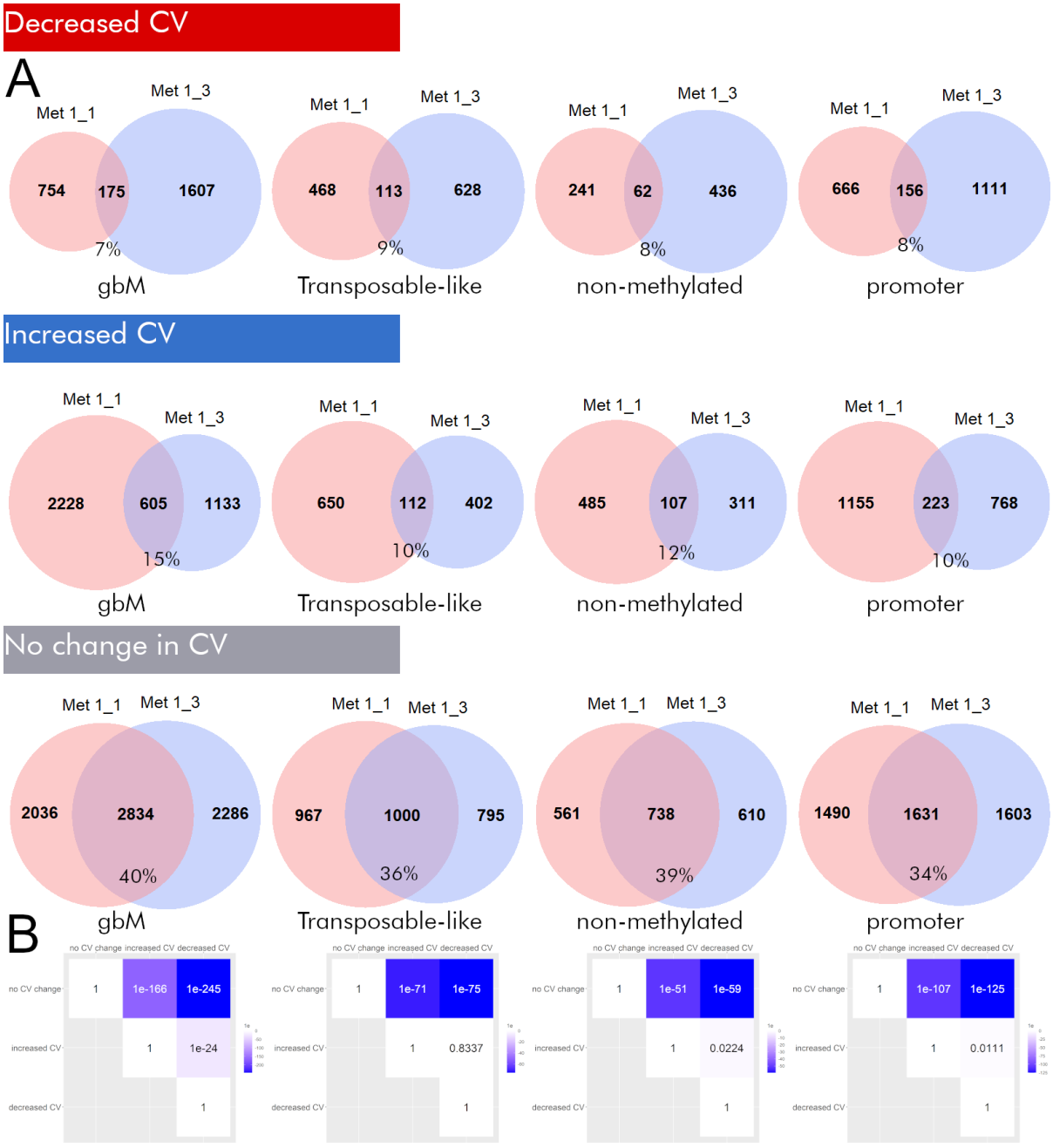

Figure 3.17: *Comparison between coefficient of variation changes in met-1 mutants.*

A: Venn diagrams, depicting relative sizes of each category as shown in **figure 3.16A**, as well as overlaps between them. The number shows the amount of genes in either the overlap or in the mutant alone. Red circles represent WT-1 to *met1-1* comparison, blue WT-3 to *met1-3* comparison. Percentage amount represents the fraction of genes shared between the two samples. Genes are split into 4 categories, and comparisons are made separately for each of the three CV states (decreased, increased, or not significantly changed).

B: The statistical relationship between overlaps for each category in panel A, calculated using Fisher's Exact Test. Numbers on the matrix depict p-values.

4. Discussion

4.1 Gene expression variability patterns.

Within both the long-term measurement, mock drought (DM), and the short-term measurement, mock high light (HLM), genes were identified to greatly vary in the CV calculated between biological replicates across different time-points. This behaviour was expected in both series. The data suggests that, for short term measurements, gene expression was altered between time points as a result of the circadian clock (Salomé and McClung, 2004), as shown by clustering together of measurements taken closely to each-other in time (figure 3.3B, figure 3.4B). Additionally, previous research has already identified that inter-individual gene expression variability of genes can vary between specific times of day (Cortijo et al., 2019). Similarly, while the long-term measurements did not suffer from differences in variation introduced by the circadian rhythm as they were not sampled at different times, they, in turn, included effects caused by plant development (Schmid et al., 2005).

As such, in both mock drought and mock highlight, the coefficient of variation value calculated between all bio-replicates also captured differences in expression resulting from sources other than inherent variability. This effect can be analysed more closely by comparing CV of genes variable in DM or HLM against their CV within individual measurements. This comparison showed that many genes were variable in both DM and their individual measurements (figure 3.4A), albeit in HLM this percentage was smaller (figure 3.4B). To correct this, only genes variable in both DM and HLM were analysed (figure 3.4C). This way, the gene expression variability

caused by differing timescales and the presence of circadian rhythm should be removed, leaving “true” underlying variability. The effectiveness of this is evidenced by significantly increased average CV within the measurement of genes from the short timescale sample (figure 3.5C).

The use of microarray sequencing data, as opposed to RNA-seq, imposes certain limitations. RNA-seq is more accurate in analysis of genes with low or very high expression (Zhao et al., 2014), and therefore provides a greater range of readings, which would allow for much greater accuracy in calculation of variation. In spite of this advantage, microarray-based datasets were selected for this analysis, as, historically, microarrays have been less costly than RNA-seq. As such, a larger amount of bio-replicates can be analysed with the same amount of resources, meaning the availability of microarray-based time series is greater. Moreover, the two series selected were grown in identical conditions, which enabled analysis based on consensus-list approach.

Indeed, it is highly likely that microarrays are responsible for the anomalous distribution of CV values for genes with lowest and highest mean expression values (figure 3.2D-E). The fact that only genes at extremes of the expression scale are affected suggests that it is not a consequence of CV metric alone – and, as stated above, it is those genes that microarrays have trouble analysing. Genes with high expression magnitude can reach the upper intensity limit of microarrays, at which point they are assigned very similar expression values, resulting in significantly lower CV. Genes with very low expression, meanwhile, are more susceptible to noise when analysed using microarrays, which is responsible for increasing their CV. The amount of genes affected is low enough, however, that the benefit of analysing these large datasets outweighs the limitations placed upon them by use of microarrays.

While some other design elements of the datasets used impose certain limitations upon this work, such as low number of biological replicates measured at individual time-points, their use does come with a significant positive as well. Because both mock high light (Alvarez-Fernandez et al., 2021) and mock drought (Bechtold et al., 2016) are accompanied by stress series, they could be used to investigate variable genes in abiotic stress conditions in a future work.

4.2 Gene Ontology analysis of high variability and low variability genes.

The results of gene ontology analysis indicate that, in the high variability gene-set, every enriched term present in output of both DAVID and Panther, as well as Wilcoxon analysis of either HLM or DM, is related to stress response either directly, like “response to wounding”, or indirectly, like “leaf senescence” (figure 3.7B). This is consistent with previous research, which has identified that genes responsible for stress response are highly variable (Bar-Even et al., 2006) in yeast and in *Arabidopsis* in short timescales (Cortijo et al., 2019).

Some of the terms related to stress response were found to be depleted in low variability gene-set (figure 3.8C). This gene-set was enriched for terms related to cellular housekeeping, and reproduction – cell cycle, proteolysis, cellular localisation, DNA recombination, chromatin modification. This, too, is in line with past research, which has identified that low noise genes are enriched for housekeeping functions in *Mus musculus* single-cell data (Barroso et al., 2018), in *S. cerevisiae* (Fraser et al., 2004), and, indeed, in *Arabidopsis*, in short timescales (Cortijo et al., 2019).

The effect of noise on fitness has been found to vary significantly between genes in yeast (Keren et al., 2016). Because housekeeping genes are responsible for basal functions of the cell, it stands to reason that disruptions to their

concentrations would have a significant effect on functioning of the cell. Moreover, noise has the ability to propagate through gene networks (Pedraza and van Oudenaarden, 2005) – meaning that variability in genes involved in processes such as RNA splicing, DNA metabolism and histone modification, all of which are closely related to transcription, could significantly alter the phenotype. An important caveat, however, is that the research cited above focuses not on inter-individual variability, but instead on noise, which is variability not across individuals, but across time. Additionally, varied mechanisms exist that suppress propagation of noise (Singh and Hespanha, 2009) (Siciliano et al., 2013). Nonetheless, while the causes and mechanisms may be different, the outcomes of gene product concentration departing from the norm maybe be similar both events.

Likewise, the enrichment of biological process terms related to stress response in high variability gene-set can be theorised to be the result of separate evolutionary drive. In bacteria, stress survival strategies are not only a result of actions in response to environmental difference, such as expression of stress response genes. An alternative strategy is bet hedging (Veening et al., 2008) - even within isogenic populations, within uniform environments, a subset of cells exhibits alternative phenotypes that, in the event of a sudden environmental change, can survive. A similar mechanism has been identified in yeast (Levy et al., 2012). In principle, this method can be advantageous over switching phenotypes in response to environmental changes in environments that are more constant (Kussell, 2005), although it has not been documented in plants aside from seed dormancy.

The fact that the biological process terms enriched within high variability gene-set are related to each-other, and that the same is the case in the low variability gene-set, suggests that the organism does indeed control variability of

gene expression – although the exact reasons for this control are unknown, and the mechanisms used are poorly understood.

One significant limitation of this work lies in the fact that concentration of mRNA is, by itself, incapable of accurately predicting concentration of protein (Gygi et al., 1999). Protein concentration is a function of rate of translation, which itself a result of multiple factors that include mRNA concentration, and protein decay rate (Brockmann et al., 2007). As such, it is possible that post-transcriptional mechanisms could exist that would moderate the effects of variation in mRNA concentrations. Indeed, extensive regulation networks exist between various stages of gene expression (Dahan et al., 2011). However, since mRNA concentration is connected to protein concentration (Brockmann et al., 2007), it stands to reason that significant expression variability can result in protein concentration variation.

4.3 Methylation of variable and non-variable genes.

Existing research has uncovered some potential sources of variation in gene expression between isogenic populations. Certain histone modifications were found to be correlated with high variability and low variability genes (Cortijo et al., 2019), yet the finer details, and other factors responsible for regulation of variation, are currently unknown.

The analysis conducted here focuses on two relationships – that between gene expression magnitude and methylation, and that between gene expression variability and methylation. The comparison of expression magnitude and methylation shows that, in CG context, gbM genes with medium expression tend to have a statistically higher methylation than genes with low or high expression (figure 3.13). This is consistent with existing research, which has reported the same trend

between overall cytosine methylation, which is largely composed of CG-context methylation, and gene expression (Zilberman et al., 2007).

The results of analysis of promoter methylation in the CG context too are in accord with our current knowledge. Not only did research find that the majority of genes are unmethylated in their promoter regions (Zhang et al., 2006), which is consistent with the low median on the boxplot for all three expression level categories (figure 3.13), but promoter methylation is also associated with tissue-specific expression and, overall, with lower transcription (Zhang et al., 2006). Here, this may be represented by the statistically significant difference between methylation of promoters associated with low expression genes, and those of genes with medium or high expression. The larger distribution range of methylation proportion of promoters of low expression genes suggests that, unlike the other two categories, it contains genes with both low and high methylation (figure 3.13).

The statistical analysis of genes with transposable element-like methylation is less reliable, due to very low amount of genes with high expression in this group (figure 3.12). This is a result of the filtering step that ensured only genes with non-negligible CHG or CHH classified into this group – therefore, one can conclude that this type of methylation is rarely found in highly expressed genes, which suggests they play a role in repression of gene expression.

The comparison between coefficient of variation and methylation shows that genes that are variable in DM and HLM tend to have lower methylation (figure 3.11). The effect is stronger in genes with gbM rather than in promoters. For CHG and CHH methylation, the magnitude of the effect is lesser, but it remains significant for genes with transposable element-like methylation, and for promoters.

In context of the previous analysis, the lower CG methylation of promoters associated with variable genes helps address one of the issues with use of CV. While coefficient of variation is a useful metric, it can potentially be sensitive to comparatively minor changes in expression. As such, its potential downside could be that the analysis might have preferentially selected genes with low expression, rather than those with high physiological variability. If that were the case, however, promoters of variable genes would have higher methylation than those of non-variable genes, which is not the case here. Moreover, if that error were present, genes with transposable element-like methylation would show large variability, since a significant number of them has low expression (figure 3.12), whereas this analysis indicates that significant majority are non-variable (figure 3.10).

The fact that variable genes appear to have very little CG methylation is intriguing, and so it was followed up by a gene ontology analysis to identify enriched and depleted terms among gene-sets created by splitting analysed genes by their methylation and expression variability. Both variable groups are enriched for genes responsible for stress response, which is consistent with the previous GO analysis. There is some similarity between variable non-methylated genes and non-variable non-methylated genes as well. Both of these groups are depleted for genes responsible for RNA processing. These terms, meanwhile, are enriched in non-variable methylated gene-set. This is in line with existing research, which suggests that gene body methylation is associated with constitutively-expressed genes (Zhang et al., 2006), which encode housekeeping functions such as processing of RNA. Importantly, however, non-variable non-methylated genes include “regulation of DNA-templated transcription” and “regulation of RNA biosynthetic process” as enriched BP terms. This, too, matches previous analyses, as transcription factors –

which are partly responsible for regulation of transcription - were found to be one of the most undermethylated categories of genes (Zilberman et al., 2007).

One potential limitation of this analysis lies in mismatch between the age of plants used for expression analysis and the age of plants used to determine methylation – both mock high light (Alvarez-Fernandez et al., 2021) and mock drought (Bechtold et al., 2016) were 5 weeks old at the start of the measurement, while plants used for bisulfite sequencing were 3-4 weeks old (Stroud et al., 2012). As such, while both samples were gathered from the same tissue, it is feasible that methylation of some genes might differ between these times.

To summarise, the results of these analyses support the conclusion that gene expression variability and gene body CG methylation are linked to each-other. What they can not establish, however, is whether methylation and expression variability merely co-occur, or if gene methylation is responsible for moderating expression variability.

4.4 Gene expression differences between WT and *met1* mutants.

In light of previous analysis identifying correlation between CG gene body methylation and gene expression variability, this relationship was investigated further. *Arabidopsis thaliana* serves as a particularly useful organism for such an investigation, as genome-wide methylation loss mutants are not lethal (Kankel et al., 2003), meaning that effects of methylation loss can be examined on the scale of the whole genome.

As such, WT plants were compared to *met1-1* mutants, where CG methylation was significantly reduced, and *met1-3* mutants, where CG methylation was almost entirely lost. Unlike the previous analysis, plants sequenced for expression and

plants sequenced for methylation are of the same age (Catoni et al., 2017). Genes were split into 3 categories based on methylation, instead of just 2 - genes with transposable-like methylation were isolated as before, but what was previously termed as gbM genes was split into genes with significant gbM-only methylation and non-methylated genes.

The analysis of expression data indicates that loss of methylation significantly affected the expression of some genes (figure 3.15B), which is consistent with past analyses of mutant *met1* specimens (Zhang et al., 2006). These genes were excluded from gene expression variability analysis, because a significant portion of gene expression change seen in methylation loss mutants is the result of reactivation of transposable elements (Lister et al., 2008), which by their nature would not be uniform across bio-replicates. As such, their inclusion would potentially inflate measured coefficient of variation values of analysed genes. Even for upregulated or downregulated genes unrelated to transposable elements, major change in expression implies action of mechanisms which would not be present in majority of other genes.

Overlaps with loss differently methylated regions (DMRs) were used to identify genes that lost methylation, as opposed to calculating differences in methylation proportion. As such, even genes that have been classified as unmethylated using methodology of previous analyses may still be catalogued as having lost methylation, since while majority of the gene in WT might be unmethylated, some small region of it might have methylation that is then lost in *met1* mutants. Moreover, the gene-set analysed between WT and *met1-3* is different than that analysed between WT and *met1-1*, as the two differ for the amount genes that are differentially expressed or overlap loss DMRs.

The analysis comparing CV of genes in WT against CV of genes in *met1* mutants indicates that, for a majority of genes, CV is not significantly changed for any of the three gene groups or for promoters (figure 3.16B). Curiously, while in *met1-1* a larger amount of genes has gained variability than lost it, in *met1-3* this trend is reversed, with more genes losing variability than gaining it. In order to reconcile the differences between the two samples, gene groups sorted by CV change were compared (figure 3.17A).

The statistical analysis of the results revealed that the size of overlaps for genes with increased CV was greater than that for genes with decreased CV for gbM and non-methylated genes and promoters, and that for genes with transposable element-like methylation there was no statistically significant difference (figure 3.17B). This is partially explained by existing research – transposable elements that were not mobilised by loss of CG methylation might be kept redundantly demobilised by CHG and CHH context methylation (Kato et al., 2003), which, in this context, could plausibly also redundantly buffer gene expression variability in CG methylation loss mutants.

While the increase in CV is statistically significant, there remains the question of why so many genes do not match category between groups, and why both increases and decreases in CV are seen, even in genes with significant non-CG methylation. One potential explanation lies in the fact that noise is transmitted through gene networks (Pedraza and van Oudenaarden, 2005). If the assumption is made that variability can be treated in a similar manner to noise, then change in variability of some genes, such as those that were not analysed because of their changed expression, could have a wide-ranging impact on variability of other genes that themselves are not affected by loss of methylation.

In order to identify any similarities between genes that had their CV increased in both *met1-1* and *met1-3*, a GO BP analysis was carried out. The results indicate that genes related to ribosomal localisation, as well as RNA metabolic process, were enriched. In particular, enrichment of “RNA metabolic process” fits the previous GO analyses – while a different gene-set was used previously, this term was enriched in non-variable methylated gene-set, and depleted in both the variable and non-variable non-methylated gene-sets, fitting the profile of a gene group that was methylated and non-variable in WT, but lost methylation and gained variability in *met1* mutants. Additionally, this does support the hypothesis that some of the variability is not a result of genes losing methylation, but of cascading effect through regulatory networks. Because RNA metabolism is a housekeeping function, and one that is tied to transcription, it’s not implausible that variability in genes responsible for carrying it out would translate into cell-wide effects.

4.5 Conclusion.

This study has explored inter-individual gene expression variability in *Arabidopsis thaliana*. It has found that this variability is to some degree controlled by the organism, and that distinct functions are associated with both the high variability and low variability gene-sets. Additionally, it has analysed variable genes, and found a statistically strong relationship between CG methylation and variable genes. Comparison of WT and hypomethylated mutants has shown that, among genes without differential expression, loss of methylation leads to a variety of outcomes, with statistically greater amount of genes increasing expression variability rather than decreasing. This suggests that gene body methylation is one of the mechanisms buffering gene expression variability. The results of this study could be further built upon in the future by investigating genes with individually removed methylation to

isolate the effect from transcriptional changes resulting from genome-wide hypomethylation. These could potentially be complemented by gene network analyses to confirm whether wide-ranging gene expression variability changes are the result of small numbers of variable genes, and if so, which genes they are.

What are the reasons for the relationship between gene expression variability and methylation? One option that was considered was that low methylation is associated with chromatin state that is in some way conducive to gene expression variability. However, past research has discovered that variable genes are associated with compact environments (Cortijo et al., 2019), and that, in *Arabidopsis*, CG-context methylation is associated with chromatin condensation (Zhong et al., 2021), which conflicts with the findings presented here which find variable genes are less methylated. As such, it is unlikely that methylation's effects upon chromatin state are responsible for expression variability. While other potential reasons exist, the amount of genetic mechanisms interacting with methylation necessitates follow-up studies to identify the ones responsible for changes in expression variability.

5. References

Abley, K., Formosa-Jordan, P., Tavares, H., Chan, E.Y., Afsharinafar, M., Leyser, O. and Locke, J.C. 2021. An ABA-GA bistable switch can account for natural variation in the variability of Arabidopsis seed germination time. *eLife*. **10**, p.e59485.

Alemany, A., Florescu, M., Baron, C.S., Peterson-Maduro, J. and van Oudenaarden, A. 2018. Whole-organism clone tracing using single-cell sequencing. *Nature*. **556**(7699), pp.108–112.

Alvarez-Fernandez, R., Penfold, C.A., Galvez-Valdivieso, G., Exposito-Rodriguez, M., Stallard, E.J., Bowden, L., Moore, J.D., Mead, A., Davey, P.A., Matthews, J.S., and others 2021. Time-series transcriptomics reveals a BBX32-directed control of acclimation to high light in mature Arabidopsis leaves. *The Plant Journal*. **107**(5), pp.1363–1386.

Anderson, S.J., Kramer, M.C., Gosai, S.J., Yu, X., Vandivier, L.E., Nelson, A.D.L., Anderson, Z.D., Beilstein, M.A., Fray, R.G., Lyons, E. and Gregory, B.D. 2018. N6-Methyladenosine Inhibits Local Ribonucleolytic Cleavage to Stabilize mRNAs in Arabidopsis. *Cell Reports*. **25**(5), pp.1146-1157.e3.

Aristizabal, M.J., Anreiter, I., Halldorsdottir, T., Odgers, C.L., McDade, T.W., Goldenberg, A., Mostafavi, S., Kobor, M.S., Binder, E.B., Sokolowski, M.B. and O'Donnell, K.J. 2020. Biological embedding of experience: A primer on epigenetics. *Proceedings of the National Academy of Sciences*. **117**(38), pp.23261–23269.

Ascenzi, R. and Gantt, J.S. 1997. A drought-stress-inducible histone gene in *Arabidopsis thaliana* is a member of a distinct class of plant linker histone variants. *Plant Molecular Biology*. **34**(4), pp.629–641.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. and Sherlock, G. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*. **25**(1), pp.25–29.

Aufsatz, W., Mette, M.F., van der Winden, J., Matzke, A.J.M. and Matzke, M. 2002. RNA-directed DNA methylation in Arabidopsis. *Proceedings of the National Academy of Sciences*. **99**(Supplement 4), pp.16499–16506.

Baek, D., Jiang, J., Chung, J.-S., Wang, B., Chen, J., Xin, Z. and Shi, H. 2011. Regulated AtHKT1 Gene Expression by a Distal Enhancer Element and DNA Methylation in the Promoter Plays an Important Role in Salt Tolerance. *Plant and Cell Physiology*. **52**(1), pp.149–161.

Balaban, N.Q., Merrin, J., Chait, R., Kowalik, L. and Leibler, S. 2004. Bacterial persistence as a phenotypic switch. *Science*. **305**(5690), pp.1622–1625.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y. and Barkai, N. 2006. Noise in protein expression scales with natural protein abundance. *Nature Genetics*. **38**(6), pp.636–643.

Barroso, G.V., Puzovic, N. and Dutheil, J.Y. 2018. The Evolution of Gene-Specific Transcriptional Noise Is Driven by Selection at the Pathway Level. *Genetics*. **208**(1), pp.173–189.

Bayramoglu, B., Toubiana, D., van Vliet, S., Inglis, R.F., Shnerb, N. and Gillor, O. 2017. Bet-hedging in bacteriocin producing Escherichia coli populations: the single cell perspective. *Scientific Reports*. **7**(1), p.42068.

Bechtold, U., Penfold, C.A., Jenkins, D.J., Legaie, R., Moore, J.D., Lawson, T., Matthews, J.S.A., Vialet-Chabrand, S.R.M., Baxter, L., Subramaniam, S., Hickman, R., Florance, H., Sambles, C., Salmon, D.L., Feil, R., Bowden, L., Hill, C., Baker, N.R., Lunn, J.E., Finkenstädt, B., Mead, A., Buchanan-Wollaston, V., Beynon, J., Rand, D.A., Wild, D.L., Denby, K.J., Ott, S., Smirnoff, N. and Mullineaux, P.M. 2016. Time-Series Transcriptomics Reveals That AGAMOUS-LIKE22 Affects Primary Metabolism and Developmental Processes in Drought-Stressed Arabidopsis. *The Plant Cell*. **28**(2), pp.345–366.

Beelman, C.A. and Parker, R. 1995. Degradation of mRNA in eukaryotes. *Cell*. **81**(2), pp.179–183.

Berardini, T.Z., Mundodi, S., Reiser, L., Huala, E., Garcia-Hernandez, M., Zhang, P., Mueller, L.A., Yoon, J., Doyle, A., Lander, G., Moseyko, N., Yoo, D., Xu, I., Zoeckler, B., Montoya, M., Miller, N., Weems, D. and Rhee, S.Y. 2004. Functional Annotation of the Arabidopsis Genome Using Controlled Vocabularies. *Plant Physiology*. **135**(2), pp.745–755.

Bird, A. 2007. Perceptions of epigenetics. *Nature*. **447**(7143), pp.396–398.

Bishop, A.L., Rab, F.A., Sumner, E.R. and Avery, S.V. 2007. Phenotypic heterogeneity can enhance rare-cell survival in 'stress-sensitive' yeast populations. *Molecular Microbiology*. **63**(2), pp.507–520.

Blake, W.J., KÆrn, M., Cantor, C.R. and Collins, J.J. 2003. Noise in eukaryotic gene expression. *Nature*. **422**(6932), pp.633–637.

Bolger, A.M., Lohse, M. and Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**(15), pp.2114–2120.

Bolstad, B.M. 2004. *Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization*. PhD Thesis, University of California, Berkeley.

Bolstad, B.M., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R.A. and Speed, T.P. 2005. Quality Assessment of Affymetrix GeneChip Data *In*: R. Gentleman, V. Carey, W. Huber, R. Irizarry and S. Dudoit, eds. *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer, pp.33–47.

Bolstad, B.M., Irizarry, R.A., Astrand, M. and Speed, T.P. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics (Oxford, England)*. **19**(2), pp.185–193.

Boyer, J.S. 1982. Plant Productivity and Environment. *Science*. **218**(4571), pp.443–448.

Brettschneider, J., Collin, F., Bolstad, B.M. and Speed, T.P. 2008. Quality assessment for short oligonucleotide microarray data. *Technometrics*. **50**(3), pp.241–264.

Brockmann, R., Beyer, A., Heinisch, J.J. and Wilhelm, T. 2007. Posttranscriptional Expression Regulation: What Determines Translation Rates? *PLOS Computational Biology*. **3**(3), p.e57.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K.O., Di Narzo, A.F., Yan, P., Hodgson, J.G., Weinrich, S., Bosman, F., Roth, A. and Delorenzi, M. 2013. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology*. **231**(1), pp.63–76.

Campbell, D.R., Sosenski, P. and Raguso, R.A. 2019. Phenotypic plasticity of floral volatiles in response to increasing drought stress. *Annals of Botany*. **123**(4), pp.601–610.

Cao, Z. and Grima, R. 2020. Analytical distributions for detailed models of stochastic gene expression in eukaryotic cells. *Proceedings of the National Academy of Sciences*. **117**(9), pp.4682–4692.

Carey, J.N., Mettert, E.L., Roggiani, M., Myers, K.S., Kiley, P.J. and Goulian, M. 2018. Regulated Stochasticity in a Bacterial Signaling Network Permits Tolerance to a Rapid Environmental Change. *Cell*. **173**(1), pp.196-207.e14.

Carlson, K.D., Fernandez-Pozo, N., Bombarely, A., Pisupati, R., Mueller, L.A. and Madlung, A. 2017. Natural variation in stress response gene activity in the allopolyploid Arabidopsis suecica. *BMC Genomics*. **18**(1), p.653.

Carlson, M. 2021. *GO.db: A set of annotation maps describing the entire Gene Ontology*.

Carlson, M. 2015. *TxDb.Athaliana.BioMart.plantsmart22: Annotation package for TxDb object(s)*.

Catoni, M., Griffiths, J., Becker, C., Zabet, N.R., Bayon, C., Dapp, M., Lieberman-Lazarovich, M., Weigel, D. and Paszkowski, J. 2017. DNA sequence properties that predict susceptibility to epiallelic switching. *The EMBO Journal*. **36**(5), pp.617–628.

Catoni, M., Tsang, J.M., Greco, A.P. and Zabet, N.R. 2018. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Research*. **46**(19), pp.e114–e114.

Catoni, M. and Zabet, N.R. 2021. Analysis of Plant DNA Methylation Profiles Using R. *Methods in Molecular Biology (Clifton, N.J.)*. **2250**, pp.219–238.

Cedar, H. and Bergman, Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*. **10**(5), pp.295–304.

Chen, X., Schönberger, B., Menz, J. and Ludewig, U. 2018. Plasticity of DNA methylation and gene expression under zinc deficiency in Arabidopsis roots. *Plant and Cell Physiology*. **59**(9), pp.1790–1802.

Chinnusamy, V., Schumaker, K. and Zhu, J. 2004. Molecular genetic perspectives on cross-talk and specificity in abiotic stress signalling in plants. *Journal of Experimental Botany*. **55**(395), pp.225–236.

Choi, P.J., Cai, L., Frieda, K. and Xie, X.S. 2008. A Stochastic Single-Molecule Event Triggers Phenotype Switching of a Bacterial Cell. *Science*. **322**(5900), pp.442–446.

Conn, V. and Conn, S.J. 2019. SplintQuant: a method for accurately quantifying circular RNA transcript abundance without reverse transcription bias. *RNA*. **25**(9), pp.1202–1210.

Cooke, M.S., Evans, M.D., Dizdaroglu, M. and Lunec, J. 2003. Oxidative DNA damage: mechanisms, mutation, and disease. *The FASEB Journal*. **17**(10), pp.1195–1214.

Cortijo, S., Aydin, Z., Ahnert, S. and Locke, J.C. 2019. Widespread inter-individual gene expression variability in *Arabidopsis thaliana*. *Molecular systems biology*. **15**(1), p.e8591.

Cui, J., shen, N., Lu, Z., Xu, G., Wang, Y. and Jin, B. 2020. Analysis and comprehensive comparison of PacBio and nanopore-based RNA sequencing of the Arabidopsis transcriptome. *Plant Methods*. **16**(1), p.85.

Dahan, O., Gingold, H. and Pilpel, Y. 2011. Regulatory mechanisms and networks couple the different phases of gene expression. *Trends in Genetics*. **27**(8), pp.316–322.

Denkena, J., Johannes, F. and Colomé-Tatché, M. 2021. Region-level epimutation rates in Arabidopsis thaliana. *Heredity*. **127**(2), pp.190–202.

Dietrich, J.-E. and Hiiragi, T. 2007. Stochastic patterning in the mouse pre-implantation embryo. *Development*. **134**(23), pp.4219–4231.

Donald, R.G. and Cashmore, A.R. 1990. Mutation of either G box or I box sequences profoundly affects expression from the Arabidopsis rbcS-1A promoter. *The EMBO Journal*. **9**(6), pp.1717–1726.

Du, J., Johnson, L.M., Groth, M., Feng, S., Hale, C.J., Li, S., Vashisht, A.A., Gallego-Bartolome, J., Wohlschlegel, J.A., Patel, D.J. and Jacobsen, S.E. 2014. Mechanism of DNA Methylation-Directed Histone Methylation by KRYPTONITE. *Molecular Cell*. **55**(3), pp.495–504.

Duveau, F., Hodgins-Davis, A., Metzger, B.P., Yang, B., Tryban, S., Walker, E.A., Lybrook, T. and Wittkopp, P.J. 2018. Fitness effects of altering gene expression noise in Saccharomyces cerevisiae N. Barkai & K. J. Verstrepen, eds. *eLife*. **7**, p.e37272.

Dykhuizen, D.E., Dean, A.M. and Hartl, D.L. 1987. Metabolic Flux and Fitness. *Genetics*. **115**(1), pp.25–31.

Elgart, V., Jia, T. and Kulkarni, R. 2010. Quantifying mRNA Synthesis and Decay Rates Using Small RNAs. *Biophysical Journal*. **98**(12), pp.2780–2784.

Ellis, E.C. and Ramankutty, N. 2008. Putting people in the map: anthropogenic biomes of the world. *Frontiers in Ecology and the Environment*. **6**(8), pp.439–447.

Elowitz, M.B., Levine, A.J., Siggia, E.D. and Swain, P.S. 2002. Stochastic Gene Expression in a Single Cell. *Science*. **297**(5584), pp.1183–1186.

FAO, IFAD, UNICEF, WFP, and WHO 2020. *The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets.* [Online]. Rome: FAO. [Accessed 29 December 2020]. Available from: http://www.fao.org/documents/card/en/c/ca9692en.

Fay, J.C., McCullough, H.L., Sniegowski, P.D. and Eisen, M.B. 2004. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. *Genome Biology*. **5**(4), p.R26.

Finnegan, E.J., Peacock, W.J. and Dennis, E.S. 1996. Reduced DNA methylation in Arabidopsis thaliana results in abnormal plant development. *Proceedings of the National Academy of Sciences*. **93**(16), pp.8449–8454.

Finnegan, J.E. and Dennis, E.S. 1993. Isolation and identification by sequence homology of a putative cytosine methyltransferase from *Arabidopsis thaliana*. *Nucleic acids research*. **21**(10), pp.2383–2388.

Fraser, H.B., Hirsh, A.E., Giaever, G., Kumm, J. and Eisen, M.B. 2004. Noise Minimization in Eukaryotic Gene Expression. *PLOS Biology*. **2**(6), p.e137.

Fray, R.G. and Simpson, G.G. 2015. The Arabidopsis epitranscriptome. *Current Opinion in Plant Biology*. **27**, pp.17–21.

Gao, H., Brandizzi, F., Benning, C. and Larkin, R.M. 2008. A membrane-tethered transcription factor defines a branch of the heat stress response in *Arabidopsis thaliana*. *Proceedings of the National Academy of Sciences*. **105**(42), pp.16398–16403.

Gautier, L., Cope, L., Bolstad, B.M. and Irizarry, R.A. 2004. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*. **20**(3), pp.307–315.

The Gene Ontology Consortium 2019. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Research*. **47**(D1), pp.D330–D338.

Golding, I., Paulsson, J., Zawilski, S.M. and Cox, E.C. 2005. Real-time kinetics of gene activity in individual bacteria. *Cell*. **123**(6), pp.1025–1036.

Golem, S. and Culver, J.N. 2003. Tobacco mosaic virus Induced Alterations in the Gene Expression Profile of *Arabidopsis thaliana*. *Molecular Plant-Microbe Interactions*. **16**(8), pp.681–688.

van der Graaf, A., Wardenaar, R., Neumann, D.A., Taudt, A., Shaw, R.G., Jansen, R.C., Schmitz, R.J., Colomé-Tatché, M. and Johannes, F. 2015. Rate, spectrum, and evolutionary dynamics of spontaneous epimutations. *Proceedings of the National Academy of Sciences*. **112**(21), pp.6676–6681.

Green, E.L. 1941. Genetic and Non-Genetic Factors Which Influence the Type of the Skeleton in an Inbred Strain of Mice. *Genetics*. **26**(2), pp.192–222.

Grote, S. 2020. *GOfuncR: Gene ontology enrichment using FUNC*.

Gygi, S.P., Rochon, Y., Franza, B.R. and Aebersold, R. 1999. Correlation between Protein and mRNA Abundance in Yeast. *Molecular and Cellular Biology*. **19**(3), pp.1720–1730.

Hagai, T., Chen, X., Miragaia, R.J., Rostom, R., Gomes, T., Kunowska, N., Henriksson, J., Park, J.-E., Proserpio, V., Donati, G., Bossini-Castillo, L., Vieira Braga, F.A., Naamati, G., Fletcher, J., Stephenson, E., Vegh, P., Trynka, G., Kondova, I., Dennis, M., Haniffa, M., Nourmohammad, A., Lässig, M. and Teichmann, S.A. 2018. Gene expression variability across cells and species shapes innate immunity. *Nature*. **563**(7730), pp.197–202.

Han, Q., Bartels, A., Cheng, X., Meyer, A., An, Y.-Q.C., Hsieh, T.-F. and Xiao, W. 2019. Epigenetics Regulates Reproductive Development in Plants. *Plants*. **8**(12), p.564.

Hannah, M.A., Heyer, A.G. and Hincha, D.K. 2005. A Global Survey of Gene Regulation during Cold Acclimation in *Arabidopsis thaliana*. *PLOS Genetics*. **1**(2), p.e26.

Hofmeister, B.T., Lee, K., Rohr, N.A., Hall, D.W. and Schmitz, R.J. 2017. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. *Genome Biology*. **18**(1), p.155.

Hong, L., Dumond, M., Zhu, M., Tsugawa, S., Li, C.-B., Boudaoud, A., Hamant, O. and Roeder, A.H.K. 2018. Heterogeneity and Robustness in Plant Morphogenesis: From Cells to Organs. *Annual Review of Plant Biology*. **69**(1), pp.469–495.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*. **37**(1), pp.1–13.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*. **4**(1), pp.44–57.

Inde, Z. and Dixon, S.J. 2018. The Impact of Non-Genetic Heterogeneity on Cancer Cell Death. *Critical reviews in biochemistry and molecular biology*. **53**(1), pp.99–114.

Johannes, F. and Schmitz, R.J. 2019. Spontaneous epimutations in plants. *New Phytologist*. **221**(3), pp.1253–1259.

Johnston, I.G. and Bassel, G.W. 2018. Identification of a bet-hedging network motif generating noise in hormone concentrations and germination propensity in Arabidopsis. *Journal of The Royal Society Interface*. **15**(141), p.20180042.

Kankel, M.W., Ramsey, D.E., Stokes, T.L., Flowers, S.K., Haag, J.R., Jeddeloh, J.A., Riddle, N.C., Verbsky, M.L. and Richards, E.J. 2003. Arabidopsis MET1 Cytosine Methyltransferase Mutants. *Genetics*. **163**(3), pp.1109–1122.

Kato, M., Miura, A., Bender, J., Jacobsen, S.E. and Kakutani, T. 2003. Role of CG and Non-CG Methylation in Immobilization of Transposons in Arabidopsis. *Current Biology*. **13**(5), pp.421–426.

Keren, L., Hausser, J., Lotan-Pompan, M., Vainberg Slutskin, I., Alisar, H., Kaminski, S., Weinberger, A., Alon, U., Milo, R. and Segal, E. 2016. Massively Parallel Interrogation of the Effects of Gene Expression Levels on Fitness. *Cell*. **166**(5), pp.1282-1294.e18.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J. and Harter, K. 2007. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *The Plant Journal*. **50**(2), pp.347–363.

Kimura, H. 2013. Histone modifications for human epigenome analysis. *Journal of Human Genetics*. **58**(7), pp.439–445.

Klepikova, A.V., Logacheva, M.D., Dmitriev, S.E. and Penin, A.A. 2015. RNA-seq analysis of an apical meristem time series reveals a critical point in *Arabidopsis thaliana* flower initiation. *BMC Genomics*. **16**(1), p.466.

Kolodziejczyk, A.A., Kim, J.K., Tsang, J.C.H., Ilicic, T., Henriksson, J., Natarajan, K.N., Tuck, A.C., Gao, X., Bühler, M., Liu, P., Marioni, J.C. and Teichmann, S.A. 2015. Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation. *Cell Stem Cell*. **17**(4), pp.471–485.

Komorowski, M., Miękisz, J. and Stumpf, M.P.H. 2013. Decomposing Noise in Biochemical Signaling Systems Highlights the Role of Protein Degradation. *Biophysical Journal*. **104**(8), pp.1783–1793.

Kotte, O., Volkmer, B., Radzikowski, J.L. and Heinemann, M. 2014. Phenotypic bistability in Escherichia coli's central carbon metabolism. *Molecular Systems Biology*. **10**(7).

Krueger, F. and Andrews, S.R. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics*. **27**(11), pp.1571–1572.

Krzywinski, M. 2020. Designing for Color Blindess. [Accessed 21 September 2021]. Available from: http://mkweb.bcgsc.ca/colorblind/index.mhtml#page-container.

Kunkel, T.A. and Bebenek, K. 2000. DNA replication fidelity. *Annual review of biochemistry*. **69**(1), pp.497–529.

Kussell, E. 2005. Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments. *Science*. **309**(5743), pp.2075–2078.

Lamesch, P., Berardini, T.Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D.L., Garcia-Hernandez, M., Karthikeyan, A.S., Lee, C.H., Nelson, W.D., Ploetz, L., Singh, S., Wensel, A. and Huala, E. 2012. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*. **40**(D1), pp.D1202–D1210.

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

Law, C.W., Chen, Y., Shi, W. and Smyth, G.K. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*. **15**(2), p.R29.

Law, J.A. and Jacobsen, S.E. 2010. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics*. **11**(3), pp.204–220.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T. and Carey, V.J. 2013. Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*. **9**(8), p.e1003118.

Lebrigand, K., Magnone, V., Barbry, P. and Waldmann, R. 2020. High throughput error corrected Nanopore single cell transcriptome sequencing. *Nature Communications*. **11**(1), p.4025.

Lehner, B. 2013. Genotype to phenotype: lessons from model organisms for human genetics. *Nature Reviews Genetics*. **14**(3), pp.168–178.

Lestas, I., Vinnicombe, G. and Paulsson, J. 2010. Fundamental limits on the suppression of molecular fluctuations. *Nature*. **467**(7312), pp.174–178.

Leuendorf, J.E., Frank, M. and Schmülling, T. 2020. Acclimation, priming and memory in the response of *Arabidopsis thaliana* seedlings to cold stress. *Scientific Reports*. **10**(1), p.689.

Levy, S.F., Ziv, N. and Siegal, M.L. 2012. Bet Hedging in Yeast by Heterogeneous, Age-Correlated Expression of a Stress Protectant. *PLOS Biology*. **10**(5), p.e1001325.

Li, H., Ruan, J. and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research*. **18**(11), pp.1851–1858.

Li, W., Notani, D. and Rosenfeld, M.G. 2016. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics*. **17**(4), pp.207–223.

Li, X., Kim, Y., Tsang, E.K., Davis, J.R., Damani, F.N., Chiang, C., Hess, G.T., Zappala, Z., Strober, B.J., Scott, A.J., Li, A., Ganna, A., Bassik, M.C., Merker, J.D., Hall, I.M., Battle, A. and Montgomery, S.B. 2017. The impact of rare variation on gene expression across tissues. *Nature*. **550**(7675), pp.239–243.

Liang, Z., Shen, L., Cui, X., Bao, S., Geng, Y., Yu, G., Liang, F., Xie, S., Lu, T., Gu, X. and Yu, H. 2018. DNA N6-Adenine Methylation in *Arabidopsis thaliana*. *Developmental Cell*. **45**(3), pp.406-416.e3.

Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R. 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in Arabidopsis. *Cell*. **133**(3), pp.523–536.

Littau, V.C., Burdick, C.J., Allfrey, V.G. and Mirsky, S.A. 1965. The role of histones in the maintenance of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America*. **54**(4), pp.1204–1212.

Liu, N., Fromm, M. and Avramova, Z. 2014. H3K27me3 and H3K4me3 Chromatin Environment at Super-Induced Dehydration Stress Memory Genes of *Arabidopsis thaliana*. *Molecular Plant*. **7**(3), pp.502–513.

Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E.G., Van Drogen, A., Borel, C., Frank, M., Germain, P.-L., Bludau, I., Mehnert, M., Seifert, M., Emmenlauer, M., Sorg, I., Bezrukov, F., Bena, F.S., Zhou, H., Dehio, C., Testa, G., Saez-Rodriguez, J., Antonarakis, S.E., Hardt, W.-D. and Aebersold, R. 2019. Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nature Biotechnology*. **37**(3), pp.314–322.

Lun, A.T.L., McCarthy, D.J. and Marioni, J.C. 2016. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Research*. **5**, p.2122.

Luo, G.-Z., MacQueen, A., Zheng, G., Duan, H., Dore, L.C., Lu, Z., Liu, J., Chen, K., Jia, G., Bergelson, J. and He, C. 2014. Unique features of the m 6 A methylome in *Arabidopsis thaliana*. *Nature Communications*. **5**(1), p.5630.

Marí-Ordóñez, A., Marchais, A., Etcheverry, M., Martin, A., Colot, V. and Voinnet, O. 2013. Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics*. **45**(9), pp.1029–1039.

Marondedze, C., Thomas, L., Gehring, C. and Lilley, K.S. 2019. Changes in the Arabidopsis RNA-binding proteome reveal novel stress response mechanisms. *BMC Plant Biology*. **19**(1), p.139.

Mathieu, O., Reinders, J., Čaikovski, M., Smathajitt, C. and Paszkowski, J. 2007. Transgenerational Stability of the Arabidopsis Epigenome Is Coordinated by CG Methylation. *Cell*. **130**(5), pp.851–862.

Matsui, A., Ishida, J., Morosawa, T., Mochizuki, Y., Kaminuma, E., Endo, T.A., Okamoto, M., Nambara, E., Nakajima, M., Kawashima, M., Satou, M., Kim, J.-M., Kobayashi, N., Toyoda, T., Shinozaki, K. and Seki, M. 2008. Arabidopsis Transcriptome Analysis under Drought, Cold, High-Salinity and ABA Treatment Conditions using a Tiling Array. *Plant and Cell Physiology*. **49**(8), pp.1135–1149.

Mattick, J.S. 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*. **25**(10), pp.930–939.

McAdams, H.H. and Arkin, A. 1997. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences*. **94**(3), pp.814–819.

Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., Lewis, S. and Thomas, P.D. 2010. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Research*. **38**(suppl_1), pp.D204–D210.

Miura, T. 2005. Developmental regulation of caste-specific characters in social-insect polyphenism. *Evolution & Development*. **7**(2), pp.122–129.

Morrongiello, J.R., Bond, N.R., Crook, D.A. and Wong, B.B.M. 2012. Spatial variation in egg size and egg number reflects trade-offs and bet-hedging in a freshwater fish. *Journal of Animal Ecology*. **81**(4), pp.806–817.

Mourrain, P., Béclin, C., Elmayan, T., Feuerbach, F., Godon, C., Morel, J.-B., Jouette, D., Lacombe, A.-M., Nikic, S., Picault, N., Rémoué, K., Sanial, M., Vo, T.-A. and Vaucheret, H. 2000. Arabidopsis SGS2 and SGS3 Genes Are Required for Posttranscriptional Gene Silencing and Natural Virus Resistance. *Cell*. **101**(5), pp.533–542.

Nagalakshmi, U., Waern, K. and Snyder, M. 2010. RNA-Seq: A Method for Comprehensive Transcriptome Analysis. *Current Protocols in Molecular Biology*. **89**(1).

Naouar, N., Vandepoele, K., Lammens, T., Casneuf, T., Zeller, G., Hummelen, P.V., Weigel, D., Rätsch, G., Inzé, D., Kuiper, M., Veylder, L.D. and Vuylsteke, M. 2009. Quantitative RNA expression analysis with Affymetrix Tiling 1.0R arrays identifies new E2F target genes. *The Plant Journal*. **57**(1), pp.184–194.

Newman, J., Ghaemmaghami, S., Ihmels, J., Breslow, D., Noble, M., DeRisi, J. and Weissman, J. 2006. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*. **441**, pp.840–6.

Nicolas, D., Phillips, N.E. and Naef, F. 2017. What shapes eukaryotic transcriptional bursting? *Molecular BioSystems*. **13**(7), pp.1280–1290.

Oliver, R.P. and Ipcho, S.V.S. 2004. Arabidopsis pathology breathes new life into the necrotrophs-vs.-biotrophs classification of fungal pathogens. *Molecular Plant Pathology*. **5**(4), pp.347–352.

Ossowski, S., Schneeberger, K., Lucas-Lledo, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D. and Lynch, M. 2010. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. *Science*. **327**(5961), pp.92–94.

Parekh, S., Ziegenhain, C., Vieth, B., Enard, W. and Hellmann, I. 2016. The impact of amplification on differential expression analyses by RNA-seq. *Scientific Reports*. **6**(1), p.25533.

Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D., Hall, A.J., Barton, G.J. and Simpson, G.G. 2020. Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification. *eLife*. **9**, p.e49658.

Pedersen, T.L. 2020. *patchwork: The Composer of Plots* [Online]. Available from: https://CRAN.R-project.org/package=patchwork.

Pedraza, J.M. and van Oudenaarden, A. 2005. Noise Propagation in Gene Networks. *Science, New Series*. **307**(5717,), pp.1965–1969.

Quintero-Cadena, P., Lenstra, T.L. and Sternberg, P.W. 2020. RNA Pol II Length and Disorder Enable Cooperative Scaling of Transcriptional Bursting. *Molecular Cell*. **79**(2), pp.207-220.e8.

R Core Team 2020. *R: A Language and Environment for Statistical Computing* [Online]. Vienna, Austria: R Foundation for Statistical Computing. Available from: https://www.R-project.org/.

Rahbari, R., Wuster, A., Lindsay, S.J., Hardwick, R.J., Alexandrov, L.B., Turki, S.A., Dominiczak, A., Morris, A., Porteous, D., Smith, B., Stratton, M.R. and Hurles, M.E. 2016. Timing, rates and spectra of human germline mutation. *Nature genetics*. **48**(2), pp.126–133.

Rahmstorf, S. and Coumou, D. 2011. Increase of extreme events in a warming world. *Proceedings of the National Academy of Sciences*. **108**(44), pp.17905–17909.

Raj, A. and van Oudenaarden, A. 2008. Stochastic gene expression and its consequences. *Cell*. **135**(2), pp.216–226.

Ranjith-Kumar, C.T., Miller, W., Sun, J., Xiong, J., Santos, J., Yarbrough, I., Lamb, R.J., Mills, J., Duffy, K.E., Hoose, S., Cunningham, M., Holzenburg, A., Mbow, M.L., Sarisky, R.T. and Kao, C.C. 2007. Effects of Single Nucleotide Polymorphisms on Toll-like Receptor 3 Activity and Expression in Cultured Cells*. *Journal of Biological Chemistry*. **282**(24), pp.17696–17705.

Rayirath, P., Benkel, B., Mark Hodges, D., Allan-Wojtas, P., MacKinnon, S., Critchley, A.T. and Prithiviraj, B. 2009. Lipophilic components of the brown seaweed, Ascophyllum nodosum, enhance freezing tolerance in *Arabidopsis thaliana*. *Planta*. **230**(1), pp.135–147.

Rentel, M.C., Lecourieux, D., Ouaked, F., Usher, S.L., Petersen, L., Okamoto, H., Knight, H., Peck, S.C., Grierson, C.S., Hirt, H. and Knight, M.R. 2004. OXI1 kinase is necessary for oxidative burst-mediated signalling in Arabidopsis. *Nature*. **427**(6977), pp.858–861.

Ripley, L.S. 1990. Frameshift Mutation: Determinants of Specificity. *Annual Review of Genetics*. **24**(1), pp.189–211.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W. and Smyth, G.K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*. **43**(7), p.e47.

Roberfroid, S., Vanderleyden, J. and Steenackers, H. 2016. Gene expression variability in clonal populations: Causes and consequences. *Critical Reviews in Microbiology*. **42**(6), pp.969–984.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L. and Pachter, L. 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*. **12**(3), p.R22.

Roeder, R.G. 1996. The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*. **21**(9), pp.327–335.

Salomé, P.A. and McClung, C.R. 2004. The *Arabidopsis thaliana* Clock. *Journal of Biological Rhythms*. **19**(5), pp.425–435.

Santuari, L., Sanchez-Perez, G.F., Luijten, M., Rutjens, B., Terpstra, I., Berke, L., Gorte, M., Prasad, K., Bao, D., Timmermans-Hereijgers, J.L.P.M., Maeo, K., Nakamura, K., Shimotohno, A., Pencik, A., Novak, O., Ljung, K., van Heesch, S., de Bruijn, E., Cuppen, E., Willemsen, V., Mähönen, A.P., Lukowitz, W., Snel, B., de Ridder, D., Scheres, B. and Heidstra, R. 2016. The PLETHORA Gene Regulatory Network Guides Growth and Cell Differentiation in Arabidopsis Roots. *The Plant Cell*. **28**(12), pp.2937–2951.

Sapan, C.V., Lundblad, R.L. and Price, N.C. 1999. Colorimetric protein assay techniques. *Biotechnology and Applied Biochemistry*. **29**(2), pp.99–108.

Saze, H., Mittelsten Scheid, O. and Paszkowski, J. 2003. Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nature genetics*. **34**, pp.65–9.

Schlichting, C.D. 1986. The Evolution of Phenotypic Plasticity in Plants. *Annual Review of Ecology and Systematics*. **17**(1), pp.667–693.

Schmid, M., Davison, T., Henz, S., Pape, U., Demar, M., Vingron, M., Schölkopf, B., Weigel, D. and Lohmann, J. 2005. A gene expression map of *Arabidopsis thaliana* development. *Nature genetics*. **37**, pp.501–6.

Schmiedel, J.M., Carey, L.B. and Lehner, B. 2019. Empirical mean-noise fitness landscapes reveal the fitness impact of gene expression noise. *Nature Communications*. **10**(1), p.3180.

Schmitt, J., Stinchcombe, J.R., Heschel, M.S. and Huber, H. 2003. The Adaptive Evolution of Plasticity: Phytochrome-Mediated Shade Avoidance Responses1. *Integrative and Comparative Biology*. **43**(3), pp.459–469.

Sclep, G., Allemeersch, J., Liechti, R., De Meyer, B., Beynon, J., Bhalerao, R., Moreau, Y., Nietfeld, W., Renou, J.-P., Reymond, P., and others 2007. CATMA, a comprehensive genome-scale resource for silencing and transcript profiling of Arabidopsis genes. *BMC bioinformatics*. **8**(1), pp.1–13.

Serra, L., Arnaud, N., Selka, F., Rechenmann, C., Andrey, P. and Laufs, P. 2018. Heterogeneity and its multiscale integration in plant morphogenesis. *Current opinion in plant biology*. **46**, pp.18–24.

Shahryary, Y., Symeonidi, A., Hazarika, R.R., Denkena, J., Mubeen, T., Hofmeister, B., van Gurp, T., Colomé-Tatché, M., Verhoeven, K.J.F., Tuskan, G., Schmitz, R.J. and Johannes, F. 2020. AlphaBeta: computational inference of epimutation rates and spectra from high-throughput DNA methylation data in plants. *Genome Biology*. **21**(1), p.260.

Shen, L., Liang, Z., Gu, X., Chen, Y., Teo, Z.W.N., Hou, X., Cai, W.M., Dedon, P.C., Liu, L. and Yu, H. 2016. N6-Methyladenosine RNA Modification Regulates Shoot Stem Cell Fate in Arabidopsis. *Developmental Cell*. **38**(2), pp.186–200.

Siciliano, V., Garzilli, I., Fracassi, C., Criscuolo, S., Ventre, S. and di Bernardo, D. 2013. miRNAs confer phenotypic robustness to gene networks by suppressing biological noise. *Nature Communications*. **4**(1), p.2364.

Simons, A.M. and Johnston, M.O. 2006. Environmental and Genetic Sources of Diversification in the Timing of Seed Germination: Implications for the Evolution of Bet Hedging. *Evolution*. **60**(11), pp.2280–2292.

Singh, A. and Hespanha, J.P. 2009. Optimal Feedback Strength for Noise Suppression in Autoregulatory Gene Networks. *Biophysical Journal*. **96**(10), pp.4013–4023.

Skupsky, R., Burnett, J.C., Foley, J.E., Schaffer, D.V. and Arkin, A.P. 2010. HIV Promoter Integration Site Primarily Modulates Transcriptional Burst Size Rather Than Frequency. *PLOS Computational Biology*. **6**(9), p.e1000952.

Smith, Z.D., Gu, H., Bock, C., Gnirke, A. and Meissner, A. 2009. High-throughput bisulfite sequencing in mammalian genomes. *Methods (San Diego, Calif.)*. **48**(3), pp.226–232.

Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S.W., Tavaré, S., Deloukas, P., Hurles, M.E. and Dermitzakis, E.T. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science (New York, N.Y.)*. **315**(5813), pp.848–853.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D.J. and Jacobsen, S.E. 2014. Non-CG methylation patterns shape the epigenetic landscape in Arabidopsis. *Nature Structural & Molecular Biology*. **21**(1), pp.64–72.

Stroud, H., Greenberg, M.V.C., Feng, S., Bernatavichute, Y.V. and Jacobsen, S.E. 2013. Comprehensive analysis of silencing mutants reveals complex regulation of the Arabidopsis methylome. *Cell*. **152**(1–2), pp.352–364.

Stroud, H., Hale, C.J., Feng, S., Caro, E., Jacob, Y., Michaels, S.D. and Jacobsen, S.E. 2012. DNA methyltransferases are required to induce heterochromatic re-replication in Arabidopsis. *PLoS genetics*. **8**(7), p.e1002808.

Sun, S., Zhu, J., Ma, Y. and Zhou, X. 2019. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. *Genome Biology*. **20**(1), p.269.

Thirumalaikumar, V.P., Gorka, M., Schulz, K., Masclaux-Daubresse, C., Sampathkumar, A., Skirycz, A., Vierstra, R.D. and Balazadeh, S. 2020. Selective autophagy regulates heat stress memory in *Arabidopsis* by NBR1-mediated targeting of HSP90 and ROF1. *Autophagy*., pp.1–16.

Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A. and Narechania, A. 2003. PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*. **13**(9), pp.2129–2141.

Tollervey, J.R. and Lunyak, V.V. 2012. Epigenetics: Judge, jury and executioner of stem cell fate. *Epigenetics*. **7**(8), pp.823–840.

Trontin, C., Tisné, S., Bach, L. and Loudet, O. 2011. What does Arabidopsis natural variation teach us (and does not teach us) about adaptation in plants? *Current Opinion in Plant Biology*. **14**(3), pp.225–231.

Tsien, R.Y. 1998. The Green Fluorescent Protein. *Annual Review of Biochemistry*. **67**(1), pp.509–544.

Tsugeki, R., Kochieva, E.Z. and Fedoroff, N.V. 1996. A transposon insertion in the Arabidopsis SSR16 gene causes an embryo-defective lethal mutation. *The Plant Journal*. **10**(3), pp.479–489.

Tunnacliffe, E. and Chubb, J.R. 2020. What Is a Transcriptional Burst? *Trends in Genetics*. **36**(4), pp.288–297.

United Nations Department of Economic and Social Affairs, Population Division 2019. *World population prospects Highlights, 2019 revision Highlights, 2019 revision*.

Uphoff, S., Lord, N.D., Okumus, B., Potvin-Trottier, L., Sherratt, D.J. and Paulsson, J. 2016. Stochastic activation of a DNA damage response causes cell-to-cell mutation rate variation. *Science*. **351**(6277), pp.1094–1097.

Vallejo, A.J., Yanovsky, M.J. and Botto, J.F. 2010. Germination variation in *Arabidopsis thaliana* accessions under moderate osmotic and salt stresses. *Annals of Botany*. **106**(5), pp.833–842.

Veening, J.-W., Stewart, E.J., Berngruber, T.W., Taddei, F., Kuipers, O.P. and Hamoen, L.W. 2008. Bet-hedging and epigenetic inheritance in bacterial cell development. *Proceedings of the National Academy of Sciences*. **105**(11), pp.4393–4398.

Veneziano, D., Nigita, G. and Ferro, A. 2015. Computational Approaches for the Analysis of ncRNA through Deep Sequencing Techniques. *Frontiers in Bioengineering and Biotechnology*. **3**.

Vidal, R., Frangione, B., Rostagno, A., Mead, S., Révész, T., Plant, G. and Ghiso, J. 1999. A stop-codon mutation in the BRI gene associated with familial British dementia. *Nature*. **399**(6738), pp.776–781.

Voliotis, M. and Bowsher, C.G. 2012. The magnitude and colour of noise in genetic negative feedback systems. *Nucleic Acids Research*. **40**(15), pp.7084–7095.

Wang, C., Kim, T., Gao, D., Vaglenov, A. and Kaltenboeck, B. 2007. Rapid high-yield mRNA extraction for reverse-transcription PCR. *Journal of biochemical and biophysical methods*. **70**(3), pp.507–509.

Wang, J., Tian, L., Lee, H.-S., Wei, N.E., Jiang, H., Watson, B., Madlung, A., Osborn, T.C., Doerge, R.W., Comai, L. and Chen, Z.J. 2006. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics*. **172**(1), pp.507–517.

Wang, W., Vinocur, B. and Altman, A. 2003. Plant responses to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance. *Planta*. **218**(1), pp.1–14.

Weinberger, L.S., Burnett, J.C., Toettcher, J.E., Arkin, A.P. and Schaffer, D.V. 2005. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell*. **122**(2), pp.169–182.

Wickham, H. 2016. *ggplot2: Elegant Graphics for Data Analysis* [Online]. Springer-Verlag New York. Available from: https://ggplot2.tidyverse.org.

Widdicombe, W.D. and Thelen, K.D. 2002. Row Width and Plant Density Effects on Corn Grain Production in the Northern Corn Belt. *Agronomy Journal*. **94**(5), pp.1020–1023.

Yan, H., Liu, Y., Zhang, K., Song, J., Xu, W. and Su, Z. 2019. Chromatin State-Based Analysis of Epigenetic H3K4me3 Marks of Arabidopsis in Response to Dark Stress. *Frontiers in Genetics*. **10**.

Yang, H., Chang, F., You, C., Cui, J., Zhu, G., Wang, L., Zheng, Y., Qi, J. and Ma, H. 2015. Whole-genome DNA methylation patterns and complex associations with gene structure and expression during flower development in Arabidopsis. *The Plant Journal*. **81**(2), pp.268–281.

Zabet, N.R., Catoni, M., Prischi, F. and Paszkowski, J. 2017. Cytosine methylation at CpCpG sites triggers accumulation of non-CpG methylation in gene bodies. *Nucleic Acids Research*. **45**(7), pp.3777–3784.

Zhang, K., Sridhar, V.V., Zhu, J., Kapoor, A. and Zhu, J.-K. 2007. Distinctive Core Histone Post-Translational Modification Patterns in *Arabidopsis thaliana*. *PLOS ONE*. **2**(11), p.e1210.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S.W.-L., Chen, H., Henderson, I.R., Shinn, P., Pellegrini, M., Jacobsen, S.E. and Ecker, J.R. 2006. Genome-wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis. *Cell*. **126**(6), pp.1189–1201.

Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. and Liu, X. 2014. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLOS ONE*. **9**(1), p.e78644.

Zhong, Z., Feng, S., Duttke, S.H., Potok, M.E., Zhang, Y., Gallego-Bartolomé, J., Liu, W. and Jacobsen, S.E. 2021. DNA methylation-linked chromatin accessibility affects genomic architecture in *Arabidopsis*. *Proceedings of the National Academy of Sciences*. **118**(5), p.e2023347118.

Zilberman, D., Gehring, M., Tran, R., Ballinger, T. and Henikoff, S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. *Nature genetics*. **39**, pp.61–69.