Chapter 1

Chapter 5: Global effect of G Quadruplex stabilisation on gene expression:

Introduction:

As was shown in Chapter 4, the distribution of PG4s around and within genic regions is not uniform. In the *Arabidopsis thaliana* genome, template stranded PG4s are enriched in the 5' UTRs and promoter proximal regions, whilst coding strand PG4s are enriched in 3' UTRs and promoter distal regions. These enrichments do not appear to be explained simply by the GC content of these sequences, nor by the requirement to code for particular amino acids. These features may be deliberately conserved, indicating a biological function for PG4s within gene bodies. As G4s are formed from single

stranded DNA, it has been previously hypothesised that coding strand G4s might function to promote transcription by competing with double stranded DNA to produce regions of open chromatin which could easily become transcription bubbles (Fig 1.1a, Rhodes & Lipps 2015). G4s in the coding strand also have an opportunity to form in mRNA and regulate stability, splicing or translation. Template stranded G4s which occur downstream of translocating RNA Polymerase II (Pol II), on the other hand, might cause blockages which prevent elongation, causing downregulation of the gene (Fig. 1.1b, Rhodes & Lipps 2015).

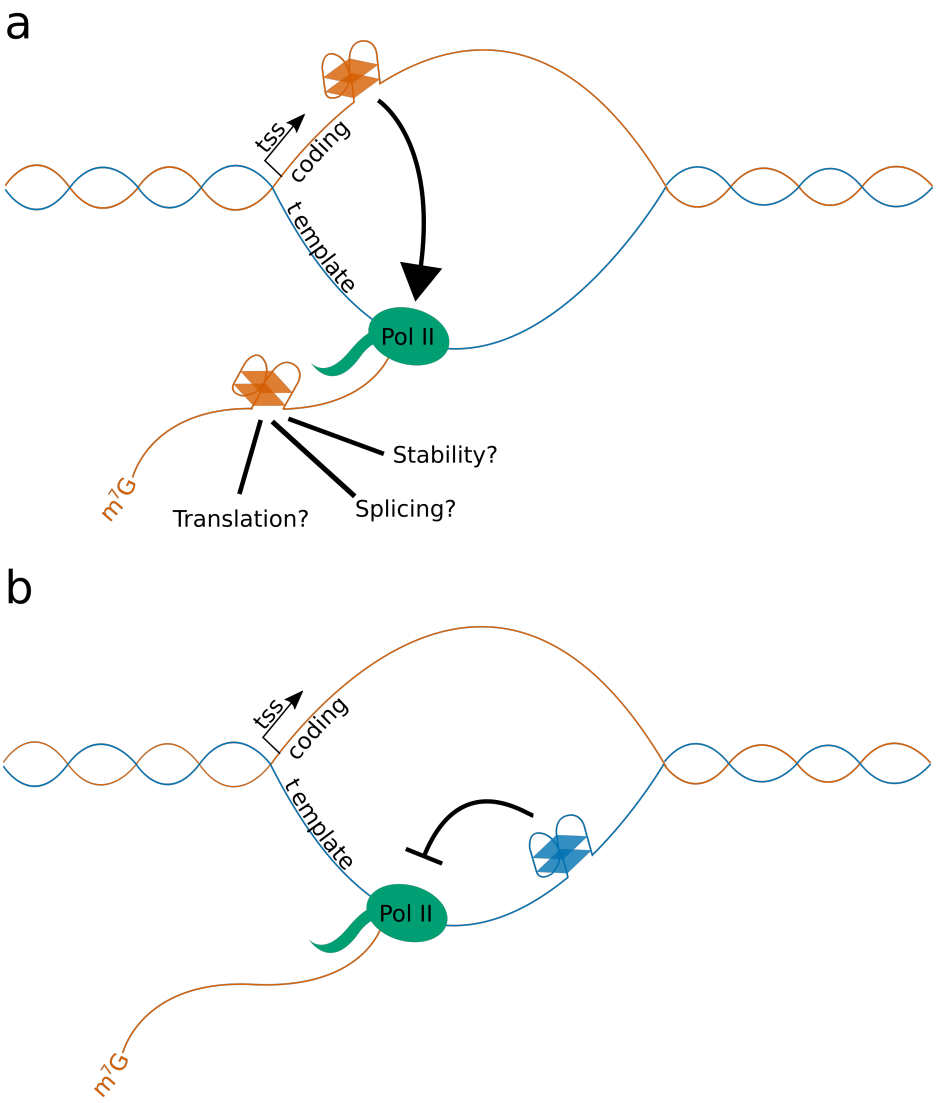

Figure 1.1: **Possible Mechanisms for DNA and RNA G4 function in transcription and gene expression.** Adapted from Figure 4. Rhodes & Lipps 2015. **a)** Possible mechanism for function of G4s located in the coding strand. Since G4s form from single stranded DNA, G4s in the coding strand may promote melting of double stranded DNA, increasing transcription levels. G4s which form in the coding strand of the exonic DNA of a gene will also be present in the mRNA produced from that locus, and G4s which form in the coding strand of introns will be present in pre-mRNA. These RNA G4s might also influence gene expression through alteration of pre-mRNA splicing, mRNA stability, or translation. **b)** Possible mechanism for the function of G4s located in the template strand. The template strand of genes is scanned by RNA Polymerase II during transcription, and G4s which form in ahead of the transcription complex may cause slowing or stalling if they cannot be correctly resolved. RNA Polymerase II translocation speed is linked to a number of co-transcriptional processes, including splicing.

Arguably the best *in vivo* evidence for G4 formation was conducted using an antibody raised against G4 structures, referred to as BG4. BG4 was used to visualise G4s in human cancer cells by immunofluorescence (Biffi et al. 2013). G4 foci were identified at both the telomeres (which are highly GC rich and have been shown to form G4s *in vitro*), and in interstitial regions which contain actively transcribed euchromatin. The G4 density of the cells was also seen to fluctuate throughout the cell cycle, with the greatest number of foci appearing during S phase, when the DNA is decondensed to allow replication to occur.

The BG4 antibody has been futher used to conduct ChIPseq experiments in human cell lines, in which the DNA fragments to which the antibody binds were enriched and subsequently sequenced (Hänsel-Hertsch et al. 2016). BG4 ChIPseq peaks were found to overlap with regions of open chromatin which are sensitive to DNAse digestion. These regions are commonly found around the transcriptional start sites of genes and are associated with actively transcription or promoter proximal pausing. Interestingly, genes in which BG4 peaks strengthened after treatment with HDAC inhibitors (which cause relaxation of heterochromatin) also saw a corresponding increase in gene expression, suggesting that promoter G4 formation may have a positive impact upon gene expression.

Another method by which the effect of G4s on transcription has been studied is through stabilisation with G4 binding ligands. These ligands generally bind to G4s through external hydrophobic pi-stacking above a planar tetrad, or through intercalation between the inner faces of tetrads. Treatment of

yeast species *Saccharomyces cerevisiae* with the G4 stacking ligand N-methyl-mesoporphyrin was shown to upregulate the expression of genes containing coding strand G4s in their promoters (Hershman et al. 2008). Rodriguez et al. showed that treatment of human cells with Pyridostatin caused DNA damage at PG4 containing regions in gene bodies, caused by both replication dependent and transcription dependent damage. These genes were also down-regulated in expression, suggesting that stabilised G4s caused arrest of Pol II. G4s have also been shown to cause pausing of DNA and RNA polymerases *in vitro*, in the presence and absence of G4 binding ligands (ref).

Here we conduct a global study of G4 stabilisation in the model plant *Arabidopsis thaliana*, using the G4 binding ligand NMM. Previous studies have shown that treatment of Arabidopsis seedlings with NMM cause developmental defects (Nakagawa et al. 2012), suggesting an effect on gene expression. We also identify alterations in Pol II occupancy potentially caused by G4 dense transcribed regions of genes. Interestingly, Mullen et al. showed that genes involved in response to drought stress tended to be more likely to contain PG4s in their gene bodies (Mullen et al. 2012). Since G4 formation is dependent upon potassium cation concentration, the intracellular concentration of which increases during drought stress, this could constitute a regulatory mechanism of G4s. We also investigate the overlap of NMM regulated genes with drought stress responsive genes.

Methods:

Plant Growth Conditions:

For N-Methyl Mesoporphyrin (NMM) (Frontier Scientific, NMM580) treatment microarray experiments, the *Arabidopsis thaliana* Columbia (Col-0) ecotype was used. Seeds were surface sterilised, stratified for 2-3 days at 4°C and sown on vertical plates containing Murashige & Skoog (MS) agar with 0.5% sucrose and 0.8% agar. Plants were then transferred to growth cabinets with 16 hours light at 23°C. Seedlings used for expression analysis by qRT-PCR were grown for 7 days on MS plates, treated for 6 hours by flooding the plate with MS liquid media containing NMM after which roots and shoots are harvested separately.

RNA Extraction & Microarray data generation:

Total nucleic acid isolation protocol was carried out by phenol-chloroform extraction as described by White and Kaper (1989). The resulting pellets were resuspended in sterile water and stored at -80C. The RNA concentration and quality was checked using the NanoDrop 1000 Spectrophotometer (ThermoScientific).

cDNA library preparation and microarray analysis were performed by the Genomics Core Facility in the Sheffield Institute for Translational Neuroscience. An Arabidopsis Gene 1.0 ST array was used. RNA integrity and abundance was measured using an Aligent Bioanalyser 2100. Hybridization and scan-

ning procedures were conducted according to the manufacturer using the Affymetrix Gene Chip hybridisation system.

Microarray Analysis:

NMM Microarray analysis was conducted in R using the packages `oligo` and `puma`. `oligo/puma` was chosen over `oligo/limma` for this analysis as NMM treatment appeared to cause large consistent changes in gene expression which violated the assumptions used in Robust Multi-chip Averaging (RMA), namely that most genes do not change there expression and that there is no correlation between average expression and log fold change. CEL files were read into R using `oligo` but were not normalised using RMA. The `puma` baysian probabilistic method was used to normalise data and conduct differential expression analysis. `puma` Probability of Positive Log Ratio (PPLR) values were calculated for each contrast. Strongly differentially expressed genes were produced using an absolute log2 fold change threshold of 1 and a PPLR of 0.05 (or 0.95 for positively differentially expressed genes), and moderately differentially expressed genes were produced using an absolute log2 fold change threshold of 0.5 and a PPLR of 0.05. Annotation of microarray data was conducted using the `oligo getNetAffx` function and Ensembl annotations were extracted.

Analysis of previously published microarray data:

Processed Berberine expression data was downloaded from the supplementary material of Nakagawa et al. 2012. Data was generated using an Affymetrix ATH1 GeneChip array from Col-0 plants grown on MS media containing 12.5 M Berberine for 14 days.

Drought stress microarrays were downloaded from Linster et al. 2015 (GSE65414). Drought stressed plants were grown on soil for 6 weeks under with 8 hour light period, with normal watering, followed by 10 days drought stress. Data was generated using the Affymetrix Gene 1.1 ST Array.

Raw drought stress microarray data was processed in R using `oligo` and `limma`. CEL files were read into R using `oligo` and quantile normalised & median polished using Robust Multi-chip averaging. Linear modelling was then performed using `limma`, and p values were adjusted for multiple testing using Benjamini Hochberg correction. Strongly differentially expressed genes were produced using an absolute log2 fold change threshold of 1 and a FDR of 0.05, and moderately differentially expressed genes were produced using an absolute log2 fold change threshold of 0.5 and a FDR of 0.05. Annotation of microarray data was conducted using the `oligo` `getNetAffx` function and Ensembl annotations were extracted.

Genome and Annotations used:

All analyses were performed using the TAIR10 genome, downloaded in fasta format from arabidopsis.org, and the Araport11 genome annotation, down-

loaded in GTF format from araport.org. Annotations were filtered to obtain protein coding genes only using the CGAT `gtf2gtf` script. To obtain sets of genic features such as exons, introns, CDS and UTRs, CGAT `gtf2gtf` was also used. Overlapping exons from different isoforms of the same gene were flattened to produce non-overlapping exons, and bed files of exons, CDS, 5' UTRs and 3' UTRs were generated from these flattened exons using `awk`. Bed files of introns were created using CGAT `gtf2gtf` to generate exon "complementation". Bed files of whole gene bodies were generated using CGAT `gtf2gtf` to merge all intervals into a single interval spanning the entire gene.

PG4 prediction:

G Quadruplex predictions in the TAIR10 genome were carried out using an in-house script (g4predict) which utilises the Quadparser method (Huppert & Balasubramanian 2005). Results were filtered using a dynamic programming approach, commonly used in interval scheduling, to produce the greatest number of non-overlapping PG4s. Scripts can be found on GitHub at https://github.com/mparker2/g4predict. To count G4s per gene, the bed files containing PG4s were overlapped with bed files generated from Araport11 for exon, intron, CDS, 5' UTR, 3' UTR and full gene bodies. This was done using `bedtools intersect` in count mode. PG4s on the template and coding strands of gene features were counted separately. For multi-exon genes, counts for different exons were summed using `awk` scripts, and counts were normalised by length to get PG4 densities per kb. Barplots of average PG4 density for various gene features and gene sets were produced in python

using `pandas` and `seaborn`. Errorbars for these plots are estimated 68% confidence intervals generated using 1000 bootstrapped samples. Statistical hypothesis testing was done using the Mann-Whitney U test.

Maximal PG4 densities were calculated using a sliding window of 200bp generated using `bedtools makewindows`, with a step size of 5bp. `bedtools map` was used to count the number of PG4s overlapping each window. The score of the maximum scoring 200bp window overlapping the transcribed body (exons and introns) of a gene was assigned as the maximal PG4 density of the gene, using `bedtools map`. Coding and template strand densities were calculated separately. Pointplots of average expression change during NMM treatment for genesets with different maximal PG4 densities were produced in python using `pandas` and `seaborn`. Errorbars for these plots are estimated 68% confidence intervals generated using 1000 bootstrapped samples. Statistical hypothesis testing was done using the Mann-Whitney U test.

For analyses where G4seeqer was used, PG4 predictions were conducted on the TAIR10 genome using the G4seeqer command line tool. A step size of 5bp and G4Hunter threshold of 0.75 were used. All intervals tested using G4seeqer were output regardless of neural network score (i.e. a threshold of 0 was used). Maximum G4seeqer score overlapping each gene body (exons and introns) was calculated using `bedtools map`. Coding and template strand scores were calculated separately.

Self Organising Map Analysis:

Loop lengths and total loop lengths for each Quadparser 2 tetrad PG4 in the TAIR10 genome were extracted from bed files output by g4predict. PG4s which did not overlap with gene bodies were discarded. Self Organising Maps were trained in R using the package `kohonen` on loop length data. 36 clusters were used. To identify enrichment of specific clusters of PG4s in NMM downregulated genes, the total number of PG4s from each cluster overlapping the geneset was calculated, and compared to an expected number of overlaps computed by permuting PG4s amongst all genes. Genes were weighted by length such that a 2kb gene was twice as likely to be assigned PG4s as a 1kb gene. Coding and template strand PG4s were permuted separately. The log fold change between observed and expected overlap was then calculated for each cluster. SOM plots were made in Python using `matplotlib`.

Venn diagrams:

Venn diagrams of geneset overlaps were produced in Python using the package `matplotlib_venn`. Statistical hypothesis testing of overlaps was conducted using hypergeometric tests.

Pol II ChIP-tiling array analysis:

RNA Polymerase II ChIP-tiling array data was downloaded from Chodavarapu et al. 2012 (GSE21673). Plants were grown under 24hr light on soil for 10-14 days before being harvested. ChIP was conducted using Abcam

ab817 Pol II antibody and tiling arrays used were Affymetrix Arabidopsis Tiling 1.0R Array. Pol II occupancy tracks were generated from CEL files in R using `STARR`. Cyclic loess method was used for probe intensity normalisation. The enrichment ratio of PolII signal intensity over control was calculated and saved in BigWig format using `rtracklayer`. Metagene profiles for all genes were produced using CGAT `bam2geneprofile`. Gene profiles of merged exons (without introns) were produced using 100 bins across the gene body, with an upstream and downstream extension of 500bp at 10bp resolution (i.e. binned in 10bp intervals).

To compared Pol II occupancy of G4 containing genes with non-G4 containing genes, genesets with max G4seeqer scores greater than 0.95 and less than 0.05, or maximal PG4 density greater than 2 or equal to zero were used. Metagene profile matrices were read into Python using `pandas` and averaged profiles for each geneset were generated using `seaborn` bootstrapping to estimate central tendency and confidence intervals. Bootstrapped profiles were smoothed using a moving average of 20. 1000 iterations were used for all bootstraps. Profiles were normalised so that the absolute area under the curve was equal to one.

GRO/RNA seq analysis:

Global Run On (GRO) and RNA sequencing data from Hetzel et al. 2016 (GSE83108) was downloaded from the European Nucleotide Archive (ENA). Quality control analyses were performed using `FastQC` and `fastq-screen`. Mapping to the TAIR10 genome with splice junction annotations from Ara-

port11 was conducted using `STAR` with default parameters, and generated BAM files were sorted using `samtools`. Exonic read counts per gene were then counted using `featureCounts`. Read counts were normalised for library depth in R using `DESeq2` and log2 transformed to get log counts per million (logCPM). The ratio of GROseq to RNAseq reads was calculated by subtracting the average RNAseq logCPM from the average GROseq logCPM. Scatter plots of RNAseq logCPM vs GROseq logCPM were generated in Python using `seaborn`.

To contrast GRO/RNA seq ratios of G4 containing genes with non-G4 containing genes, genesets with max G4seeqer scores greater than 0.95 and less than 0.05, or maximal PG4 density greater than 2 or equal to zero were used. Overlayed histograms and kernel density estimate plots of GRO/RNA seq ratio for these genesets were generated using `seaborn`. Statistical hypothesis testing was conducted using Welch's unpaired T-test.

Results:

NMM causes global change in gene expression:

In order to test the effect of G4 stabilisation on gene expression in Arabidopsis, we conducted a microarray analysis using RNA from 7 day old seedlings treated with NMM for 6 hours. Control samples were treated with DMSO for 6 hours. We found 858 and 1098 genes were differentially upregulated and downregulated respectively using a log fold change threshold of 1 (PPLR < 0.05). When a less stringent log fold change threshold of 0.5 was used (PPLR < 0.05), downregulated genes outnumbered upregulated by a ratio of 2:1 (3882 downregulated, 1930 upregulated), suggesting that NMM has a global effect on gene expression unlikely to be caused by a single transcription factor. An MA plot of average gene expression vs log fold change, with lowess curve fitting, showed that there appeared to be a slight skew towards downregulation in genes with higher average expression (Fig 1.2a). To more clearly visualise this skew we binned genes by average expression quartile (Fig 1.2b). This showed there was indeed a relationship between average expression and expression change upon NMM treatment, with highly expressed genes tending to be more downregulated by NMM treatment. This global pattern, which violates some of the assumptions that are usually used in microarray normalisation and analysis, suggests a widespread effect of NMM directly upon either transcription or mRNA stability.

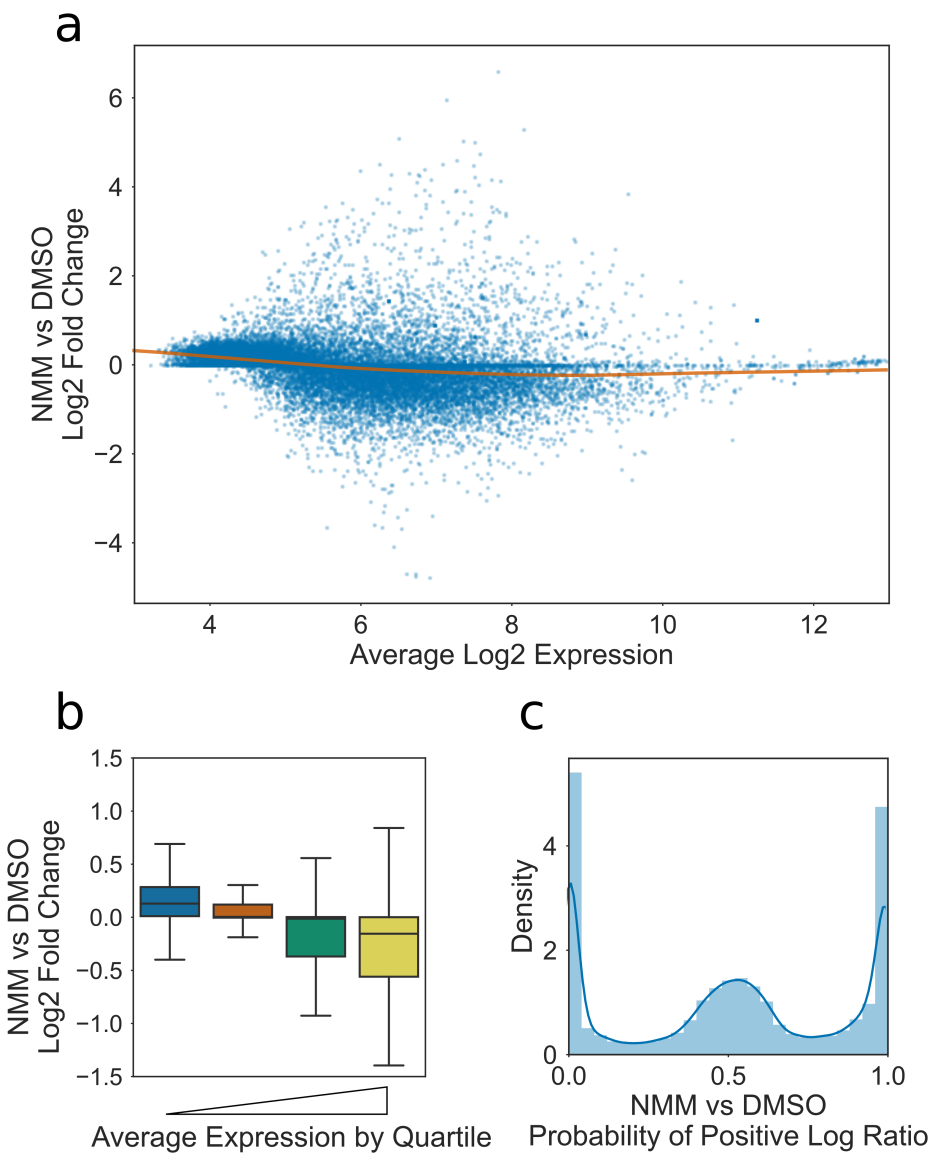

Figure 1.2: **Global effect of NMM on Gene Expression. a)** MA plot showing relationship between average gene expression and Log2 fold change in expression upon treatment with NMM. Orange line is lowess curve fit showing slight negative correlation between expression and fold change for genes in the expression range 4-8. **b)** Boxplot of Log2 fold change in expression upon treatment with NMM, cut on quartiles by average expression. Lowest expressed 25% is leftmost, and highest expressed is rightmost. Lower quartile, median, and upper quartile are at 4.6, 5.4, and 6.5, respectively. **c)** Histogram and kernel density estimate showing distribution of Probability of Positive Log Ratio (PPLR) values. PPLRs which tend towards zero represent negatively differentially expressed genes, whilst PPLRs which tend towards one represent positively differentially expressed genes.

To support the hypothesis that NMM alters gene expression through G4 stabilisation, we correlated our results with processed data from a Berberine treatment array (Fig 1.3a) (Nakagawa et al. 2012). Berberine is another G4 stabilising drug, but with a very different structure and method of action (intercalation with G4s rather than hydrophobic stacking). Despite the differences in structure of the two drugs, and the very different conditions (plants were grown on Berberine for 14 days, compared with 6 hour treatment of 7 day old seedlings with NMM), the log fold changes from our data correlated well with the Berberine dataset (Pearsons R: 0.43, Spearman's : 0.44). There was a strong overlap between the genes downregulated by NMM and those downregulated by Berberine (Fig 1.3b, p=1.1e-36). These results suggest that the main effects on gene expression were through G4 interaction, with any off target effects being less significant contributors.

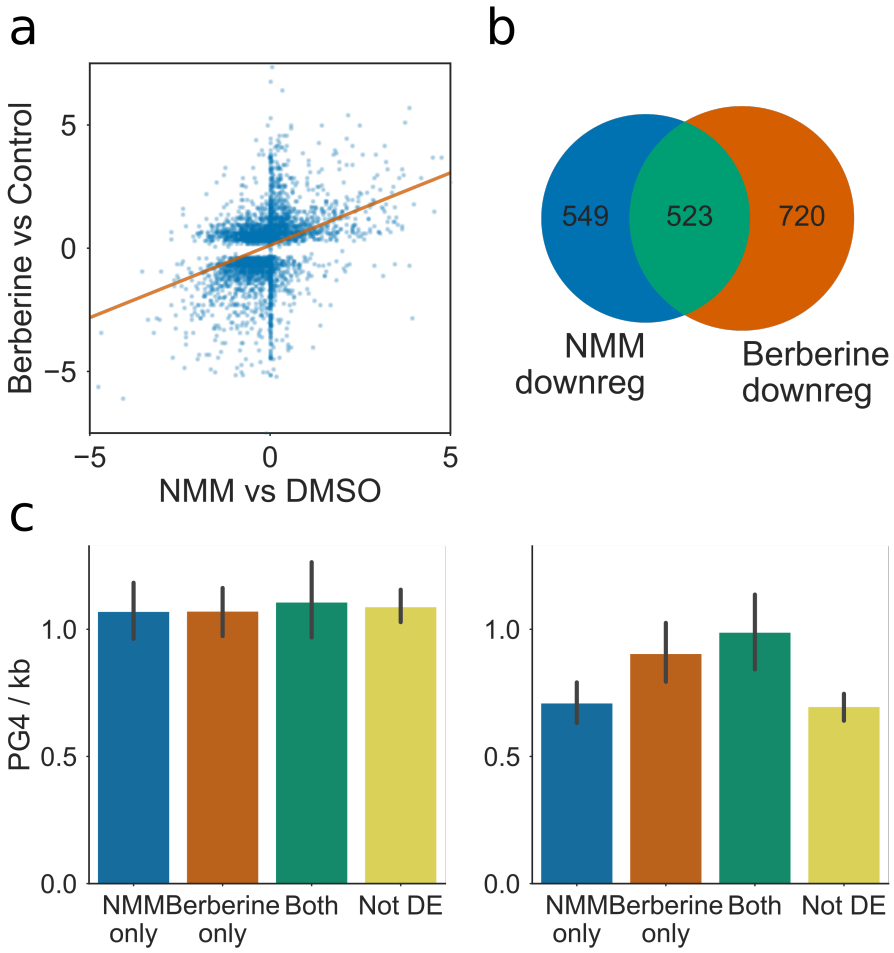

Figure 1.3: **Comparison of gene expression during NMM treatment with expression during Berberine treatment. a)** Scatter plot with regression line showing the correlation in expression change for NMM vs DMSO and Berberine vs Control. Processed Berberine data was taken from supplementary information of Nakagawa et al. 2012, however only differentially regulated genes were reported. **b)** Venn diagram reporting the overlap of genes downregulated by NMM with those downregulated by Berberine. **c)** Bar plot showing the average exonic PG4 densities of NMM and Berberine downregulated genesets, on the coding and template strands, respectively. Both genesets show an greater exonic G4 density on the template strand than genes not regulated by either drug, however genes which are regulated by both drugs had the greatest average exonic PG4. Bar colours match set colours from Fig 3b. Errorbars are 68% confidence intervals for mean generated using 1000 bootstrapped samples.

Genes downregulated by NMM are enriched in two tetrad PG4s:

We next investigated whether NMM regulated genes were enriched for PG4s. We first measured the density of PG4s in different regions of each gene (i.e. promoters, exons, introns, CDS and UTRs) and looked at the differences between the differentially expressed gene sets at 6 hours NMM treatment. We did not see a strong difference in three tetrad PG4 density in our gene sets, for any gene feature (data not shown). This is likely due to the low density of these PG4s in Arabidopsis. We discovered a striking enrichment of the 2 tetrad PG4s on the template strand of genes which were down-regulated by NMM, with approximately 50% more genes containing PG4s than expected for a gene set of that size. This enrichment occured most specifically in the CDS and 5' UTR regions of genes (Fig 1.4). Genes which were very strongly downregulated by NMM (logFC $<$ -1) tended to contain large numbers of PG4s throughout their exonic bodies, particularly in coding regions and in the 5' UTR, whilst moderately downregulated genes (logFC $<$ -0.5) tended to have greater concentration of PG4s in 5' UTRs.

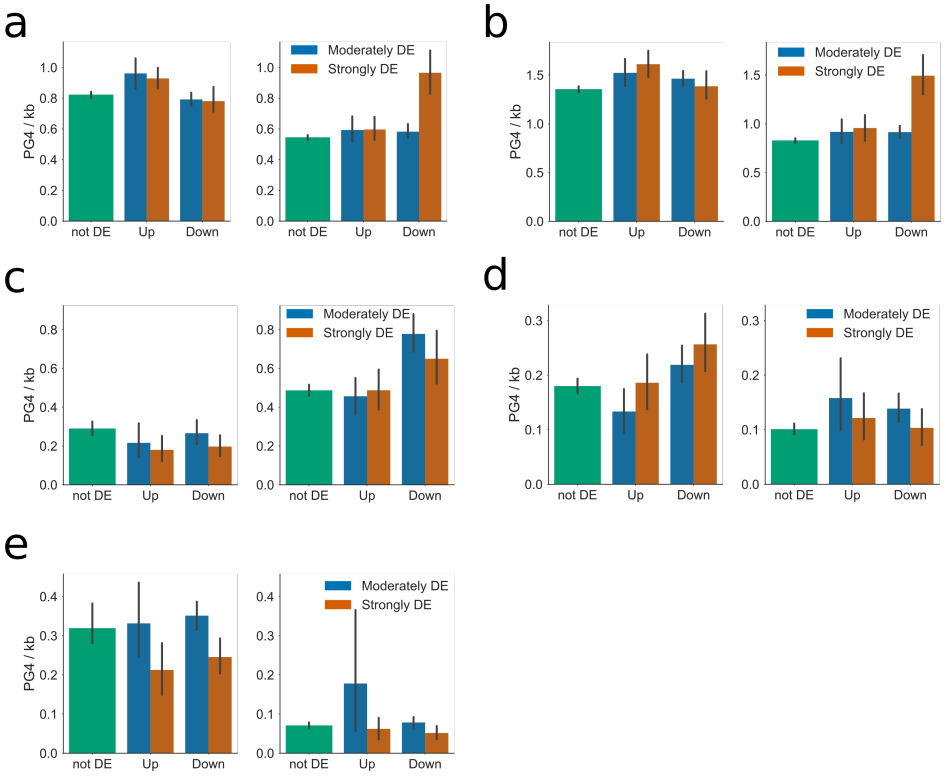

Figure 1.4: **Distribution of PG4s in genes differentially regulated by NMM.** Bar plots showing the average PG4 densities of genes up or downregulated by NMM, for **a)** full gene body (exons and introns), **b)** coding regions, **c)** 5' UTR, **d)** 3' UTR, and **e)** introns, respectively. In each figure, left and right panels represent coding and template strand, respectively. Genesets are separated into three categories by strength of regulation: green: not differentially expressed, blue: moderately differentially expressed (PPLR < 0.05, logFC > 0.5), orange: strongly differentially expressed (PPLR < 0.05, logFC > 1). Errorbars are 68% confidence intervals for mean generated using 1000 bootstrapped samples. Genes which are strongly downregulated by NMM tend to have higher PG4 densities on the template strand of coding regions and 5' UTRs, whilst moderately downregulated genes tend to have greater PG4 density on the template strand of their 5' UTRs.

The PG4 density of genes downregulated by both NMM and Berberine was also calculated (Fig 1.3c). We found that whilst the gene sets downregulated by either treatment were enriched in PG4s, those which were in the intersection of the two sets had the greatest average exonic PG4 density. This is further evidence that these drugs are regulating gene expression through G4 stabilisation.

Previous studies have shown that NMM is highly selective towards parallel G4s (Kreig et al. ..), and can induce anti-parallel G4s to switch structure (Nicoludis et al. 2012). G4s with short loop lengths are more likely to form parallel structures (ref). In order to test if NMM was selective towards G4s with particular loop lengths, we used Self Organising Maps to cluster all predicted Arabidopsis 2 tetrad PG4s into 36 groups, based on the length of each loop 1-3, and the total loop length (Fig 1.5b). Each cluster contained between approximately 1000 and 8000 PG4s (Fig 1.5a). We then analysed the relative enrichment of each PG4 cluster on each strand, within genes which were downregulated by NMM, compared to permuted profiles across all genes. No particular PG4 cluster was strongly enriched on the coding strand of downregulated genes (Fig 1.5c). One cluster, however, was strongly enriched on the template strand (Fig 1.5d). This cluster contained PG4s with very short loop lengths of 1-2bp and a total loop length of 5-6bp. This conformed well with our prior expectations, as G4s with short loop lengths are known to form propeller-like parallel G4s of the kind favoured by NMM (Nicoludis et al. 2012).

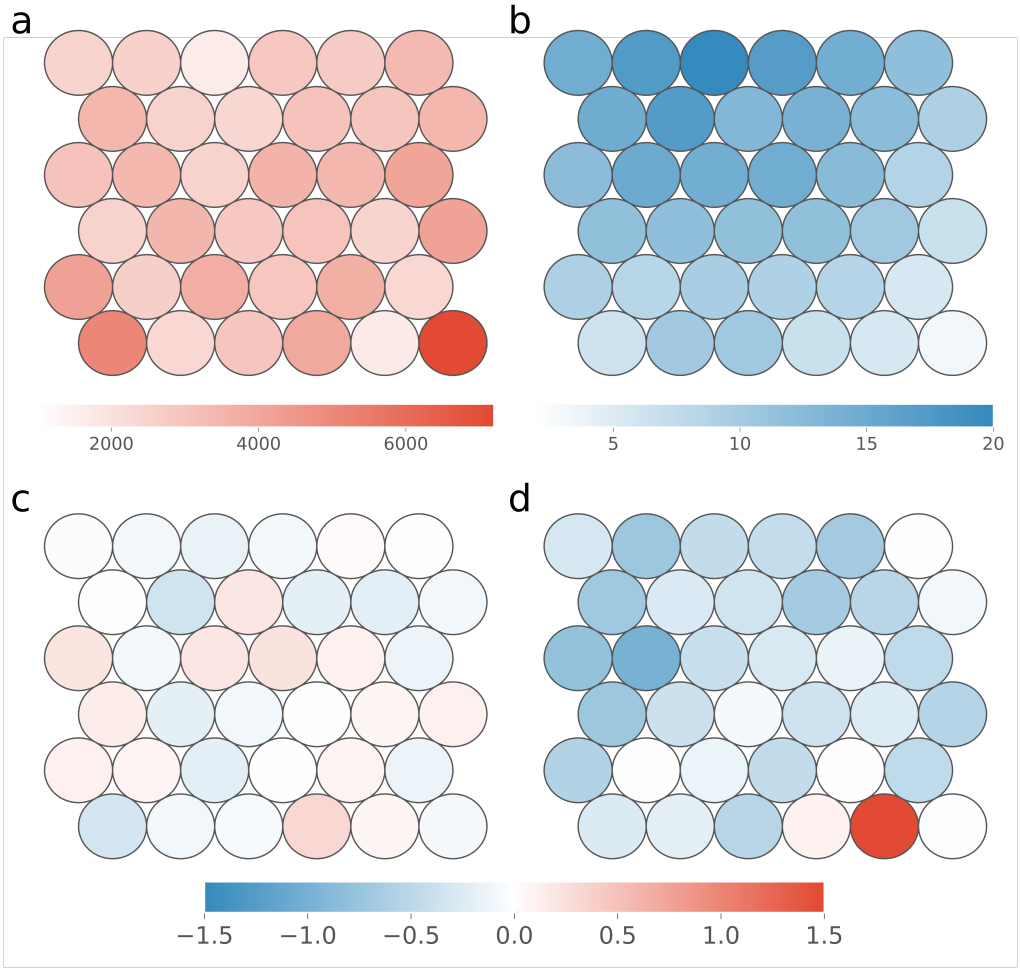

Figure 1.5: **Self Organising Maps demonstrate NMM downregulated genes contain specific PG4 types.** Self Organising Map (SOM) plots for clustering of Quadparser predicted two tetrad PG4s by loop length. In each figure, each circle represents a cluster of similar PG4s. **a)** SOM plot coloured by cluster size. Each cluster contained between 1000 and 8000 PG4s. **b)** SOM plot coloured by total length in bases of all loops. **c)** and **d)** Log2 fold enrichment of each cluster in the gene bodies of genes downregulated by NMM, on the coding and template strands, respectively. Log2 fold enrichments were generated by comparing actual overlap of PG4s with downregulated genes, with expected overlap when PG4s were permuted amongst all genes.

NMM downregulation is correlated with maximal G4 density:

Previous studies have suggested that G4s cause the largest effect on gene expression when grouped in clusters. This may be due to an increase in the likelihood of a single G4 being formed at any one time, or through increased polymorphy of G4 formation. To identify whether NMM regulated genes tend to contain G4 clusters, we used a sliding window of 200bp to count two tetrad G4 density across the whole transcribed region of each gene, including introns. Each gene was then assigned the maximum density score for the gene. Genes were then binned by their maximal density, and expression under NMM was calculated (Fig. 1.6). We found that genes with higher maximal G4 density tended to have more negative log fold changes during NMM treatment. This suggests that clusters of G4s do have a stronger effect on gene expression, and that a single region of high G4 density may be sufficient to cause downregulation of an otherwise G4 free gene during NMM treatment.

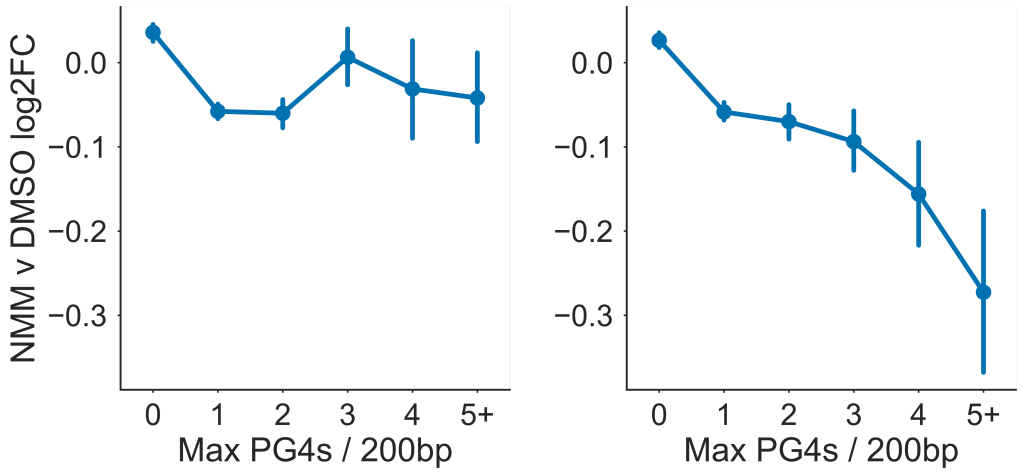

Figure 1.6: **NMM regulates genes with large maximal PG4 density** Point plot showing mean log fold change in gene expression during NMM treatment for genes binned by "maximal PG4 density", defined as the greatest concentration of PG4 motifs within a 200bp sliding window anywhere in the body of the gene (i.e. exon or intron). Left and right panels depict coding and template strands, respectively. Errorbars are 68% confidence intervals for mean generated using 1000 bootstrapped samples.

G Quadruplexes may cause downregulation by Pol II stalling:

Since transcriptional downregulation by NMM stabilised G4s appears to occur most strongly for genes which have many G4s in the gene body, and because this effect is specific to the template strand, we hypothesised that this could be a result of RNA Polymerase (Pol II) stalling. Pol II is the RNA polymerase which is responsible for transcribing all protein coding genes. The Pol II complex scans along the template strand and uses complementary base pairing to produce an mRNA copy which corresponds to the sequence of the coding strand. Since only the template strand is read directly, this might explain why coding strand G4s do not cause downregulation, since they do not form blockages which prevent the elongation of Pol II. To test whether G4s cause blockages which slow or stall the elongation of Pol II in the abscence of stabilisation by NMM, we reanalysed Pol II ChIP tiling array data (Chodvarapu et al. 2012). Metagene profiles of the transcriptional start and end sites showed an accumulation of Pol II at the transcriptional termination site (TTS) (Fig 1.7, orange profiles). This was surprising as it is in disagreement with Pol II occupancy profiles in human cell lines, where there is generally a much larger peak of paused Pol II at the start of the gene. We contrasted this result with occupancy profiles for G4 dense genes (Fig 1.7, blue profiles). For genes which contained at least one G4 dense region, measured either by G4Seeqer score greater than 0.95, or by maximum Quadparser G4 density per 200bp of greater than 3, we found that there was greater Pol II occupancy at the TSS and in the TSS proximal part of the gene body. Greater Pol II

occupancy can be explained by either of two factors: increased initiation and transcription in G4 dense genes, or slower Pol II elongation. Since G4 dense genes do not have higher expression than non-G4 containing genes, we suggest that template strand G4s cause a reduction in Pol II speed. This may result in abortive transcription or alteration of co-transcriptional processes such as splicing.

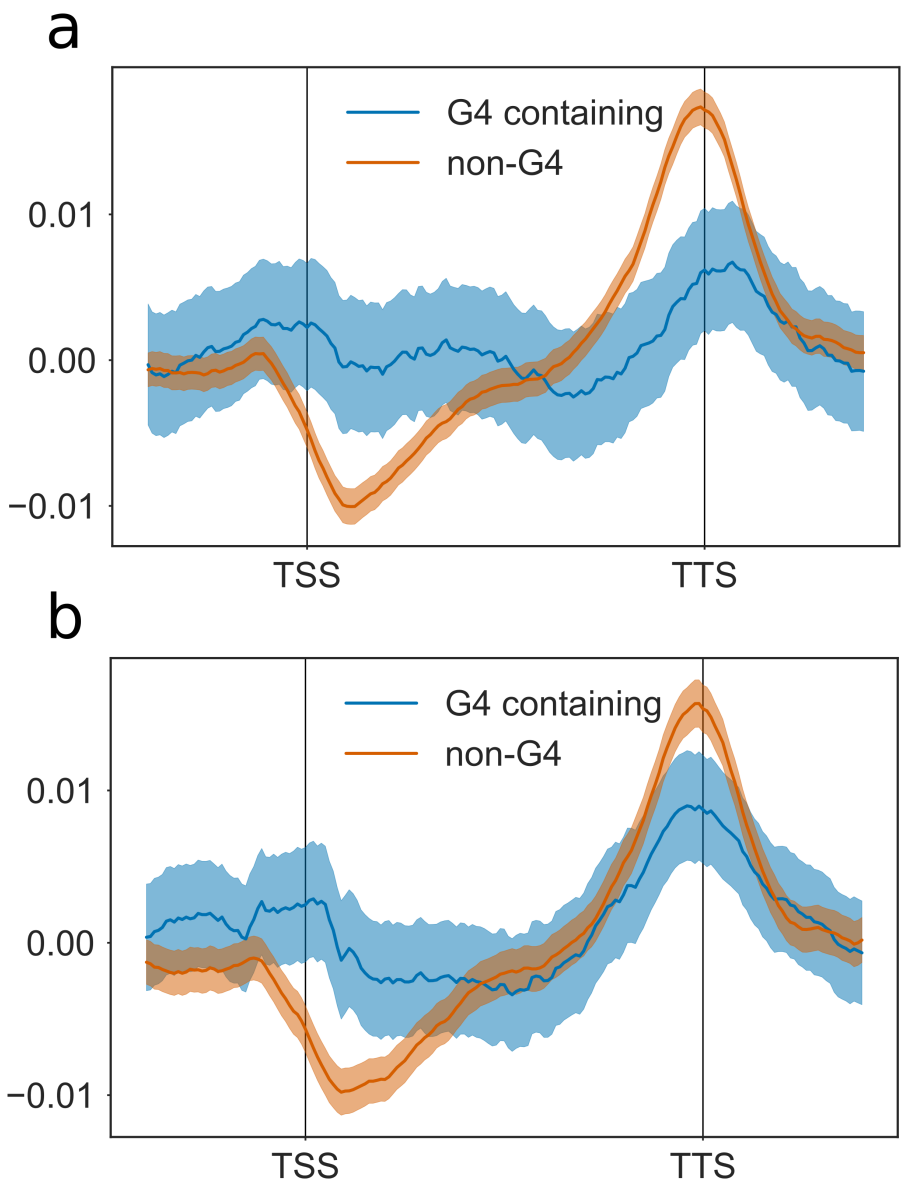

Figure 1.7: **PG4 dense genes have altered RNA Polymerase II occupancy** Metagene profiles showing RNA Polymerase II (Pol II) occupancy across binned exonic gene bodies. Profiles are made up of 500bp upstream region (in 10bp bins), 100 gene body bins, and 500bp downstream region (in 10bp bins). **a)** Metagene profile for genes containing PG4 predicted by G4Seeqer (max G4seeqer score > 0.9), vs. genes containing no G4s (max G4seeqer score < 0.1). **b)** Metagene profile for genes containing two tetrad maximal PG4 density per 200bp of 3 or greater, vs. genes with maximal PG4 density of zero (contain no PG4s). Averages were generated using 1000 bootstrapped samples from each geneset. Bootstrapped samples are normalised such that the absolute area under the curve was equal to one. Errorbars are 68% confidence intervals for bootstrapped means.

G Quadruplex stalling may result in abortive transcription.

To test whether template G4 containing genes might produce abortive transcripts that are degraded by the exosome rather than maturing to mRNAs, we analysed publicly available GRO-seq data (Hetzel et al. 2016). Since GRO-seq captures nascent RNA irrespective of its stability, and RNAseq captures stable RNAs, the ratio between the normalised read counts in GROseq vs RNAseq represents an estimate of the amount of unstable products produced at each gene locus (Fig. 1.8a). We found that the largest difference in ratio was between non-coding and protein coding RNAs, with non-coding RNAs having much greater GRO/RNA ratios (data not shown). This is likely explained by the higher rate of modification of many ncRNAs, e.g. tRNAs, which prevent reverse transcription and sequencing by RNAseq. Other ncRNAs such as natural antisense RNAs may also be unstable and degraded quickly.

G4 predictions were calculated using G4Seeqer and the score of the maximum score region within the transcribed body of each gene was assigned to the gene. A G4 containing set was produced using genes which contained a maximum template strand G4seeqer score of more than 0.95, and a G4 negative set was produced using genes with a maximum score of only 0.05 or less. We found a small but significant positive increase in GRO/RNA ratio for G4 dense genes ($p = 0.009$), suggesting that some abortive transcripts are produced from these genes (Fig. 1.8b, right panel). In contrast, genes with high scoring G4 regions on the coding strand did not have greater GRO/RNA

ratios (p=0.4) (Fig. 1.8b, left panel).

We conducted the same analysis using genes containing three or more two tetrad PG4s in 200bp maximal density clusters, as described above (Fig. 1.8c). For these genes, we did not find any significant increase in GRO/RNA ratio, suggesting that these two tetrad G4s are not sufficiently stable to cause abortive transcription. Since these genes do have higher promoter proximal Pol II occupancy, we suggest that elongation occurs more slowly across the genes, but does not cause premature termination.

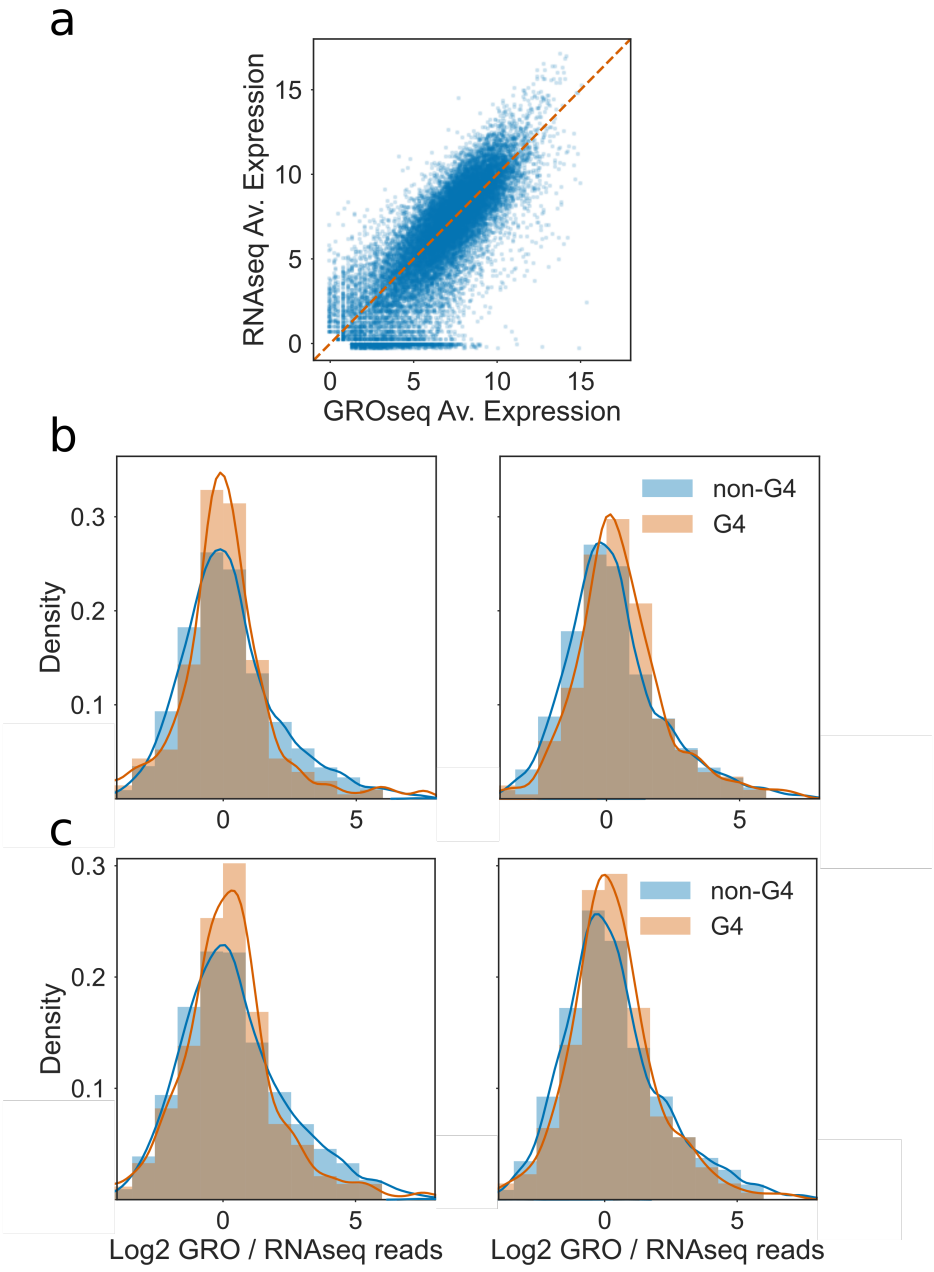

Figure 1.8: **Log Ratio of GRO / RNA seq counts per million detects abortive transcription of PG4 dense genes. a)** Scatter plot showing measured expression in log2 counts per million (logCPM) for each gene from GRO-seq vs. RNAseq datasets. Genes which fall below and to the right of the orange line have positive log2 GRO/RNA ratios. **b)** Histogram and kernel density estimates of GRO/RNA ratio for genes containing PG4 predicted by G4Seeqer (max G4seeqer score > 0.9) in orange, vs. genes containing no G4s (max G4seeqer score < 0.1) in blue. Left and right panels represent coding and template strand G4 rich genes, respectively. **c)** Histogram and kernel density estimates of GRO/RNA ratio for genes containing two tetrad maximal PG4 density per 200bp of 3 or greater in orange, vs. genes with maximal PG4 density of 0 (contain no PG4s) in blue. Left and right panels represent coding and template strand G4 rich genes, respectively.

G4 dense genes are modulated by environmental stress and correlate with NMM treatment:

Because G4s require potassium cations or other divalent cations for formation, it has been noted by previous studies that they might form more readily under stress conditions in plants such as drought stress, when the intracellular concentrations of such ions is increased (Mullen et al. 2012). Mullen et al. showed that genes involved in response to drought stress tended to be more likely to contain G-Quadruplexes in their gene bodies. To more closely examine this hypothesis we reanalysed microarrays conducted using RNA from drought stressed plants (Linster et al. 2015) to determine whether differentially expressed genes contained enrichments or depletion of PG4 structures within various gene regions. As with experiments conducted on the NMM microarray, genes were considered to be moderately differentially expressed if they underwent a log change in expression of greater than 0.5 fold (FDR < 0.05). Genes which changed in expression by more than 1 logfold (FDR < 0.05) were considered strongly differentially expressed. 2947 and 491 genes showed moderate or strong upregulation, respectively, and 2572 and 984 genes showed moderate or strong downregulation. These gene sets were then analysed for two tetrad G4 density using the Quadparser method. As with NMM downregulated genes, we found that genes downregulated during drought stress tended to have greater numbers of template strand PG4s in exonic regions, specifically in CDS and 5' UTR regions (Fig 1.9b-c). Interestingly, we also found that genes which were upregulated during drought stress were more likely to have PG4s in the coding strand of their 3' UTRs

(Fig 1.9d).

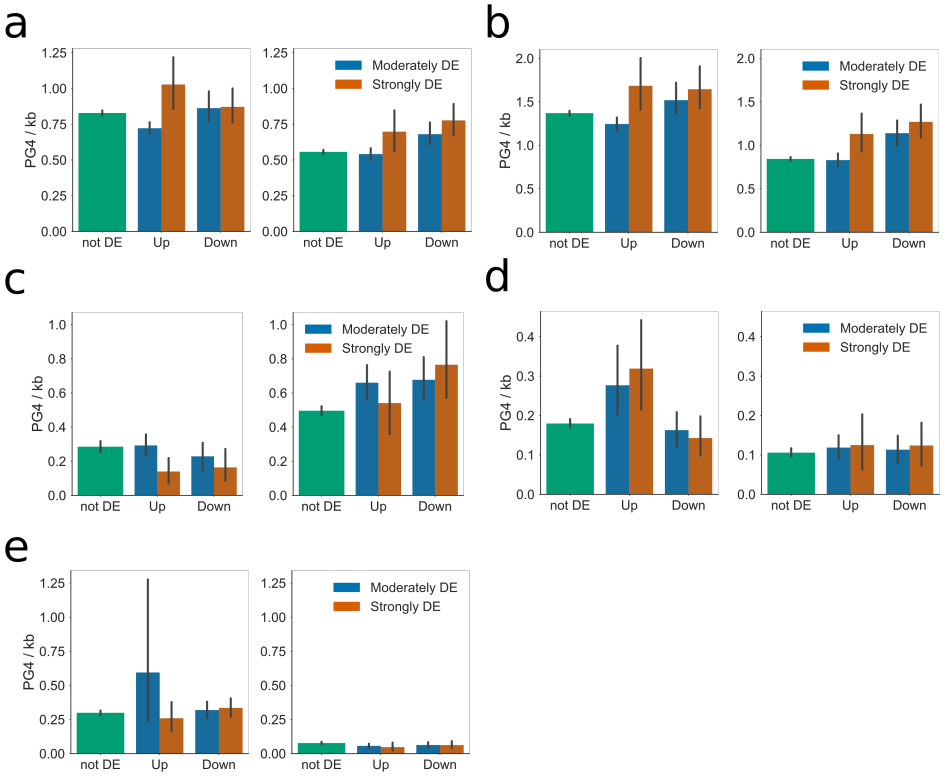

Figure 1.9: **Distribution of PG4s in genes differentially regulated by Drought stress.** Bar plots showing the average PG4 densities of genes up or downregulated by Drought stress, for **a)** full gene body (exons and introns), **b)** coding regions, **c)** 5' UTR, **d)** 3' UTR, and **e)** introns, respectively. In each figure, left and right panels represent coding and template strand, respectively. Genesets are separated into three categories by strength of regulation: green: not differentially expressed, blue: moderately differentially expressed (PPLR < 0.05, logFC > 0.5), orange: strongly differentially expressed (PPLR < 0.05, logFC > 1). Errorbars are 68% confidence intervals for mean generated using 1000 bootstrapped samples. Genes which are downregulated by drought stress tend to have higher PG4 densities on the template strand of coding regions and 5' UTRs. Genes which are upregulated by drought stress are more likely to contain PG4s in their 3' UTRs.

Next we investigated whether there were any similarities in the expression profiles of NMM treated seedlings and drought stressed plants. We found a strong overlap between genes moderately downregulated by NMM and those moderately downregulated by drought stress (p = 1.7e-196) (Fig 1.10a). Analysis of the PG4 density of these gene sets and the overlap between them showed that the genes which were downregulated in both experiments tended to be the most PG4 rich, particularly in the 5' UTR of the gene (Fig 1.10b). This suggests that these genes could indeed be regulated through the same mechanism of G4 stabilisation.

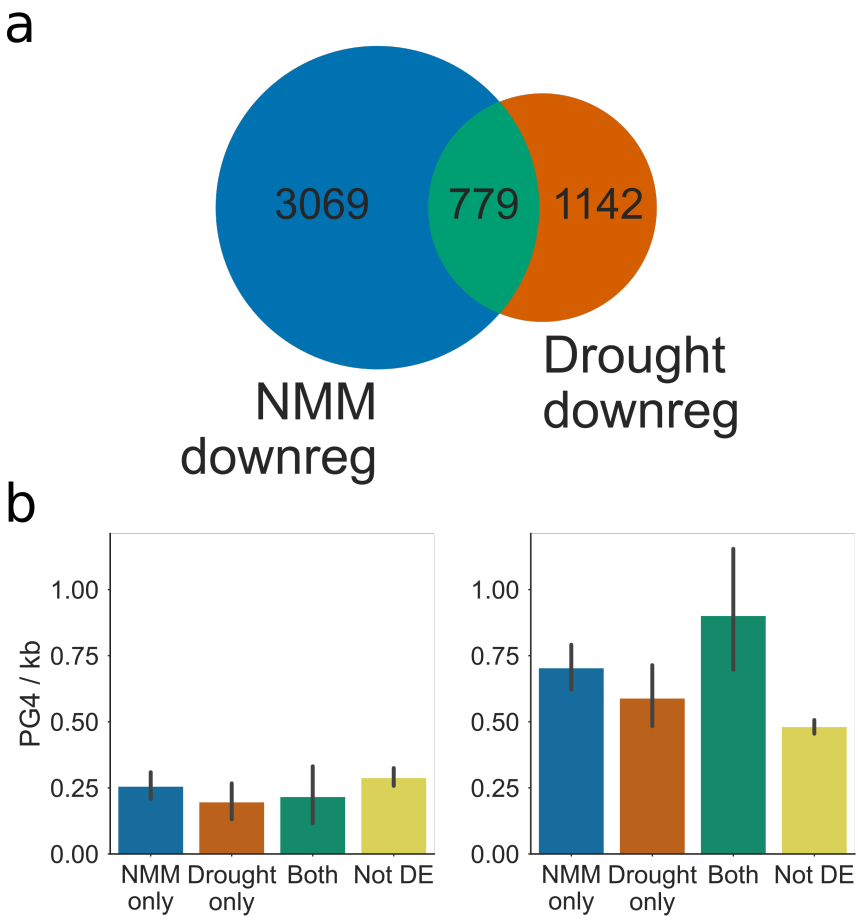

Figure 1.10: **Overlap of genes downregulated by NMM with those downregulated by Drought stress. a)** Venn diagram reporting the overlap of genes downregulated by NMM with those downregulated by drought stress. **c)** Bar plot showing the average PG4 densities in 5' UTRs of NMM and drought stress downregulated genesets. Left and right panels show densities on the coding and template strands, respectively. Both genesets show an greater exonic G4 density on the template strand than genes not regulated by either drug, however genes which are regulated by both drugs had the greatest average PG4 density in 5' UTRs. Bar colours match set colours from Fig 3b. Errorbars are 68% confidence intervals for mean generated using 1000 bootstrapped samples.

Discussion:

We have conducted the first in detail analysis of the effects of G4 stabilisation in a higher plant, Arabidopsis thaliana. To determine how gene expression is altered by G4 stabilisation, we carried out a microarray analysis of plants treated with the G4 binding drug, NMM. NMM treatment had very strong effects on gene expression, particularly in genes which contained large numbers of parallel two tetrad G4s in the transcribed gene body. Moderately downregulated genes had large numbers of G4s in their 5' UTRs, whilst very strongly downregulated genes contained large numbers of G4s throughout the exonic regions of the gene. This is contrary to evidence from human systems which has suggested G4s influence transcription most when located in the promoter region. Furthermore, we find that the effect on gene expression is entirely strand dependent: only template stranded G4s strongly affect expression. Many previous studies have suggested that two tetrad G4s are not biologically relevant due to their relative instability compared to three tetrad G4s. The majority of these studies have been focussed upon their role in human biology. However, several other studies have shown that two tetrad G4s are stable *in vitro* and at physiological temperatures. In organisms that exist at lower temperatures it is entirely credible that two tetrad G4s may play a role. Indeed, here we find that the strongest effect of NMM treatment is on genes predicted to form only two tetrad G4s.

Since the effect of NMM on gene expression mainly appears to be confined to genes with template stranded G4s, G4s will not be present in the mRNA of downregulated genes. Any direct binding of NMM to G4s must therefore

occur in the DNA. The most likely explanation for a template stranded effect is that stabilised G4s interact with the elongating Pol II, which uses the non-coding strand of genes as a template for RNA polymerisation. Since G4s have previously been shown to cause polymerase stalling in vitro, we investigated the Pol II occupancy profiles of G4 containing genes. This data showed that G4 containing genes had higher Pol II density at the TSS proximal end of the gene, and lower density at the TTS. An increase in Pol II occupancy could be explained by one of two factors: more Pol II molecules binding and initiating transcription; or a reduction in Pol II speed. We suggest that G4s in the template strand block the elongation process and cause Pol II to slow down, resulting in a higher Pol II occupancy upstream of the G4 dense region. The Pol II data analysed is captured in the absence of any G4 stabilising drugs, indicating that G4 dependent Pol II slowing is a commonly occurring phenomenon. We hypothesise that two tetrad G4s form naturally in genes, causing Pol II slowing, but still creating full length products. Changes in Pol II speed may have consequential effects for co-transcriptional processes such as mRNA splicing. In some cases blockages may cause premature termination, resulting in truncated products. Analysis of the ratio of GRO/RNAseq read counts suggests that G4 dense genes do indeed produce slightly more unstable products than genes containing no G4s. In the presence of NMM, however, stabilised G4s are likely to become too difficult for the transcription complex to unwind, and cause greater levels of premature termination, the products of which are presumably degraded. The result of this is the dramatic downregulation seen in the microarray.

Previous studies of PG4 localisation in Arabidopsis have highlighted a greater number of two tetrad PG4s in genes annotated as responsive to drought than expected given the distribution across all genes (Mullen et al. 2012). Since intracellular potassium levels are increased during drought stress, it is possible that the stability of G4s could be increased. To investigate this potential G4 dependent regulatory mechanism, we analysed microarray data from drought stressed plants. We found that genes which are downregulated by drought stress contained more PG4s in the template strand, particularly in 5' UTRs and to a lesser extent in coding regions. This result matched closely to the enrichment of G4s in genes downregulated by NMM. Indeed, when we studied the overlap between drought stress downregulated genes and NMM downregulated genes, we found a strong overlap. Furthermore, the genes which were downregulated in both conditions were those which had the greatest PG4 density in their 5' UTRs. We suggest that during drought stress G4s in these genes form more strongly, causing blockages that pause Pol II, downregulating the expression of the gene. Finally, we found that genes upregulated by drought stress tended to contain higher levels of G4s in their 3' UTRs. This effect was not replicated by NMM treatment, suggesting an alternative mechanism of action. Since the 3' UTR is known to be an important regulator of mRNA stability and translation, we speculate these G4s form more strongly in the mRNA during drought stress and recruit some G4 binding factor which could enhance the stability of the mRNA.