Chapter 1

Introduction

The G-Quadruplex

What is a Quadruplex?

The genome is often referred to as the "blueprint" for life. This metaphor suggests a static set of instructions, which simply encodes data and does not change its form. In reality, however, chromosomes are highly dynamic structures which are able to undergo various covalent modifications to the DNA and proteins, as well as change topologies on a global and local scale. At the global level, changes in how the chromatin is packed result in the closing off or opening up of specific regions, changing the level of transcription of the genes contained within them. Genes which are far apart in sequence space can be brought together through looping to allow co-regulation. At the smaller scale, the DNA itself is able to fold into a number of different shapes, including various types of duplex, triplex, and quadruplex. Examples include B-DNA (the classic double helix), A-DNA (duplex), R-loops (triplex), i-motifs (quadruplex) and G-Quadruplexes. These structures have different relative stabilities depending on the conditions of their local environment, e.g. the local concentration of solutes, complementary RNAs, or stabilising proteins, the level of molecular crowding, or the intracellular pH. Furthermore, some structures form only in DNA containing specific sequence attributes, such as high GC content. Whilst non-B forms of DNA have been known to form *in vitro* for some time, it is only recently

that evidence of their formation *in vivo*, and their effects on biological systems, has begun to accumulate.

One of the more well studied non-B DNA/RNA forms is the G-Quadruplex (G4), a Guanine rich, four stranded DNA helix. The properties of guanine which allow G4s to form were first hinted at by German chemist Ivar Bang in 1910, more than four decades before Watson, Crick and Franklin deduced the structure of the double helix (Bang, 1910). Bang noted that guanosine nucleotides in concentrations of around 25 mg/ml will form a viscous gel. It was not until 1962, however, that Gellert et al. were able to use the technique of X-ray diffraction to identify the interactions which caused this property. They noted that guanine monomers were able to interact to form square, planar quartets, which then stacked to form a helical structure (Gellert et al. 1962).

Many more recent studies have shown conclusively that far from being just an unusual property of monomeric guanine, square planar "G-quartets" can form from polymeric DNA, so long as it contains a high enough local concentration of guanines. The guanines in a quartet interact through non-canonical Hoogsteen base pairing, with two hydrogen bonds occuring between each adjacent base (i.e. each base participates in four hydrogen bonds, for a total of eight bonds per quartet) (Fig 1.1a). These quartets can then stack on top of one another through hydrophobic interactions (Fig 1.1b). The number of stacked quartets in a G4 is usually referred to as the number of tetrads. The stability of G4s is generally proportional to the number of tetrads they contains. Mono or divalent cations, usually potassium, fit into the central channel of the G4. These sit equidistant from each guanine of each quartet, and between each pair of adjacent tetrads. The number of potassium cations is therefore one less than the number of tetrads (Fig 1.1b). Potassium is strongly stabilising of G4s, and so the stability of G4s is dependent upon potassium concentration.

The structure of G4s is highly polymorphic. Quadruplexes can form between multiple molecules (intermolecular) or from the same molecule (intramolecular). Intermolecular G4s tend to be less common *in vivo* than *in vitro* however as the effective concentration of G4 forming motifs *in vivo* is generally much lower than under experimental *in vitro* conditions. On top of this, G4s can fold with DNA strands in the same orientation (parallel), in different orientations

(anti-parallel) or in a mixed hybrid conformation (Fig 1.1 c). Which conformation is chosen is dependent upon the sequence from which the G4 is formed, as well as environmental conditions. The loops which connect the G-rich "pillars" of the G4 can be connected laterally (resulting in anti-parallel conformation), diagonally (also anti-parallel) or through a "propeller" like fold (resulting in parallel strands). An anti-parallel G4 with only lateral loops is generally referred to as a "chair" like G4, whilst an anti-parallel G4 with a diagonal loop is referred to as "basket" like (Fig 1.1 c). Many G4 forming sequences will fold into multiple conformations with different rates, resulting in two or more subpopulations of folded G4 from molecules of the same sequence.

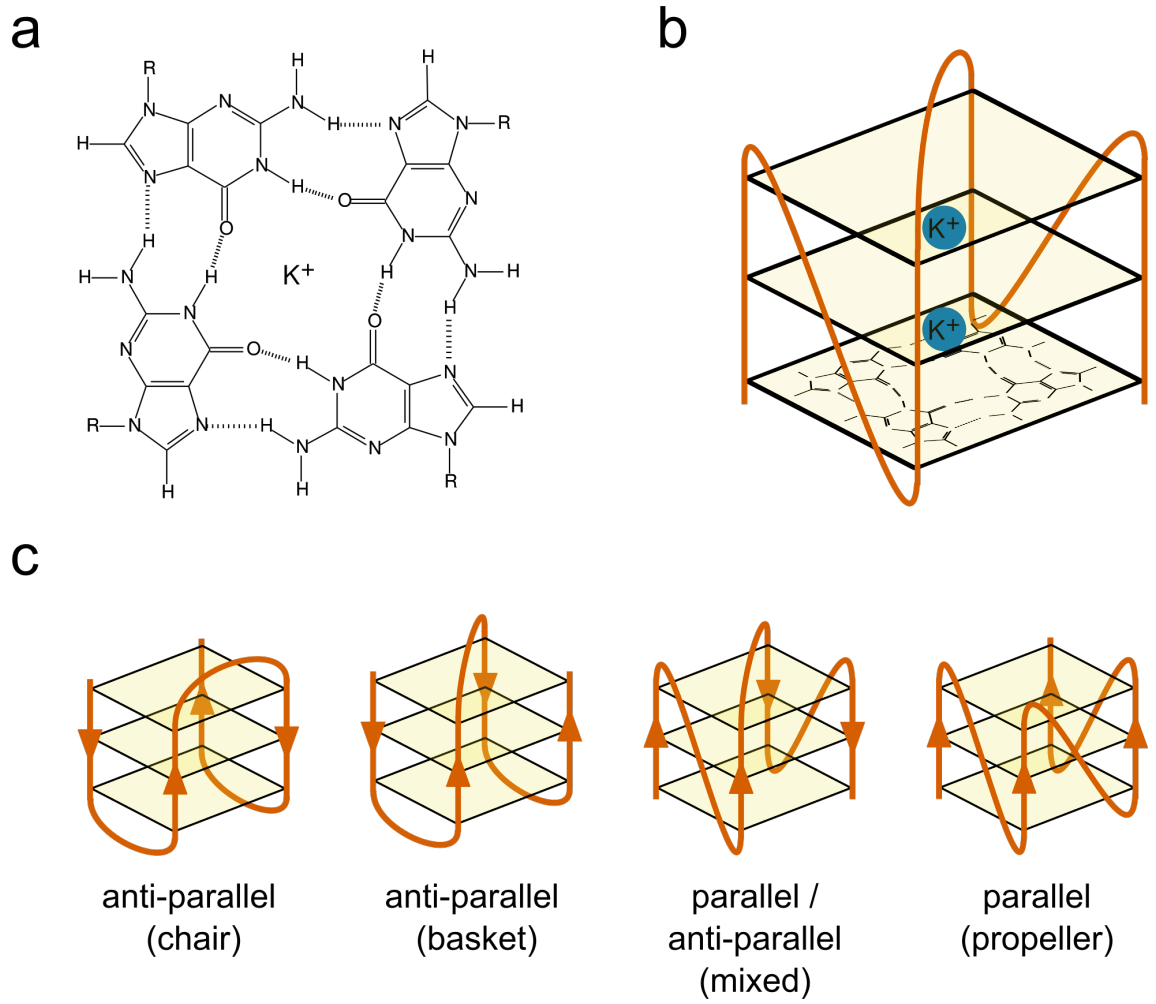

Figure 1.1: Structure of a G-Quadruplex a) The molecular structure of a G-quartet. Four Guanosines (only guanine base is shown, sugar-phosphate is represented as R) interact through Hoogsteen base pairing around a central monovalent cation. **b)** Cartoon showing the basic structure of a three tetrad G4. Three G-quartets are stacked through interactions between delocalised pi electrons. The structure is made up of four pillars of homopolymeric G-runs (shown in orange) joined by loop sequences. Potassium cations (shown in blue) sit between each tetrad. **c)** Cartoons showing how loop arrangement can contribute to the structural polymorphism of G4s. Loops can be lateral, diagonal or propeller like, resulting in anti-parallel, anti-parallel, and parallel G4s respectively. Anti-parallel G4s with all lateral loops are referred to as "chair"-like, whilst anti-parallel G4s with a diagonal loop are referred to as basket like. G4s can also contain a mix of parallel and anti-parallel strands.

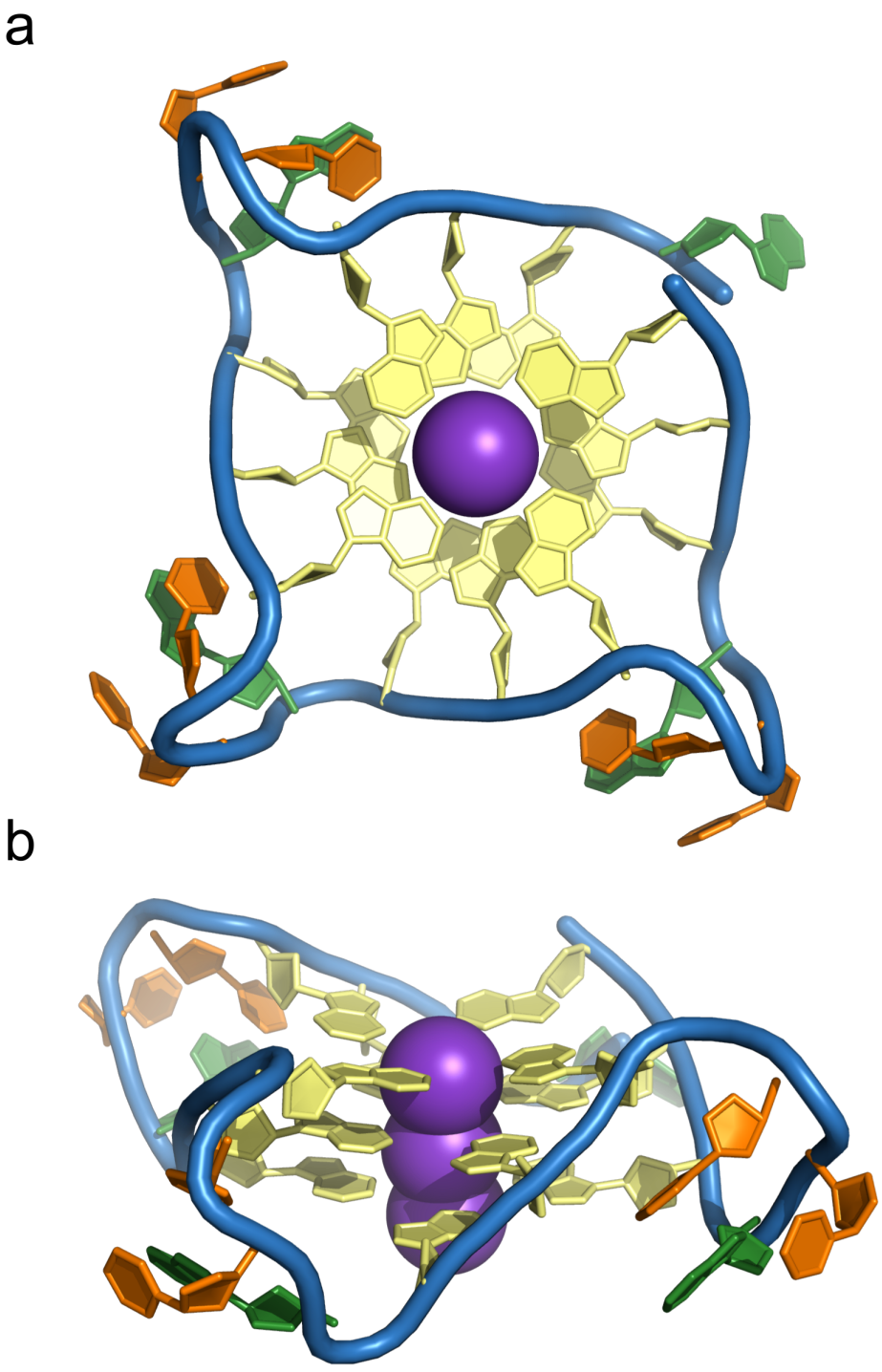

Figure 1.2: The Human Telomeric Sequence forms a G-Quadruplex a) top and **b)** side view of a parallel G4 structure solved by X-ray crystallography (PDB: 1KF1, Parkinson et al. 2002). The sequence used is (TTAGGG)4, corresponding to four of the human telomeric repeat. Gs are coloured in yellow, As in green and Ts in orange. Potassium cations are coloured in purple.

G-Quadruplex Prediction from Sequence

Genomic G4s form in sequences with high G4 content and high GC skew. Because of the dependence of G4 structure on sequence, it is theoretically possible to predict the genome-wide prevalence of G4s from sequence information alone, assuming that other conditions such as potassium concentration are held constant. The first attempt to characterise putative G4s (PG4s) at whole genome scale was conducted by Huppert and Balasubramanian in 2005. They formulated rules describing the general patterns that PG4 forming sequences tend to follow. Their first observation was that intermolecular G4s are unlikely to be common *in vivo* due to low strand concentration of the DNA. They also noted that the pillars of the G4 tended to be formed from contigous guanine homopolymers, or G-runs. There have to be four such G-runs in close proximity to create a PG4, and the length of the shortest G-run will determine the maximum number of stacked tetrads which can be formed. A minimum number of 3 tetrads was suggested for prediction: whilst 2 tetrad G4s are possible, they are less stable. Finally, they suggested that to make folding of the G4 favourable, the length of the loop sequences connecting the G-runs should be relatively short. They suggested, using evidence from molecular modelling and CD spectroscopy, an upper limit of 7bp. Again, whilst loops of much longer length are possible, they were thought likely to be unstable. Their observations were combined to create the folding pattern $G_X N_{1-7} G_X N_{1-7} G_X N_{1-7} G_X$, where $X \geq 3$. This was named the Quadparser method (Huppert and Balasubramanian 2005), and can be applied to search genomes using simple regular expression machinery.

The Quadparser method has been successfully used to identify G4s in many organisms, however the it is not perfect and results in many false negatives as well as false positives. Various adjustments can and have be made to the pattern, including increasing loop length to a maximum of 12bp, allowing two tetrad PG4s, and allowing short bulges in G-runs. These tend to increase the recall of the method but also greatly increase the number of false positives. Other methods have been proposed, such as G4Hunter (Bedrat et al. 2016), which allow PG4s to be given a numeric score based on the GC content and skew of the sequence. G4Hunter is generally performed using a sliding window between 20 and 40bp in length, and is evaluated

for each window by the following method:

```
score = 0
for base, run_length in run_length_encode(sequence):
    if base == 'G':
        score += min(run_length, 4) ** 2
    elif base == 'C':
        score -= min(run_length, 4) ** 2
    else:
        pass
score /= len(sequence)
```

Sequences which have high PG4 forming ability on the positive strand will therefore be given strong positive scores, whilst sequences with PG4 forming ability on the negative strand will be given strong negative scores. A threshold value is chosen below which to filter out non-PG4 forming sequences. Bedrat et al argued that this method was an improvement over the Quadparser technique because it was more flexible, however it also results in a much greater number of false positives when applied to a whole genome, since there are no constraints on how the G-runs are arranged in the windowed sequence.

Methods used in the characterisation of G-Quadruplexes

Gellert et al. first characterised G4 structure using X-ray diffraction of fibres from dehydrated guanine gels (Gellert et al. 1962). Since then, biophysical techniques have become key in the study of G4 structures *in vitro*. Since the advent of chemical DNA oligonucleotide synthesis in the 1980s, it has become relatively cheap to order high purity single stranded oligonucleotides for PG4 sequences, and produce micromolar concentration solutions which can be probed by CD spectroscopy or NMR.

Circular dichroism (CD) spectroscopy utilises the difference in absorbance of circularly polarised light by molecules with chiral structures (i.e. with non-superimposable mirror images). Parallel and anti-parallel G4s both exhibit unique CD absorbance spectra which are distinct from the spectra of disordered single stranded DNA. Solutions containing multiple subpopulations of different parallel and antiparallel G4s will produce spectra which are more complex to interpret, but are still clearly distinct from unordered DNA. Melting temperatures of G4s can be determined using CD or UV spectroscopy temperature gradients measured at 295nm.

Nuclear magnetic resonance spectroscopy (NMR), specifically Proton-exchange spectroscopy (1H-NMR), can also be used to identify G4 DNA. NMR is conducted in deuterated water, since deuterium has a spin of 1 and therefore does not contributed to the NMR spectrum. In single or double stranded DNA, imino protons in the guanine nucleotides will be exchanged with the solvent on short timescales, resulting in loss of the imino proton signal as they are replaced with deuterated protons. In G4 DNA, on the other hand, imino protons are located centrally within the G4 structure, and are therefore protected from exchange. This means that the 1H spectra can be used to distiguish . This does not however identify whether the G4 has parallel or anti-parallel topology. Further characterisation can be conducted using Nuclear Overhauser effect spectroscopy (NOESY) to identify spatial relationships between protons in the G4.

Finally, since folded G4s with short loops and flanking sequence are relatively globular, a number of G4 structures have been crystallised from oligomers. The structures of these crystals have then been solved by X-ray crystallography.

Studies of the structure of G4s has led to the development of a variety of G4-binding ligands. These have a wide range of structures and bind to the various G4 topologies with different strengths. The major classes of G4 binding ligands are: external stacking ligands, which have delocalised pi electron systems, and stack on the the hydrophobic surfaces of the outer G4 tetrads; intercalating ligands, which bind in the groove between stacked tetrads; and external groove binding ligands, which insert into the groove of the G4 helix, or interact with loop sequences (Chen et al 2014). Small molecules which bind G4s include: porphyrins, such as N-methyl-mesoporphyrin (NMM), an external stacking ligand; Pyridostatin, another external stacking molecule specifically designed for G4 binding; and Berberine, a naturally occuring alkaloid and G4 external stacking agent (Fig 1.3). These molecules have been used to study G4s *in vitro* and also *in vivo*.

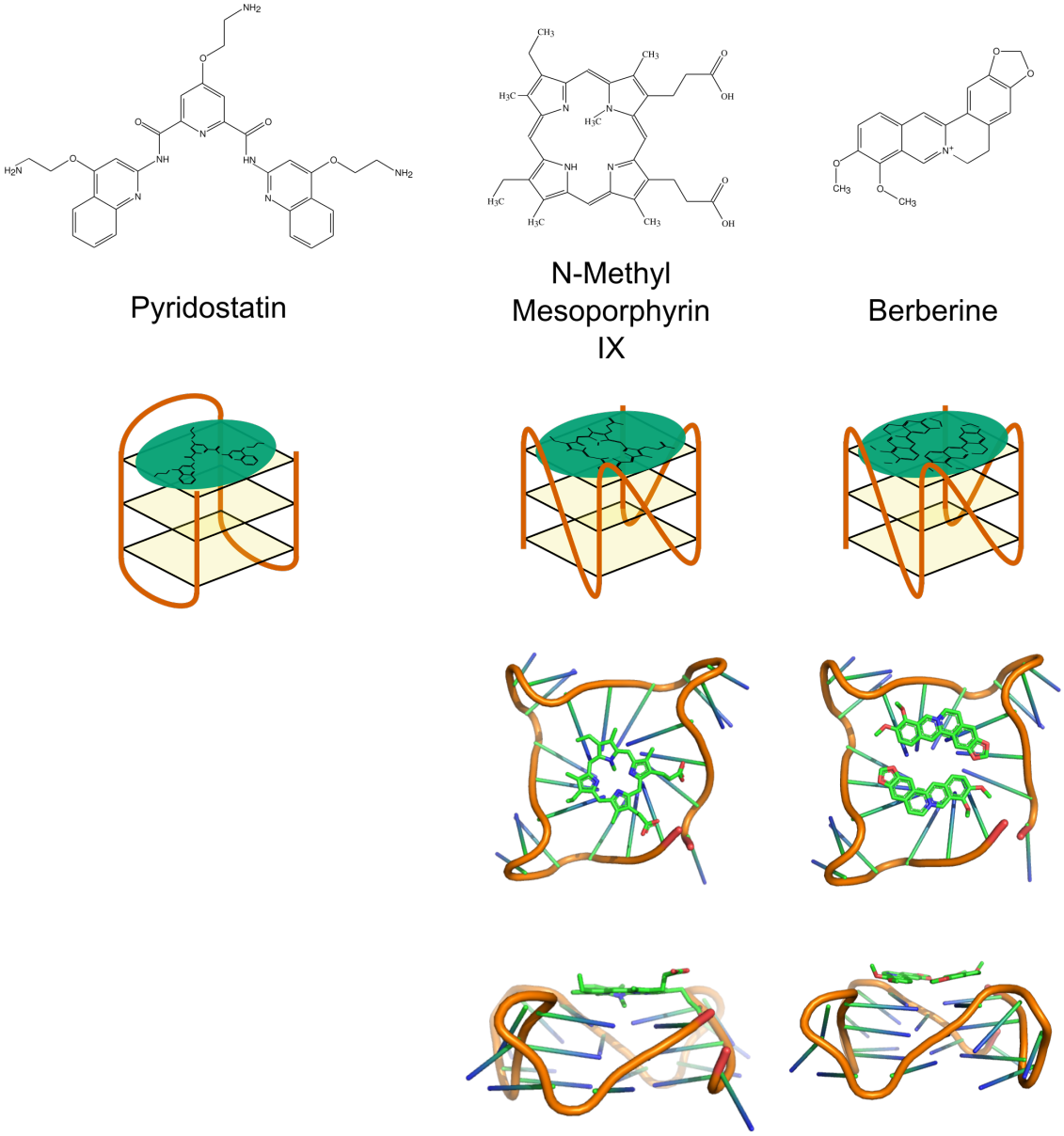

Figure 1.3: G-Quadruplex Stabilising Ligands Structures and mode of action of Pyridostatin, NMM, and Berberine. Pyridostatin is thought to bind to all G4s equally well (Müller et al 2010), whilst NMM is known to prefer parallel G4s (Nicoludis et al. 2012). Berberine has been reported to bind to both parallel and anti-parallel G4s (Bazzicalupi et al. 2013, Li et al 2017). Crystal structures of human telomeric DNA (parallel G4 form) in complex with NMM and Berberine are from PDB entries 4FXM and 3R6R respectively (Nicoludis et al. 2012, Bazzicalupi et al. 2013). Potassium ions and solvent molecules have been hidden for visualisation purposes.

Whilst biophysical methods have led to a wealth of data on G4 formation *in vitro*, biological evidence of G4 formation has been a much later development. One common approach to studying the effect of G4 stabilisation on biological processes is to treat cells or organisms with G4 binding ligands such as NMM or pyridostatin. This has been shown to have various effects on replication, genome stability, transcription and development. Naturally fluorescent or fluorophore labelled small molecules are commonly used to localise G4s by microscopy, or study their *in vitro* folding by single molecule Förster resonance energy transfer (smFRET) (Maleki et al. 2017). Biotinylated pyridostatin has also been used to pull down G4 DNA structures from human DNA (Müller et al. 2010).

Moving beyond small molecules, Biffi et al developed an antibody which specifically recognises G4 DNA using synthetic phage display technology (Biffi et al. 2013). They used the antibody, named BG4, in immunofluorescence experiments to visualise G4s in human chromatin (Biffi et al 2013). More recently, it was used in Chromatin Immunoprecipitation sequencing (ChIP-seq) experiments to identify the specific regions of human chromatin where folded G4s occur (Hänsel-Hertsch et al. 2016).

A number of techniques have been employed for whole genome or transcriptome mapping of G4s. In a method they named G4-seq, Chambers et al. introduced potassium or pyridostatin into the buffer of Illumina sequencing-by-synthesis reactions (Chambers et al. 2015). This resulted in G4 formation in the single stranded DNA fragments, which caused stalling of DNA polymerase, resulting in sequencing errors. They conducted this method on single stranded DNA derived from the human genome. Clusters of identical reads generated by bridge amplification in an Illumina flow cell were first sequenced normally. The products of the first sequencing-by-synthesis reaction were then washed off, and the clusters were resequenced in the presence of the G4 binding agents. This yielded an intial high quality, mappable read, and an error prone read. When they mapped the intitial reads to the genome, the number of sequencing errors in each position was considered an indicator of the G4 forming potential. The authors showed using this technique that only 30% of experimentally observed G4s in the human genome conform to the Quadparser motif

Yoshida et al. also used a similar method to identify G4 clusters in human genomic DNA, by

PCR amplifying sequences in the presence and absence of the G4 binding ligand telomestatin (Yoshida et al. 2018). They showed that regions which contained G4s were amplified with lower efficiency in the presence of telomestatin, due to polymerase stalling events. This resulted in a quantifiable reduction in the number of reads mapping to G4 containing regions of the genome in the drug positive samples, relative to the drug negative samples. The authors identified that regions of the genome which were predicted to form G4 clusters, rather than just single G4s, were most effective at stalling polymerase.

G4s which form in RNA can also be mapped globally, using a technique called rG4-seq, developed by Kwok et al (Kwok et al 2016). This method again utilised stalling at G4s, in the presence of potassium or pyridostatin, this time by the RNA templated DNA polymerase Reverse transcriptase (RT). They identified positions in mRNAs where a reproducible drop in reads occurred in samples where RT mediated DNA synthesis was conducted in the G4 stabilising conditions, relative to unstabilised controls.

G-Quadruplex stability prediction using Machine Learning:

A growing number of G4 forming oligonucleotide sequences have now been characterised *in vitro* by various methods, particularly by CD spectroscopy and UV melting, and the melting temperatures have been published online. The number of these is now great enough that several groups have utilised them to train machine learning models to predict G4 forming potential. Stegle et al. used a Gaussian Process model incorporating extracted features from Quadparser conforming G4s (Stegle et al. 2009). These features were the number of tetrads, the length of each of the three loops, the total loop length, and the frequencies of adenine, cytosine and guanine in the sequence, as well as the raw sequence itself. A second kernel which incorporated features about the conditions the melting temperature was acquired under, i.e. the concentrations of potassium, sodium, ammonium, and magnesium ions, was also used in the model. They trained this model on a set of 260 DNA G4 melting temperatures which were acquired from a literature search. In a cross validation experiment using 100 random 50% hold out splits, the authors were able to achieve a good level of test set accuracy with

an average of 80% of predictions within 5 degrees of the true melting temperature (Stegle et al. 2009). Furthermore, their model was interpretable, and they were able to identify tetrad number and the length of the central loop as the most important sequence features in PG4 stability. The authors employed active learning to identify candidates from human promoter sequences with high uncertainty in the model, and used CD spectroscopy to characterise them.

Whilst Stegle et al.'s model was successful on Quadparser conforming motifs, it is estimated that ~70% of G4s in the human genome do not conform to this motif (Chambers et al. 2015). More recently, Garant et al. published a method for predicting RNA G4s which was trained on a set of 368 experimentally determined sequences, 149 of which were G4 positive and 179 of which were G4 negative (Garant et al. 2017). From these sequences the trinucleotide contents were extracted, and used to train a densely connected multi-layer perceptron model, with a single hidden layer containing 35 nodes. This trinucleotide trained model had the advantage being more flexible for G4s that do not conform to the Quadparser motif. Their model achieved an average AUC score of 0.92 on hold out sets in a 5 fold cross validation experiment. It did not perform as well as the G4Hunter method (Bedrat et al. 2016), however, when tested on the rG4-seq dataset developed by Kwok et al. 2016, suggesting that some positional information is lost when sequences are converted to trinucleotide content.

Finally, the more recent efforts to produce high-throughput methods for identifying genomic G4s, such as G4-seq developed by Chambers et al. 2015, have created much better in depth datasets for training machine learning models. Sahakyan et al. used the G4-seq dataset in their model. This was a extreme gradient boosted machine model developed using `xgboost`, which regressed the percentage mismatch score of sequences from the G4-seq dataset which conform to the Quadparser method (Sahakyan et al 2017). The authors extracted features from Quadparser conforming PG4s similar to those employed by Stegle et al., including tetrad number, loop length, and mono-, di- and triunucleotide contents of the PG4, and flanking regions. This model was very successful at identifying Quadparser conforming motifs which did or did not actually form G4s, achieving a root mean squared error score of 8.14 (units used were mismatch score in G4-seq experiment, in percentage format) (Sahakyan et al. 2017).

Biological Roles of G-Quadruplexes

Genome Stability & DNA Replication

The distribution of G4s in genomes has been predicted from sequence for a wide variety of organisms, and has been experimentally determined by the techniques mentioned above for the human genome. There is conclusive evidence that G4s are not uniformly distributed throughout genomes, but tend to be clustered at functional locations. By far the strongest enrichment of G4s is seen at telomeres. G4s are also found more than would be expected in origins of replication, gene promoters, and inside gene bodies, particularly the 5' and 3' untranslated regions (UTRs). In these locations, it has been demonstrated that G4 formation has effects on the processes of DNA replication, genome stability, transcription, and translation.

Telomeres are the protein-DNA structures found at the ends of linear eukaryotic chromosomes. They consist of thousands of tandem repeats of a G-rich sequence (Moysis et al. 1998), with a single stranded overhang of around 100-200bp (Makarov et al. 1997). Due to functional limitations in templated DNA replication, the very ends of these cannot be duplicated during cell division. This means that without intervention, the chromosome will gradually shorten with each division. Telomeres therefore serve as protective caps that prevent the loss of important coding DNA from the genome. In humans, the telomeric repeat is (TTAGGG)n (Moysis et al. 1998). This has been identified through various methods as a G4 forming sequence.

Telomeres are coated in architectural proteins, called telomere end binding proteins (TEBPs), which protect the DNA from recognition by DNA damage response pathways. Giraldo & Rhodes showed that a yeast TEBP, RAP1, induces formation of G4 structures in telomeres *in vitro* (Giraldo & Rhodes, 1994). More recently, immunofluorescence experiments using the BG4 antibody showed that G4 foci overlap with fluorescent foci produced by Fluorescent *In Situ* Hybridisation (FISH) of the telomeric repeat in human HEK 293T cells (Moye et al. 2015). These results suggest that telomeric sequences do form G4s *in vivo*, which may be bound and stabilised by TEBPs. Furthermore, Moye et al. identified that telomerase, the

template-independent DNA polymerase which synthesises new telomeric repeats, is able to bind to and partially unwind parallel G4s, but not anti-parallel G4s (Moye et al. 2015). There is therefore the suggestion of a G4 regulated mechanism for telomere maintainence.

G4s seem to also play roles in other aspects of DNA replication. It is well documented that G4s are capable of stalling polymerases *in vitro*. There is also growing evidence that without the assistance of G4 unwinding helicases, G4s might cause polymerase stalling *in vivo*, also. Work by Rodriguez et al. demonstrated that treatment of human cancer cells with pyridostatin causes an increase in the DNA damage marker H2AX, suggesting an increase in double strand breaks (DSB). This damage was ameliorated by treatment with a DNA replication inhibitor, suggesting the damage was caused during replication. This is likely to be the result of replication fork collapse at G4 DNA blockages. The DNA helicase FANCJ, which is a tumour-suppressor gene often mutated in breast and ovarian cancers, is involved in DSB repair and has been shown to preferentially bind and unwind G4s (Wu & Spies, 2016). Mutation of the FANCJ ortholog DOG1 in *C. elegans* results in genome instability, and accumulation of deletions upstream of G4s (Kruisselbrink et al. 2008).

The Bloom gene (BLM) encodes a DNA helicase which, when mutated in humans, causes Bloom's Disease (German, 1993). This is characterised by a reduction in genome stability resulting from a large increase in sister chromatid exchange events (SCE). SCEs are caused by double strand breaks, which are repaired via homologous recombination (Wu, 2007). BLM has been shown to prevent SCEs by unwinding Holliday Junctions (a common form of branched duplex), and restarting collapsed replication forks (Davies et al. 2007). Work by Sun et al. first showed that BLM is also capable of binding and resolving G4s *in vitro* (Sun et al. 1998). Recently, Wietmarschen et al. used a single cell sequencing tehcnique to map the locations of SCEs in cells lacking function BLM, and found that many SCEs were at sites predicted to form G4s (Wietmarscen et al. 2018). This suggests that direct action of helicases is required to prevent genome instability at G4 loci.

In humans, DNA replication occurs from tens to hundreds of thousands of origins of replication which are found at regular distances of around 10-100kb apart (Huberman and Riggs 1968, Besnard et al. 2012). Using genome-wide mapping of replication origins by short nascent

strand sequencing, Besnard et al. identified that the majority of human origins were in GC rich regions of DNA, and that 67% overlapped with motifs conforming to the Quadparser pattern (Besnard et al. 2012). 91% of origins were associated with G4s with loop length of up to 15bp. Furthermore, they found an association between the number of G4 motifs, and the strength of usage of origins, suggesting that G4s might be a recruiting factor for replication machinery.

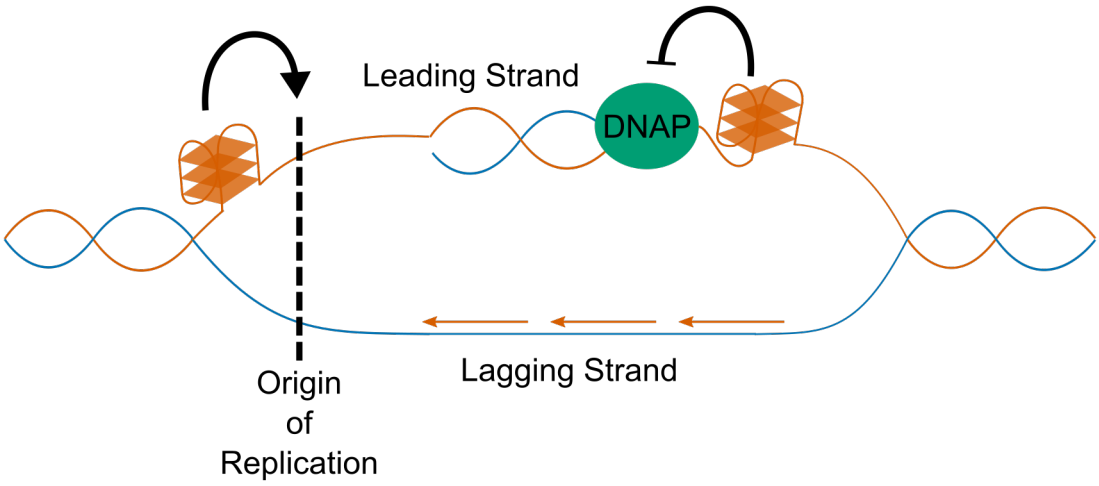

Figure 1.4: Role of G4s in DNA replication

Transcription

Transcription is the process by which DNA is copied into messenger RNA (mRNA) or non-coding RNA (ncRNA), by a DNA-templated RNA polymerase. In eukaryotic systems, all mRNA is transcribed by RNA Polymerase II (Pol II). Initiation of transcription is often catalysed by general or specific transcription factors which bind to the promoter region, upstream of the transcriptional start site (TSS). Human promoter sequences are enriched for PG4 motifs conforming to the Quadparser motif (Huppert and Balasubramanian, 2007, Eddy & Maizels 2006). Tumour suppressor genes have fewer promoter PG4s than might be expected by chance, whilst proto-oncogenes contain more than might be expected (Eddy & Maizels 2006). These also overlap with regions of open chromatin, detected by methods such as DNase Hypersensitivity sequencing (DNase-seq) or Assay for Transposable-Accessible Chromatin by Sequencing (ATAC-seq), suggesting they are often actively transcribed (Huppert and Balasubramanian, 2007). Hänsel-Hertsch et al. used the BG4 antibody to perform ChIP-seq of G4 structures in conjunction with ATAC-seq and RNA-seq, in normal human keratinocytes and an immortalised cell line. They found that BG4 peaks were indeed associated with open promoters, and were found upstream of expressed genes (Hänsel-Hertsch et al. 2016). Interestingly, many more BG4 peaks were identified in the immortalised cells than in normal cells, despite having similar levels of open chromatin. Furthermore, genes which only had a promoter BG4 peak in the immortalised cells tended to be more highly expressed in those cells compared to the normal cells, suggesting that promoter G4s may increase gene expression. This increase in expression may be the result of recruitment of positive transcription factors.

Perhaps the most well studied promoter G4 is the Nuclease Hypersensitive Element III (NHEIII) found in the promoter of the c-MYC oncogene. The NHEIII contains a number of G-rich tracts which have been shown to form G4s *in vitro* by a variety of methods (Simonsson et al. 1998, Siddiqui-Jain et al. 2002, Seenisamy et al. 2004, Ambrus et al. 2005). Formation of a G4 by this region has a strongly repressive effect on gene expression. Siddiqui-Jain et al. found that treatment of cells with the G4 stabilising agent TMPyP4 led to repression of c-MYC, whilst G4 abolishing mutations in a c-MYC promoter luciferase assay caused a three-fold

increase in expression (Siddiqui-Jain et al. 2004). Furthermore, Grand et al. identified that ~30% of human colorectal tumours have G->A transitions which abolish G4 forming potential *in vitro* (Grand et al. 2004). This was associated with increased expression, indicating that promoter G4s are key in regulation (and mis-regulation) of c-MYC. Pull-down of NHEIII binding factors by González et al. identified Nucleolin as a potential G4 interacting partner (González et al. 2009). Nucleolin is a multi-functional protein implicated in ribosome synthesis, transcription and chromatin remodelling. González et al. went on to show that Nucleolin binds to the c-MYC promoter *in vivo*, and that nucleolin overexpression results in downregulation of c-MYC gene expression.

G4s which form within the gene body may have differing effects on transcription, depending on the strand they occur in. Analysis of human gene expression data has suggested that genes containing coding strand G4s downstream of transcriptional start sites tend to have higher expression at the mRNA level than those that do not (Du et al. 2008), even when other factors such as gene function are controlled for. It has been speculated that G4 formation competes with double stranded DNA, creating single stranded "bubbles" which promote Pol II binding and transcription (Rhodes & Lipps 2015). Since the coding strand is not-directly used by Pol II, coding strand G4s will not cause polymerase stalling. G4s which form in the template strand, however, may form blockages which could slow or pause the progression of Pol II.

Transcription progresses by using the template strand as an antisense copy to replicate the coding strand sequence in mRNA. G4s which form in the template strand of gene body DNA will therefore need to be resolved before Pol II can move through them. Due to the relative stabilities of dsDNA and G4s, G4 formation may only occur in the gene body after a pioneering round of transcription, during which the DNA is in single stranded form. Rodriguez et al. showed that some of the DNA damage caused by treating cells with pyridostatin was transcription-dependent, and could be ameliorated by treating cells additionally with an inhibitor of transcription (Rodriguez et al. 2012).

Methods for estimating Pol II elongation rates across genes, such as GRO-seq or BruDRB-seq, have associated changes in speed with various features such as specific histone modifications,

exon density, and sequence features. Veloso et al. correlated elongation rates from BruDRB-seq data with GC content of genes, and found that genes with higher GC content tended to have slower elongation (Veloso et al. 2014). They hypothesised that this could be due to the greater stability of GC-rich duplexes, which have extra hydrogen bonds. It is also possible, however, that this effect could be partially due to greater numbers of G4s in GC-rich genes.

In human cells, profiling of Pol II occupancy by ChIP-seq has demonstrated that there is a large peak of paused polymerase in the first 30-60bp downstream of the TSS (Jonkers & Lis 2015). This pausing is an important checkpoint ensuring Pol II is correctly modified before elongation begins. Genes which require large and rapid increases in expression in response to environmental stresses, such as heat shock proteins, also have large amounts of paused Pol II which can be activated quickly. During initiation of transcription, Pol II is recruited to the TSS by specific or general transcription factors, and transcribes for a short distance before becoming paused. Formation of paused Pol II, referred to as the Pre-Initiation Complex (PIC), is stabilised by the Negative elongation factor (NELF) and DRB-sensitivity-inducing factor (DSIF), as well as by phosphorylation of the carboxy-terminal domain (CTD) of Pol II at Serine 5. Productive elongation can then be restarted by the action of the positive transcription elongation factor-b (P-TEFb) complex, which phosphorylates NELF and DSIF, causing the former to be released from the PIC, and the latter to switch to becoming a positive elongation factor. P-TEFb also phosphorylates the CTD at Serine 2, which is considered a hallmark modification of active Pol II. How Pol II pausing is regulated is still not clear (Jonkers & Lis 2015, Liu et al. 2015). One hypothesis is that the sequence content of promoters and promoter proximal regions may be important for regulating pausing. A number of promoter motifs, such as the GAGA motif or the downstream promoter element (DPE) have been associated with promoters with high levels of stalling (Hendrix et al. 2008).

It is well established that the promoter proximal regions of genes in many organisms have, on average, higher GC content than the rest of the gene body (Veloso et al. 2014). Eddy et al. identified that the first 200bp downstream of the TSS tends to be more GC-rich in genes with high levels of proximal pausing than in genes which do not exhibit pausing (Eddy et al. 2011). The G4-forming potential of these genes also tended to be greater on the

coding strand, meaning that G4 structures might also form in the nascent mRNA. Eddy et al. hypothesised that these 5' mRNA G4s might signal back to the polymerase to produce pausing. A similar mechanism involving an RNA hairpin has been implicated in pausing of *E coli* RNA Polymerase (Toulokhonov et al. 2007).

The enrichment of G4s in promoters and promoter proximal regions suggests that proteins involved in transcriptional complexes may bind specifically to these structures. The general transcription initiation factor complex, TFIIH, contains 11 subunits, and is required for both transcriptional processes and DNA repair through the Nucleotide Excision Repair (NER) pathway (Compe & Egly, 2012). The DNA helicases XPB and XPD are essential components of TFIIH, which catalyse the denaturation of DNA in promoters or around lesions (Coin et al. 2007). Through *in vitro* binding assays, Gray et al. identified XPB and XPD as G4 interacting proteins, which bind G4s preferentially over dsDNA (Gray et al. 2014). XPD was also found to unwind G4s *in vitro*. Gray et al. went on to perform ChIP-seq of XPD and XPB, and showed that it was enriched at TSS loci containing G4s. Approximately 40% of XPD/XPB peaks contained PG4s conforming to the Quadparser pattern (with loop lengths $>=$ 12bp). Furthermore, Hänsel-Hertsch et al. also reported a strong overlap between these XPD/XPB peaks and G4 loci observed by BG4 ChIP-seq (Hänsel-Hertsch et al. 2016). This suggests that TFIIH may be recruited to G4 containing promoters to initiate transcription.

mRNA Processing

Nascent pre-mRNA which is newly transcribed by Pol II must undergo 5' capping, splicing, RNA modification, poly-adenylation and quality control before it can mature into mRNA which is exported to the cytoplasm. Many of these processes occur co-transcriptionally and are tightly co-ordinated to prevent mistakes. Multiple independent studies in different organisms estimate that between 75%-85% of splicing is conducted in a cotranscriptional manner (Armour et al. 2011, Khodor et al. 2011, Girard et al. 2012, Tilgner et al. 2012, Windhager et al. 2012). Oesterreich et al. found that in yeast, 10% of intron splicing is complete when Pol II is only 26bp downstream of the intron acceptor site, and 50% complete when Pol II is 45bp downstream. (Oesterreich et al. 2016). By modifying Pol II to increase its speed 2.3x, they also showed that splicing could become rate limiting when elongation rate is greater. Furthermore, modifications to splice site sequences in a reporter reduced the rate of splicing, presumably by reducing the strength of recognition by snRNAs, and thereby the rate of spliceosome assembly (Oesterreich et al. 2016). This indicates that interplay of splice site strength and Pol II elongation speed determine the relative efficiency of splicing.

It has been estimated that greater than 90% of human genes undergo some form of alternative splicing (Wang et al. 2008). During this process, different donor and acceptor sites compete to be utilised. The most common form of alternative splicing is alternate donor or acceptor usage, where the other site used is constitutive. Other forms include intron retention, where splice sites are simply not used at all, or exon skipping, where constitutive donor and acceptor sites are paired such that an intervening exon is removed from the mature mRNA. Regulation of alternative splicing can occur via protein splicing factors, as well as through changes in Pol II elongation speed. Through this mechanism, alternative splicing may be linked to changes in Pol II speed as it transcribes through template stranded G4s or other DNA structures.

Aside from their effects on Pol II speed, G-rich sequences with the potential to form G4s have been implicated as important intron motifs for splicing. These PG4s are predicted in the coding strand, meaning that they could form G4s in either the DNA or the nascent pre-mRNA. Analysis of the first exon-intron boundary of human genes by Eddy & Maizels revealed that

about 50% of boundaries contain PG4s within the first 100bp of intronic sequence (Eddy & Maizels 2008). They noted that a number of hnRNP family proteins, such as hnRNP A1 and hnRNP H, bind to G-rich motifs in RNA. hnRNP A1 has been called the "swiss army knife of gene expression", due to its ability to bind both chromatin and mRNA, and its putative roles in transcription, mRNA splicing, telomere maintenance, mRNA export, and translation. Interestingly, hnRNP A1 is capable of binding and unwinding DNA G4s, and has been demonstrated to increase expression of the KRAS and c-MYC oncogenes by resolving repressive G4s in their promoters (Chu et al. 2017). Furthermore, there is evidence that hnRNPs F and H are capable of binding to G-rich RNA sequences to regulate splicing (Xiao et al. 2009). Xiao et al. identified that G-rich sequences downstream of donor splice sites with intermediate levels of homology to the snRNA U1 were strongly conserved. They also noted that the expression of genes with these intermediate splice sites and G-runs was sensitive to the knock-down of hnRNP H (Xiao et al. 2009).

A model gene for the study of this G-rich motif dependent splicing is Bcl-X, a regulator of cell death which has two major spliced forms. The dominant isoform is the longer, Bcl-XL, which is anti-apoptotic. A switch in splicing leads to formation of the shorter form, Bcl-XS, which is pro-apoptotic (Min et al. 1996). This switch involves differential donor site usage in the splicing of the second intron. Garneau et al. showed that alternative splicing of Bcl-X is mediated by hnRNP F/H binding to two exonic G-rich regions, one upstream of the Bcl-XL donor site, and another downstream of the Bcl-XS donor site (Garneau et al. 2005). Mutation of these G-runs abolished hnRNP binding and removed the effect of recombinant hnRNP F treatment on Bcl-X splicing. Weldon et al. recently demonstrated that both of the G-rich regions are capable of forming G4s *in vitro*, and that treatment of *in vitro* splicing assays with ellipticine derived G4 binding agents was able to alter the ratios of the spliced forms. The RNA recognition domain of hnRNP F binds to single stranded DNA, suggesting that G4s modulate splicing by preventing hnRNP binding (Dominguez et al. 2010).

mRNA Stability and Translation

G4s which form in mRNA have the potential to be more stable than those formed in DNA, because RNAs tend to form more complex structures and do not have to compete with double stranded forms. Furthermore, structural studies have suggested that the extra hydroxyl groups in ribose compared to deoxy-ribose allow RNA G4s to form more hydrogen bonds within the Quadruplex structure (Collie et al. 2010). This increases the enthalpic favourability of the RNA G4 whilst also reducing the entropic cost of hydrogen bonds with ordered water molecules. Kwok et al. have used rG4-seq to identify mRNA G4s that form *ex vivo* in the presence of potassium and pyridostatin, by their ability to stall RT. They found a total of 3383 G4s in mRNAs, of which 62% were contained in the 3' UTR, 16% in the 5' UTR and 21% were in the CDS (Kwok et al. 2016). Simulation of RNA folding with the RNA G4 region constrained yielded very different structures on average to folding without constraints, suggesting that G4 formation might act as a molecular switch, changing the overall structure of the mRNA (Kwok et al, 2016). RNA G4s have been implicated in a number of regulatory roles, including translational regulation in the 5' UTR, and mRNA stability in the 3' UTR.

Translational initiation in eukaryotes begins with the binding of a 43S preinitiation complex (PIC) to the 5' cap of the linear mRNA (Hinnebusch 2014). The PIC is pre-loaded with a Methionine aminoacyl-tRNA, and scans along the 5' UTR of the mRNA until the first methionine codon AUG is identified. This identification catalyses the recruitment of the large 60S subunit of the ribosome, to produce a translationally active complex (Hinnebusch 2014). The 5' UTR of the mRNA often contains structures such as hairpins which must be resolved by RNA helicases to allow the passage of the PIC. G4s which form in the 5' UTR can have similar consequences, impeding the scanning of the PIC, or, if close enough to the m7G cap, preventing the loading of the complex onto the mRNA (Bugaut & Balasubramanian 2012).

A number of oncogenes, such as NRAS, BCL-2 and VEGF have been identified as containing 5' UTR G4s (Kumari et al. 2007, Shahid et al. 2010, Morris et al. 2010). When folded, these often to have repressive effects on translation. The KRAS 5' UTR, for example, causes reduced expression when attached to a luciferase reporter, compared to the same 5' UTR with

mutations that abolish G4 folding (Kumari et al. 2007). Wolfe et al. showed that treatment of cells with silvestrol, an inhibitor of the RNA helicase EIF4A, caused ribosome stalling on the 5' UTRs of genes containing PG4s (Wolfe et al. 2014). EIF4A has previously been identified as a required helicase for the translation of transcripts with long and structurally complex 5' UTRs (Parsyan et al. 2011). This evidence suggests that 5' UTR G4s can act as negative regulations of translation initiation.

Interestingly, study of the VEGF 5' UTR has identified a G4 which acts as a positive regulator of translation (Morris et al. 2010). VEGF has an extremely long 5' UTR of greater than 1000 bases, which contains two distinct internal ribosome entry sites (IRES). IRES are sequences in the 5' UTR which can control cap independent binding and initiation of ribosome scanning. They are often found in very long and GC-rich 5'UTRs which are more highly structured, and maintain their translation during stresses such as hypoxyia (Jackson 2013). In VEGF, Morris et al. showed that a G4 structure located at IRES-A was essential for internal ribosome initiation (Morris et al. 2010). This indicates that 5' UTR G4s can have complex roles in translational regulation.

G4s in 3' UTRs have the potential to act as *cis* regulatory elements controlling translation efficiency or mRNA stability. One hypothesised method of action is through competition between G4 structures and unstructured UTRs which can bind microRNAs (miRNAs). miRNAs are short (20-25bp) regulatory RNAs which act through complementarity to their target mRNA. Hybridisation of miRNAs to their targets results in changes in mRNA degradation through P-body localisation, or translational repression# (Ambros 2004). The local structure of 3' UTRs has been identified as a key modulator of their binding efficiency (Long et al. 2007). Rouleau et al. performed a characterisation of the overlap between predicted miRNA binding sites and PG4 sequences in human 3' UTRs (Rouleau et al. 2017). The authors found that 54% of PG4s in 3' UTRs overlapped with at least one predicted miRNA target site, though the enrichment or statistical significance of this overlap were not calculated. A candidate mRNA, FADS2, which is regulated by mir331-3p, was tested for G4 formation. Rouleau et al. showed that a 3' UTR sequence which binds mir331-3p also forms a G4, and that G4 formation prevents miRNA binding *in vitro*.

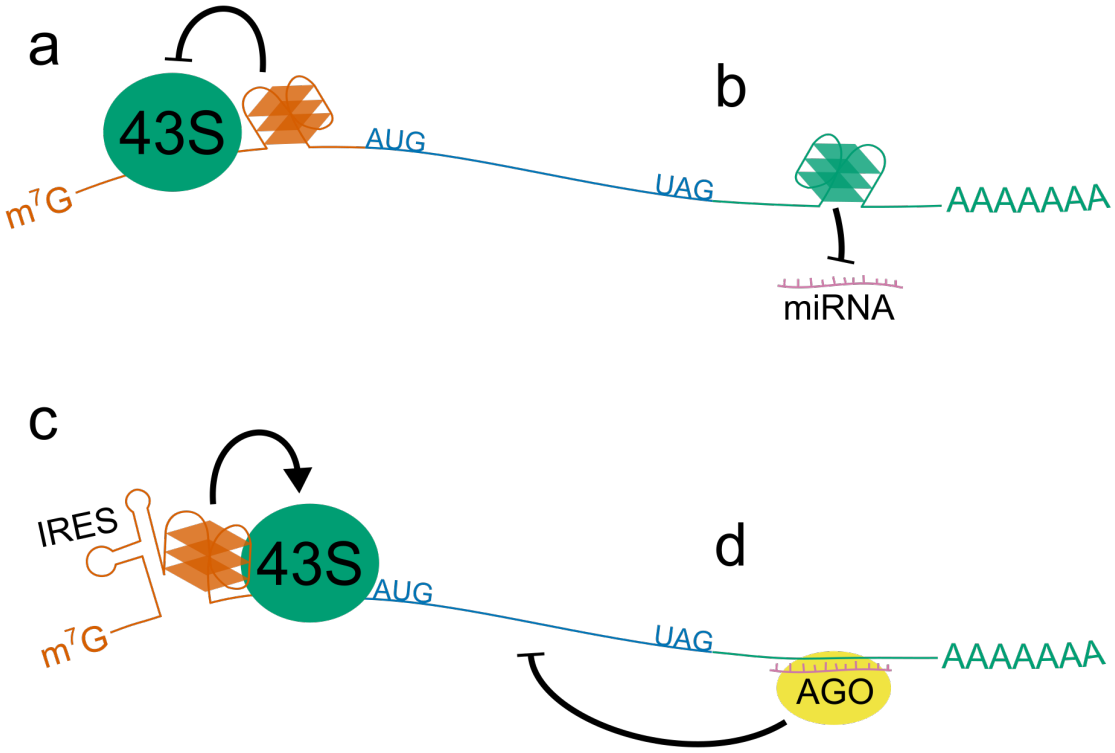

Figure 1.5: G4 function in translation and mRNA stability

Chromatin Remodelling and DNA methylation

Chromatin conformational dynamics are important regulatory mechanisms for gene expression. The major protein component of chromatin is made up of histones, which assemble into the nucleosomes around which DNA is wrapped. Chromatin remodelling is the process by which modifications are made to histones. These modifications include switching of histone isoforms, covalent modifications, and histone sliding. The latter involves the use of energy from ATP hydrolysis to translocate histones along the DNA polymer without ever removing them (Langst & Manelyte 2015).

One class of remodellers is SWI/SNF complexes, which are able to slide or eject nucleosomes. ATRX is an X-linked SWI/SNF family member, which is mutated in ATR-X syndrome, an intellectual disability disorder. ATRX has been shown to locate at heterochromatic and pericentromeric regions, as well as at telomeres (McDowell et al. 1999). During ATR-X syndrome, global DNA methylation patterns at these regions are disregulated (Gibbons et al. 2000). ATRX mutation also effects gene expression. Gibbons et al. identified gene expression changes in the alpha globin gene cluster in response to ATRX mutation (Gibbons et al. 1991). More recently, Law et al. performed ChIP-seq using an ATRX specific antibody to identify other gene targets. They found that ATRX binds specifically to G-rich tandem repeats which form G4s *in vitro*.

Chapter 2

A Recurrent Neural Network to Predict G Quadruplex Structure

Introduction:

Because of the dependence of G4 structure largely on sequence information, it is possible to make predictions about the propensity of specific sequences to form G4s using pattern matching analyses. The initial rule which was employed for putative G4 (PG4) detection in the human genome was the Quadparser method (Huppert & Balasubramanian, 2005). This is a simple regular expression following the pattern $G_X N_{1-7} G_X N_{1-7} G_X N_{1-7} G_X$ where $X \geqslant 3$. The Quadparser method was chosen based upon early Circular Dichroism and UV melting data, which suggested that G4s tended to be more stable with three or more tetrads, relatively short loop lengths, and no bulges in tetrads. These fairly strigent rules for formation mean that in general, the Quadparser method is considered to be fairly conservative, and misses a lot of sequences with high G4 forming potential.

More recently there has been a large increase in the number of available methods for predicting G4s (Bedrat et al. 2016, Hon et al. 2017, Garant et al. 2017, Sahakyan et al. 2017). The contribution from Bedrat et al., named G4Hunter, is a scoring method based on a run length encoding the input sequence. Runs of Gs score positively whilst runs of Cs score negatively. It can be used with a sliding window approach to score an entire genome, with thresholding to identify high scoring PG4s. Whilst this approach is much more tolerant of imperfections which violate the Quadparser method, it is arguably too tolerant, producing many false positives. Furthermore, it does not take into account flanking A and T sequences which may contribute to the stability of the G4, e.g. through reducing the favourability of double stranded DNA.

Some middle ground is required to improve existing methods. The current global interest in machine learning for solving biological problems has recently been brought to the field of G4 research, to attempt to solve this issue. The results of this are two new G4 prediction tools, named G4RNA screener (Garant et al. 2017), and Quadron (Sahakyan et al. 2017). The former, G4RNA screener, is a densely connected neural network which is trained on the trinucleotide contents of input sequences. The data used to train the model is a library of melting temperatures of RNA G4s obtained from a literature search. Currently this database contains only 368 sequences, however. Given the almost inconceivable number of potential

G4 forming sequences however (there are more than 10^{12} sequences which could match the original Quadparser pattern, not including flanking sequences), it is probable that this dataset does not capture all of the variety of possible G4 forming sequences. To do this, a more high throughput method for measuring G4 forming propensity is required.

A new method for sequencing of genomic G4 structures, termed G4Seq, may provide this level of throughput (Chambers et al. 2015). To create this dataset, Chamber et al. sequenced human genomic DNA in an Illumina sequencing-by-synthesis machine, in the presence or absence of G4 stabilising potassium cations, or G4 binding ligand Pyridostatin. The G4 structures caused stalling of DNA polymerase, resulting in a large number of errors in the resulting read. When the authors then mapped the resultant reads, the mismatch rate which occurred at each position in the genome could be counted to create a map of G4-forming loci. Another high throughput method for sequencing G4s, this time in human mRNAs, was developed by Kwok et al. This protocol utilises the ability of potassium or pyridostatin stabilised RNA G4s to stall reverse transcriptase, resulting in a drop off of reads in the 3'->5' direction in RNAseq data (Kwok et al. 2016).

Data from the G4Seq experiment was leveraged by Sahakyan et al. to build an extreme gradient boosted machine model, named Quadron, to predict G4 formation. Quadron initially predicts PG4s using the Quadparser method, but with extended loop lengths of 1-12. Various derived features, such as tetrad number, loop length and mono-, di- and trinucleotide content, are then extracted from these patterns and used to train a model predicting G4Seq mismatch rate. The primary drawback of this method is that the Quadparser method is still used to make initial predictions. This means that Quadron is only able to improve the precision of the existing Quadparser method by rejecting false positives that do not form G4s. Since the Quadparser method is already quite conservative, a lot of potential G4 forming sequences are still missed.

Here we present a new method for G4 prediction, which builds on the work of Bedrat et al. and Sahakyan et al. We use the G4Seq dataset as training data, however a convolutional and recurrent neural network is used to process input sequences directly, meaning fewer prior assumptions are required about what constitutes a G4 forming sequence. Our new method,

which we name G4Seeqer, performs better on both the G4Seq datasets and other datasets, including G4s immunoprecipitated from chromatin (Hänsel-Hertsch et al 2016). We also use transfer learning to apply the model learned from G4Seq data to RNA G4 prediction. Finally, we use G4Seeqer to characterise unnoticed features of the G4Seq dataset, and compare our scores to data from UV melting experiments.

Materials and Methods:

G4Hunter Algorithm:

The G4hunter algorithm from Bedrat et al 2016 was reimplemented in Cython (Python superset which can be compiled to C) with some alterations. Input sequences were run length encoded, and each run of Gs was scored as the square of length of the run, with a maximum score per run of 16. These scores were summed to give a positive strand total score. Runs of Cs were scored equivalently but in a separate negative strand score. Scores were divided by the length of the input sequence to get a normalised score.

Training Data Preprocessing:

The modified G4Hunter method was run on the hg19 genome using a window size of 50bp, a step size of 5 and a threshold of 0.75 to generate a total of 7484506 candidate G4 intervals, which were output in bed format. Intervals were increased in size by 39bp in each direction using `bedtools slop` to introduce flanking sequence information for classification. Overlapping intervals were filtered using a dynamic programming technique commonly used in the field of interval scheduling. Intervals were weighted by their G4Hunter score, such that the maximum number of high scoring non-overlapping intervals was yielded.

To ground truth score these sequences, `bedtools map` was used to intersect them with the G4Seq dataset (Chambers et al. 2015), which was downloaded from GEO (GSEXXXXX). Bedgraph files of this data contained percentage mismatch scores for each position of the human genome at 15bp resolution. The G4Seq dataset generated in the presence of potassium was chosen as it was deemed more likely to be of biological relevance than the dataset generated in the presence of Pyridostatin, a G4-binding drug.

Intervals files and corresponding mismatch scores were read into Python using `pandas` and histograms of log transformed mismatch scores were plotted using `matplotlib`. The threshold of approximately 3 for separating G4-forming and non-G4 forming sequences was chosen using

scipy to determine the local minimum in the histogram. Joint plots of percentage mismatch score against G4Hunter score were plotted using seaborn.

Since positive training examples were outweighed in the dataset by a factor of 10:1, random under-sampling of negative examples was conducted using imblanced-learn to attain a ratio of 2:1. This filtered dataset was shuffled and written to disk in bed format, and bedtools getfasta was used to extract sequences for each interval from hg19. Sequences were then one hot encoded and loaded into HDF5 format for training using h5py. For training of models on trinucleotide content, trinucleotide content statistics were extracted and loaded into HDF5 format.

Model Training and Validation:

All models were trained in Python using Keras with TensorFlow backend. The trinucleotide Multi-Layer Perceptron model contained three hidden layers with 16 units per layer. These were trained using the ADAM optimiser and binary crossentropy loss function, with a dropout rate of 0.2 on all layers. The convolutional portion of G4Seeqer was made up of two convolutional layers with 8 filters and kernel size of 3 and ReLu activation, followed by a maximum pooling layer with step size of 2. This was connected to a bidirectional Long Short Term Memory layer with 8 units and Tanh activation. The final hidden layer was a fully connected layer with 16 units, ReLu activation and a dropout rate of 0.5. G4Seeqer was trained using the RMSprop optimiser and binary crossentropy loss function.

All models were trained on 80% of the training data with 10% used for validation. Training was conducted for a maximum of 30 epochs, but with early stopping when the change in validaton loss was less than 0.0005 for more than 3 Epochs. The Trinucleotide MLP model converged to this minimum change after 8 Epochs, whilst G4Seeqer converged after 15 Epochs.

Models were validated on 10% of the total data held out for testing purposes. Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves were generated using scikit-learn and plotted with matplotlib. ROC/PR curves for G4Hunter are produced using the modified method. For comparison against Quadron, the Quadron source code was

downloaded from GitHub and installed. Since the flanking sequences required for Quadron are longer than those used for G4Seeqer, test set sequences were increased in size by 50bp in each direction to 228bp. Sequences were extracted using `bedtools getfasta` and run through Quadron. For intervals which contained multiple Quadron scoring motifs, the highest score was used. For intervals which had no motifs scored by Quadron, a score of zero was assigned.

BG4 Analysis:

NarrowPeak BED files of BG4 ChIP-seq peaks were downloaded from GEO (GSE76688, Hänsel-Hertsch et al 2016). To accomodate Quadron's flanking sequence requirements, the size of the BG4 intervals was increased by 50bp in each direction using `awk`. A BG4-negative peak set was generated using `bedtools shuffle`. Shuffling was done excluding gaps in the genome or BG4-positive peaks. Positive and negative peaks were concatenated and sequences were extracted using `bedtools getfasta`. Predictions were made on these sequences using G4Seeqer/G4Hunter/Quadron, and the maximum scoring interval per peak was assigned as the overall score of the peak. Where a model did not make any predictions in a peak, it was assigned a score of zero. Receiver Operator Characteristic (ROC) and Precision Recall (PR) curves were generated using `scikit-learn` and plotted with `matplotlib`.

rG4seq Training Data Preprocessing:

To produce training data for rG4seeqer, G4hunter windows were predicted in hg19 using a window size of 50bp and a threshold of 0.75. These were intersected with human exons using `bedtools` to get a set of 186279 putative RNA G4 forming sequences. RNA G4s identified by rG4seq in the presence of potassium (Kwok et al. 2016) were downloaded from GSE77282 (`GSE77282_K_hits.bed.gz`), and intersected with the G4hunter windows to identify RNA G4 positive and negative examples. Of the 3383 identified RNA G4s in the rG4seq dataset, 2811 (83%) overlapped with G4hunter windows. Since the ratio of negative to positive examples was extremely high, negative examples were undersampled to a ratio of 2:1

using `imbalanced-learn`. These were then shuffled and written to disk in bed format, and `bedtools getfasta` was used to extract sequences for each interval from hg19. Sequences were then one hot encoded and loaded into HDF5 format for training using `h5py`.

rG4seq Transfer Learning:

Model weights trained on the G4Seq dataset were reloaded in the same architechture for training on the rG4seq dataset. Weights from the initial convolutional layers were fixed (i.e. made un-trainable), and only LSTM weights and final dense weights were trained. The model was trained on 6014 samples (80%), validated on 752 samples (10%), and tested on 752 samples (10%). Training was conducted as for G4Seeqer, but for a maximum of 200 epochs, with early stopping after 25 epochs if validation loss did not improve. Initial learning rate was set at 0.001 but reduced by a factor of $1/3$ after 15 epochs when no reduction in validation loss was seen.

Comparison to G4RNA Screener:

Supplementary data containing 368 RNA sequences, their experimentally determined G4 forming status, and their predicted score from G4RNA Screener were downloaded from Garant et al. 2017. Sequences greater than 128bp were filtered to give a total of 347 sequences. Average rG4Seeqer and G4Seeqer predictions for these sequences was made by generating 1000 randomly paddings for each sequence, one hot encoding, and performing forward pass through the network. G4RNA Screener scores used in the ROC curve were taken directly from the supplemental material.

Mutation Mapping analysis:

Mutation mapping was applied to human promoter regions from the ENSEMBL regulatory build, which was originally generated using ChromHMM. Promoter sequences were extracted from hg38. Mutation mapping was implemented as in Alipanahi et al. 2015: candidate

sequences were edited at each position to each nucleotide, and the resulting sequences were scored for G4 formation using G4Seeqer. Heatmaps were generated in Python using `seaborn`.

G-Triplex and Hairpin analyses:

G-triplex motifs were predicted in the hg19 genome using an in house script, using the pattern $G_X N_{1-4} G_X N_{1-4} G_X$ where $3 \leqslant X \leqslant 6$. Candidate triplexes which overlapped with or were contained within Quadparser motifs (pattern $G_X N_{1-7} G_X N_{1-7} G_X N_{1-7} G_X$ where $3 \leqslant X \leqslant 6$) were subtracted to produce only triplex motifs which could not form "classical" G4s. To assign mismatch scores to these sequences, they were increased in size by 50bp in each direction using `bedtools slop`, and mismatch scores from the G4Seq dataset were mapped using `bedtools map`. Distances to next G-run were measured using Python scripts.

G-hairpin motifs were predicted in the hg19 genome using the same script, and the pattern $G_X N_{1-4} G_X$ where $4 \leqslant X \leqslant 6$. Candidate hairpins which overlapped with or were contained within Quadparser motifs or G-triplex motifs were filtered. Intervals were increased in size by 50bp in each direction using `bedtools slop`, and mismatch scores from the G4Seq dataset were mapped using `bedtools map`. Distances to next G-hairpin were measured using `bedtools closest`. Triplex and hairpin histograms and boxplots were generated in Python using `matplotlib` and `seaborn`.

Median model score experiments:

For G4Seeqer scoring of experimentally validated G4s from Guédin et al 2010., sequences were padded using randomly generated sequences to 128bp in length. Left and right padding lengths were also varied at random. 1000 randomly padded sequences were generated and scored per input sequence. Scatter plots of UV melting temperature vs. median G4Seeqer score were produced using `matplotlib`. Errorbars are 68% confidence intervals.

Synthetic sequences used for loop length and G-register experiments were 3 tetrad Quadparser conforming sequences. Each sample contained 5000 randomly generated sequences. Left

and right padding sizes and nucleotide contents were varied within random samples. Loop nucleotide contents were also varied. For G-register experiments, the extra G per run as randomly assigned to either the left of right side of the G-run. Loop lengths of 3 were used. Line plots for loop length experiments were generated using `matplotlib`, and errors are 68% confidence intervals. Boxplots were generated using `seaborn`.

G-register Experiments:

For G-register mismatch experiments, 3 tetrad Quadparser conforming G4s from the hg19 genome were identified and the corresponding mismatch score was extracted using `bedtools map`. The number of tetrads with G-register was counted. Boxplots were generated using seaborn.

Human and Mouse G4 Subpopulation Analyses:

Example Human and Mouse G4 populations were predicted in hg19 and mm10 using the Quadparser pattern with loop lengths of 3. All possible G4s conforming to this pattern were generated using python `itertools`. Venn diagrams were generated using `matplotlib_venn`. P-values were produced using hypergeometric tests. Dinucleotide complexity was defined as the total number of unique dinucleotides contained in the motif. Histograms and kernel density estimate plots were produced using `seaborn`. For visualisation of PG4 distributions, a sample of 50000 motifs were randomly selected from the total population of all possible Quadparser motifs with loop lengths of three. These were transformed into two components using UMAP dimensionality reduction, with Hamming distance as the distance metric. Sequences which appear in the human and mouse genome were extracted from the sample. 2D Kernel Density Estimate plots for the full sample, and hg19 and mm10 subsets, were generated using `seaborn`.

Results and Discussion

Candidate G4 proposal

One major drawback to any machine learning method for G4 prediction over existing regular expression or pattern based methods is the relative expense of computation. It is therefore not sensible to train and classify a neural network model on all possible input sequences from a genome. Instead we decided to use an existing method, the G4hunter algorithm proposed by Bedrat et al., to produce candidate regions which could then be labelled as true positive G4s or non-G4s, and used as input for model training.

We reimplemented the G4hunter algorithm with some minor modifications. Bedrat et al's method run length encodes the sequence of interest and scored G-runs as the square of the run length, and C-runs as the negative of the square of the run length. These scores are then summed to give an overall score for the sequence. This method was chosen as it was assumed, based on earlier work, that a high G-content on both strands would make the G-Quadruplex unfavourable compared to double stranded DNA. Since we wished to make as few assumptions as possible, and given that the G4seq dataset stems from G4 formation in *in vitro* single stranded DNA, we altered the method to produce two scores. G runs score positively on the positive strand and C runs score positively on the negative strand. This means the sequence d(GGGCCC) would yield a high score on both strands rather than a single score of zero.

To produce candidate regions for model training, we ran the modified G4hunter method on the human genome (hg19) using a window size of 50bp, a step size of 5bp and a threshold of 0.75. Unstringent values were chosen to produce a high recall, i.e. capture as many true positive G4 structures as possible. These settings produced intervals which overlapped with all PG4 sequences predicted by the Quadparser method using maximum loop lengths of 12. It also produced significantly more sequences that did not conform to the Quadparser method, some of which are likely to form G4 structures.

Training Data preprocessing

To created training sequences, the 50bp candidate regions from the G4hunter method were increased in size by 39bp in each direction to produce intervals of 128bp in length, since previous work has suggested that flanking regions are an important determinant of G4 stability. Clusters of overlapping intervals were filtered to produce only the interval with the highest G4Hunter score (in cases of ties, a random highest scoring interval was selected). This produced a total of 6237943 candidate intervals. Each candidate interval was then scored by mapping the value of the maximum scoring overlapping window from the G4seq dataset (Chambers et al. 2015), which contains percentage mismatches (%mm) from sequencing in the presence of potassium vs. absence of potassium, in 12bp windows. Regions of high %mm on the positive strand indicate a G4 structure on the negative strand, and vice versa.

Plotting the distribution of the log of the %mm scores produced a bimodal distribution with a peak around 1 (corresponding to 2-3% mismatch) and another around 3.5 (corresponding to approximately 30% mismatch) (Fig. 2.1a). We determined the local minimum in the histogram between the two peaks to be approximately 3 (around 20% mismatch), therefore we used this value to split the data into G4 positive and G4 negative subsets. This yielded more than 10 times more G4 negative sequences than positive, however (5809719 negative to 428224 positive). Since maintaining such an imbalance in the training data would produce a poor classifier, we undersampled the G4 negative class to a ratio of 2:1, yielding a total of 1284672 sequences.

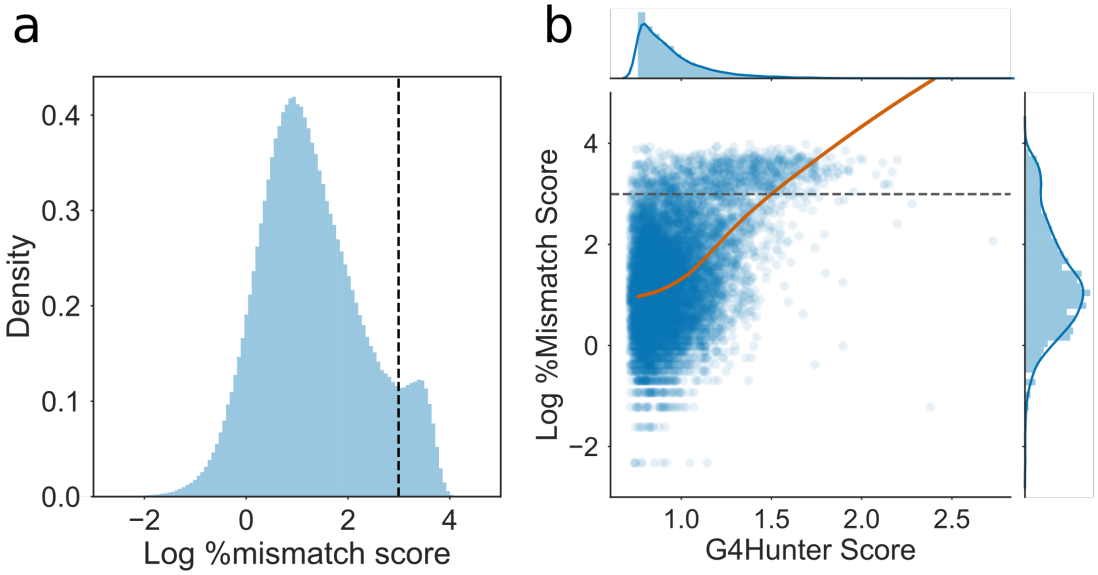

Figure 2.1: Mismatch scores of candidate sequences identified by G4Hunter: a) Histogram of log percentage mismatch score from the G4Seq dataset, for the 50bp sequences identified by G4Hunter (threshold of 0.75). Dashed line shows the threshold chosen to delimit G4 positive and G4 negative sequences. This corresponded to around 20% mismatch score. **b)** Joint plot of log mismatch score against G4Hunter score for 10000 randomly sampled sequences. Orange line shows lowess curve fit.

Model selection and training

Previously published G quadruplex prediction methods which utilise machine learning techniques (Quadron, G4RNA) have used derived features such as trinucleotide content to feed to models. These features result in the loss of some spatial information about the sequence, however. For example, the sequence GGTGGTGGTGGGGGG has the same trinucleotide composition as GGGTGGGTGGGTGGG, but is unlikely to have equivalent G4 forming propensity. Furthermore, Quadron derived features require input sequences to conform to the QuadParser regular expression, meaning that Quadron is only able to improve the precision of the QuadParser method, and not the recall. We opted for a neural network involving convolutional layers (those often used for image classification) that could make predictions directly from the sequence itself, without any derived features whatsoever. This allows us to make no assumptions about potential G4 patterns in the dataset. The overall architechture selected was a convolutional-recurrent neural network (Fig. 2.2), which has previously been used to identify regulatory motifs in DNA (DanQ paper). The architechture consists of two one dimensional convolutional layers with a kernel size of 3, to capture local features in the sequence. A maximum pooling layer then reduces the size of the feature space by half. These features are then fed to a bidirectional Long Short Term Memory (LSTM) layer, which is able to learn and recognise long distance relationships between features in the sequence. The model outputs a single value between zero and one of the probability of G4 formation. The dataset was split into three for training and testing: 1332565 sequences (~80%) were used for training, 166571 (~10%) for in-training model validation, and 166576 (~10%) for post-training model testing.

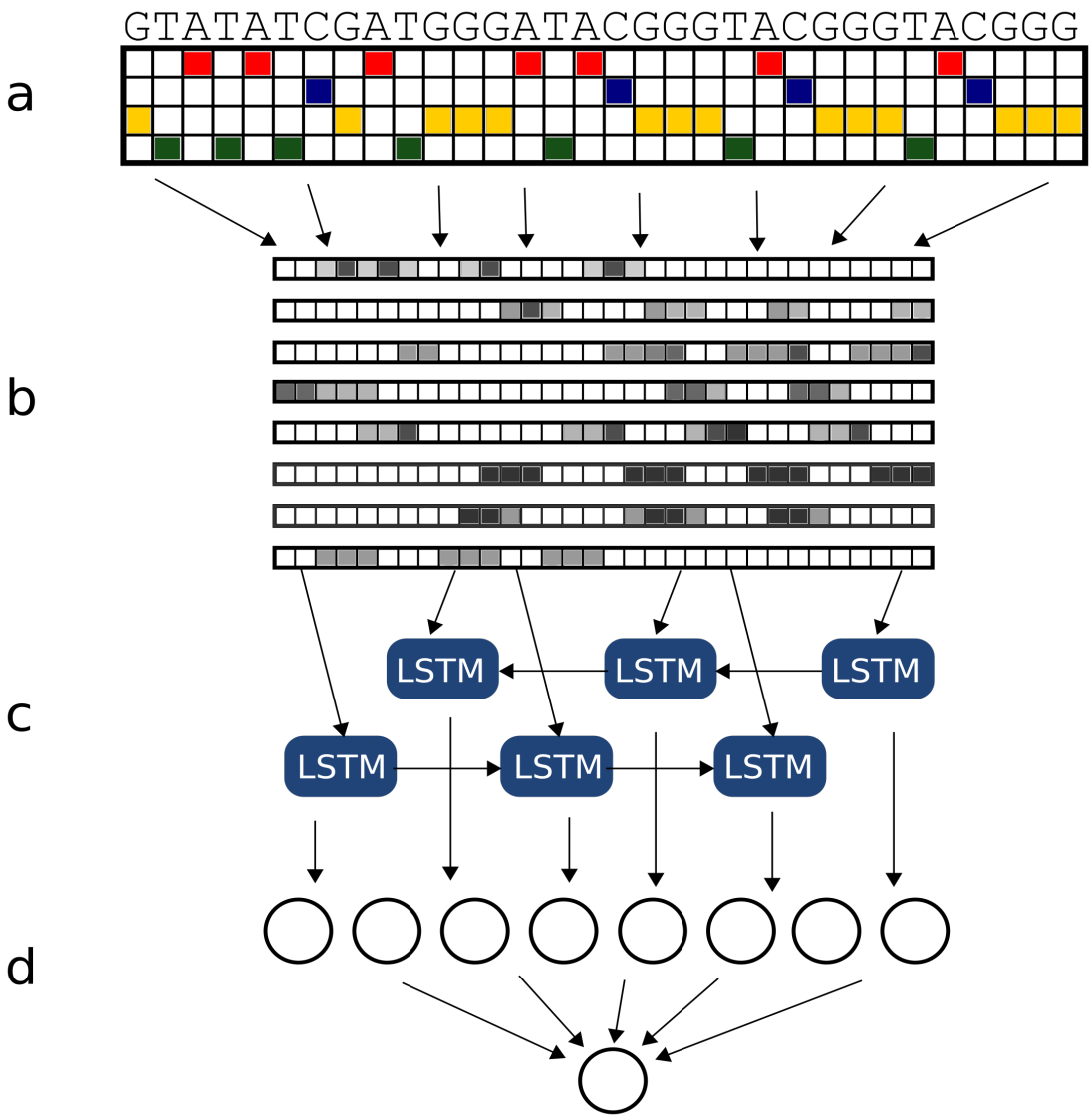

Figure 2.2: G4Seeqer architecture Adapted from Quang & Xie 2016. **a)** Sequences are one hot encoded to produce a matrix which can be processed by the neural network. **b)** Input matrices are passed through a convolutional layer. Each layer contains 8 filters which are trained to recognise local patterns on the scale of 3-6bp in size. **c)** Convolutional features are passed through a bidirectional Long Short Term Memory (LSTM) layer. This layer recognises long distance interactions between features which might combine to produce G4s. **d)** Finally, features are passed through a fully connected layer. Output from the model is a single probability of whether the sequence forms a G4.

Comparison to existing methods:

We benchmarked our technique (hereafter referred to as G4Seeqer) using the 10% of the data reserved for testing. The model was compared to our modified G4Hunter method, as well as a multi-layer perceptron model trained on trinucleotide frequencies derived from the same dataset as was used to train G4Seeqer. This model allows us to compare the methodology of G4RNA Screener (Garant et al. 2017) to our own method, since G4RNA Screener was originally trained on a database of RNA G Quadruplexes, and may not perform as well on a dataset derived from DNA. The performance of the methods were calcuated using the Receiver Operating Characteristic area under curve (ROC AUC). We found that neither G4Hunter nor the G4RNA-like method performed as well as G4Seeqer on the test dataset (AUC G4Hunter 0.82, G4RNA Screener-like 0.90, G4Seeqer 0.94) (Fig. 2.3a-b). This is likely due to the loss of sequence spatial information in the former methods.

To benchmark our method against the other G4Seq trained machine learning G4 prediction package, Quadron, we downloaded and installed the Quadron source code from Sahakyan et al. 2017. Quadron was used to score sequences from the held out test set and compared to the performance of other methods. Quadron requires larger flanking regions of 50bp, so test set intervals and sequences were increased in length by 50bp in both directions to produce test sequences of length 228. This was deemed the best way to compare the two methods, however was still not ideal since Quadron may have been trained on some or all of the test sequences. Regions which had predictions associated with them by G4Hunter and G4Seeqer but had no associated predictions from Quadron (due to not conforming to the Quadparser regular expression) were given a Quadron score of zero. The ROC curve of Quadron (AUC 0.7) ((Fig. 2.3a-b)) had interesting properties: the model is capable of producing a true positive rate of around 30% with a false positive rate of less than 1%, however shortly beyond this point in the ROC curve the curve becomes linear. This is because Quadron is only capable of scoring sequences which conform to the Quadparser method, with all other sequences being scored zero. Since these sequences only account for a 25% of all the potential PG4 forming sequences in the test set, the model does not perform well on the full dataset. Because it makes no

assumptions about the input sequence, G4Seeqer is able to make accurate predictions for all forms of PG4 in the test data, as reflected in the ROC curve.

We were interested in how G4Seeqer compared against Quadron on only sequences which conformed to the Quadparser motif. We therefore filtered our dataset for intervals on which Quadron had made a prediction, and replotted the ROC curves and Precision Recall curves for the filtered data (Fig. 2.3c-d). As expected the AUC for Quadron was much better on this filtered set (AUC 0.93), however it was still outperformed by G4Seeqer (AUC 0.94), suggesting that G4Seeqer captures the same or more explanatory information directly from the input sequence.

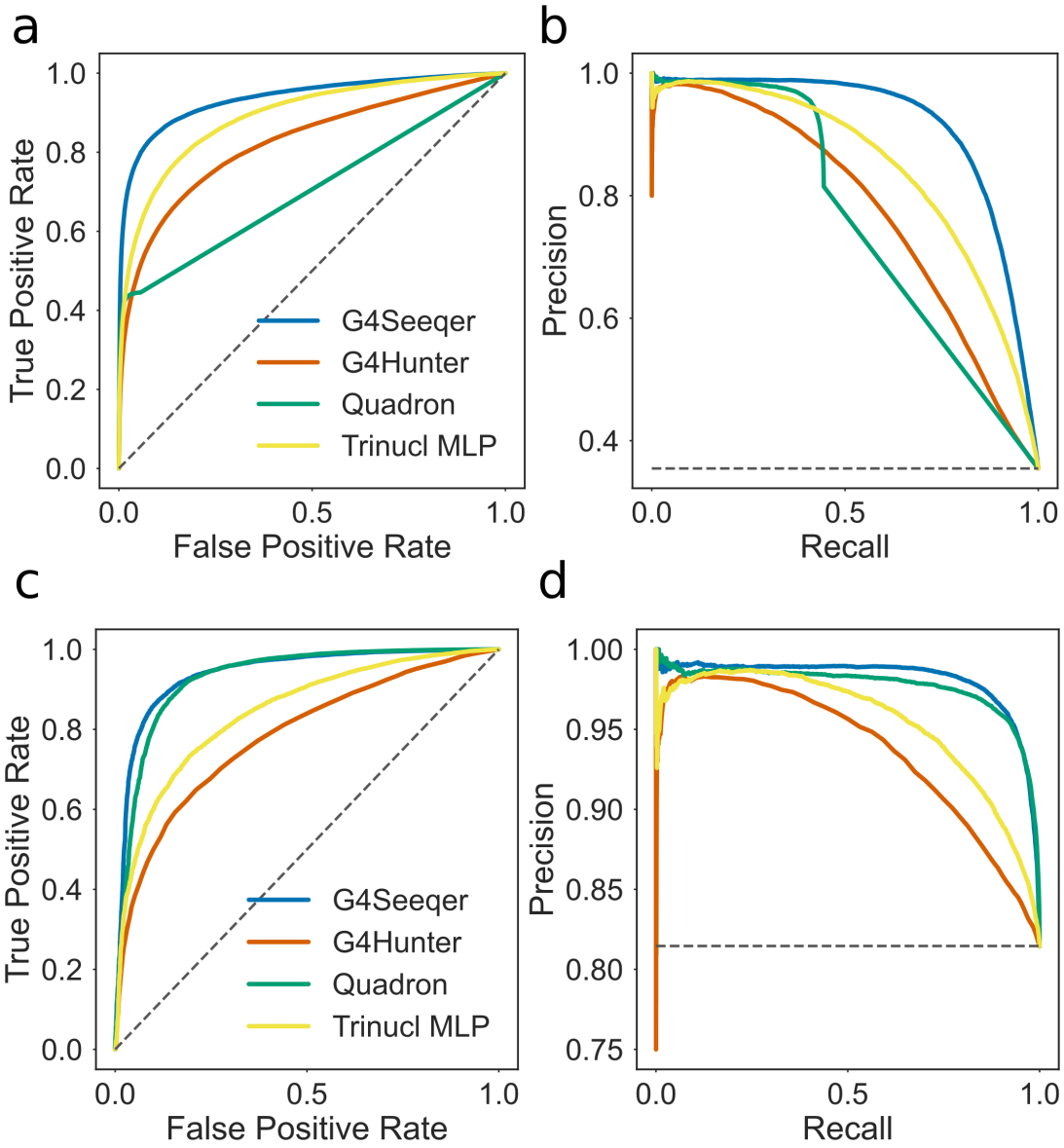

Figure 2.3: Validation curves for G4Seeqer method a) Receiver Operator Characteristic (ROC) curve showing the performance of G4Seeqer, Multi-Layer Perceptron (MLP) trained on trinucleotide contents, Quadron, a Gradient Boosted Machine model (Sahakyan et al. 2017) and the G4Hunter method (Bedrat et al 2016), on a held out test set of the G4Seq dataset (10% of total dataset). **b)** Precision-recall curves showing the performance of G4Seeqer, trinucleotide MLP, Quadron, and G4Hunter on the same dataset. **c)** ROC curve and **d)** Precision Recall curve showing the performance on sequences from the test set conforming to the Quadparser motif.

BG4 ChIP-seq data evaluation:

Further model validation was performed on G4s experimentally validated by an entirely different technique, namely G4-chromatin immunoprecipitation (BG4) (Hänsel-Hertsch et al 2016). The BG4 dataset is arguably more biologically relevant than G4seq since G4 structures are not induced by addition of potassium, and are captured from native chromatin. BG4 peak intervals were shuffled to produce a set of G4 negative sequences, and then the performance of the models was evaluated on the real and shuffled peaks. For each BG4 interval, the highest scoring overlapping prediction for each model was assigned. Any intervals with no overlapping predictions were scored zero for that model. We tested the G4Seeqer, G4Hunter and Quadron methods. As with the G4Seq test dataset, we found that Quadron performed reasonably for Quadparser conforming BG4 peaks, but was unable to identify most of the true positive BG4 peaks due to its restriction to the pattern. G4Seeqer performed better on all BG4 peaks, with an AUC of 0.71, however was only marginally better than the G4Hunter technique (AUC 0.7) (Fig. 2.4). These results suggest that the information within the G4Seq dataset, when captured by a suitable model, is predictive of G4s in an *in vivo* setting.

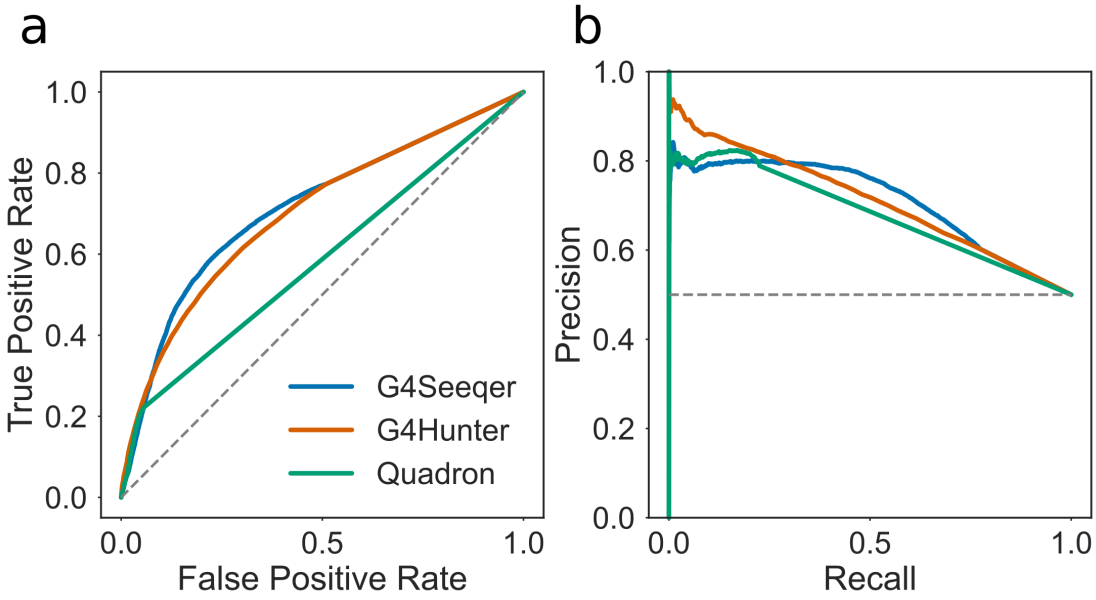

Figure 2.4: Detection by BG4 peak sequences using G4Seeqer a) Receiver Operator Characteristic (ROC) curve showing the performance of G4Seeqer, Quadron, and the G4Hunter method, on BG4 ChIP-seq peaks and randomly shuffled negative sequences.

Transfer learning on RNA G4Seq (rG4seq) Dataset:

PG4s with the same sequence are likely to have slightly different G4 forming potentials in DNA and RNA, due to the chemical differences in these molecules. The sugars which make up the backbone of RNA are riboses, which have an extra 2' hydroxyl group compared to the deoxyribose found in DNA. This extra hydroxyl group is thought to have a number of implications for G4 formation: it increases the number of backbone hydrogen bonds in the G4, increasing its enthalpic favourability and its entropic favourability (by reducing the number of coordinated water molecules). Furthermore, the 2' hydroxyl introduces steric constraints which make parallel RNA G4s much more favourable than anti-parallel ones. Given these differences, it is likely therefore that a model specifically trained on DNA G4 sequences will not perform optimally on RNA G4s.

To address this issue, we decided to retrain G4Seeqer using the rG4seq dataset produced by Kwok et al. 2016, to create an rG4Seeqer model. Data was prepared similarly to the data for G4Seeqer: candidate regions were selected from human exonic sequences using G4hunter with window size of 50bp and a threshold of 0.75. This yielded a set of 186279 putative RNA G4 forming sequences, which were increased by 39bp in each direction to get flanking sequences. These were then intersected with rG4seq hits collected under potassium stabilising conditions (Kwok et al. 2016). Of the 3383 identified RNA G4s in the rG4seq dataset, 2811 (83%) overlapped with G4hunter windows. rG4seq negative examples were undersampled with a ratio of 2:1 to yield 7518 training samples. 80% of these were used for training, with 10% for validation and 10% held out for testing.

Because of the significantly smaller size of the rG4seq derived training set, we found that the method for training which yielded most optimal results was transfer learning from the G4Seeqer model. Weights of the initial convolutional feature extraction layers were therefore held constant, and only the weights of the LSTM layers (which find long range interactions) and dense output layers were retrained.

Testing was first conducted on the held out set of 752 sequences using G4Hunter, G4Seeqer and the newly trained rG4Seeqer. G4Seeqer significantly outperformed G4Hunter (AUC 0.9

vs 0.83 respectively), suggesting that the information extracted from the G4Seq dataset is applicable to the rG4seq dataset (Figure 2.6). rG4Seeqer outperformed both methods, however (AUC 0.95), demonstrating that domain specific information is better for predicting RNA G4s in the rG4seq dataset (Figure 2.6).

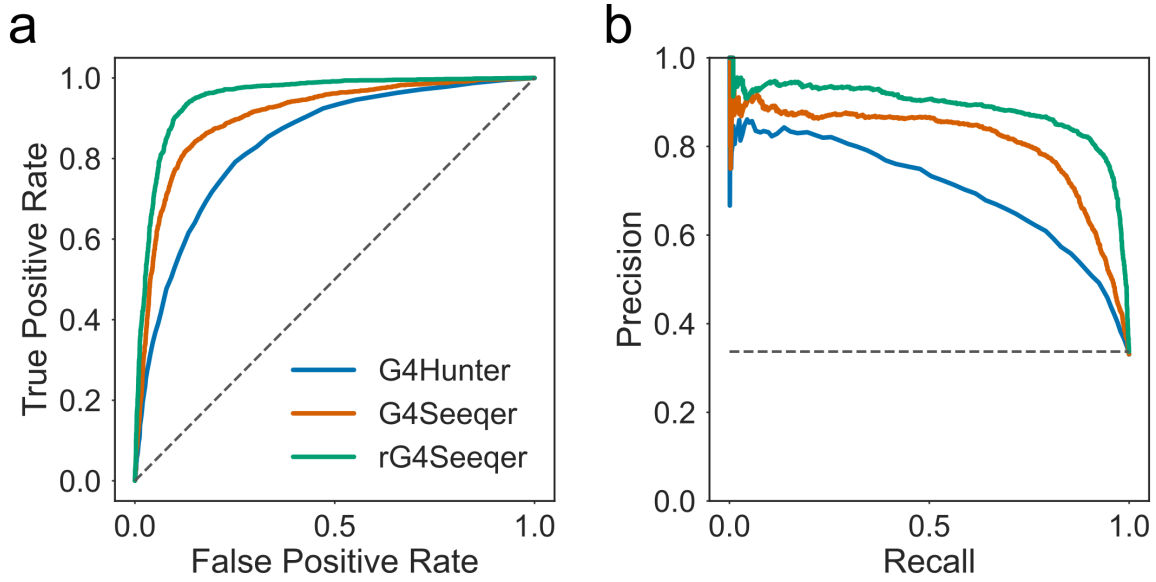

Figure 2.5: Validation curves for rG4Seeqer method a) Receiver Operator Characteristic (ROC) curves showing the performance of rG4Seeqer, G4Seeqer and the G4Hunter method (Bedrat et al 2016), on a held out test set of the rG4Seq dataset (10% of total dataset). **b)** Precision-recall curves showing the performance of rG4Seeqer, G4Seeqer, and G4Hunter on the same dataset.

We sought to test rG4Seeqer on G4s identified by a variety of physical methods, using the set of RNA G4s curated by Garant et al. for their model, G4RNA Screener (Garant et al. 2017). We used 347 sequences from this dataset, of which 169 sequences were G4 postive and 178 were G4 negative. G4Seeqer and rG4Seeqer predictions were calculated by padding with random sequences to a length of 128bp before one hot encoding. This was conducted 1000 times for each sequence and the mean score was taken. G4RNA screener scores were taken directly from the supplemental information of Garant et al. 2017. G4RNA screener was found to perform best on the dataset (AUC 0.91), perhaps unsuprisingly since it was trained directly on the sequences. Perhaps more importantly, rG4Seeqer significantly outperformed G4Seeqer on the dataset (AUC 0.89 vs 0.82), showing that the rG4seq trained model generalises better to RNA G4 sequences than the G4seq trained model.

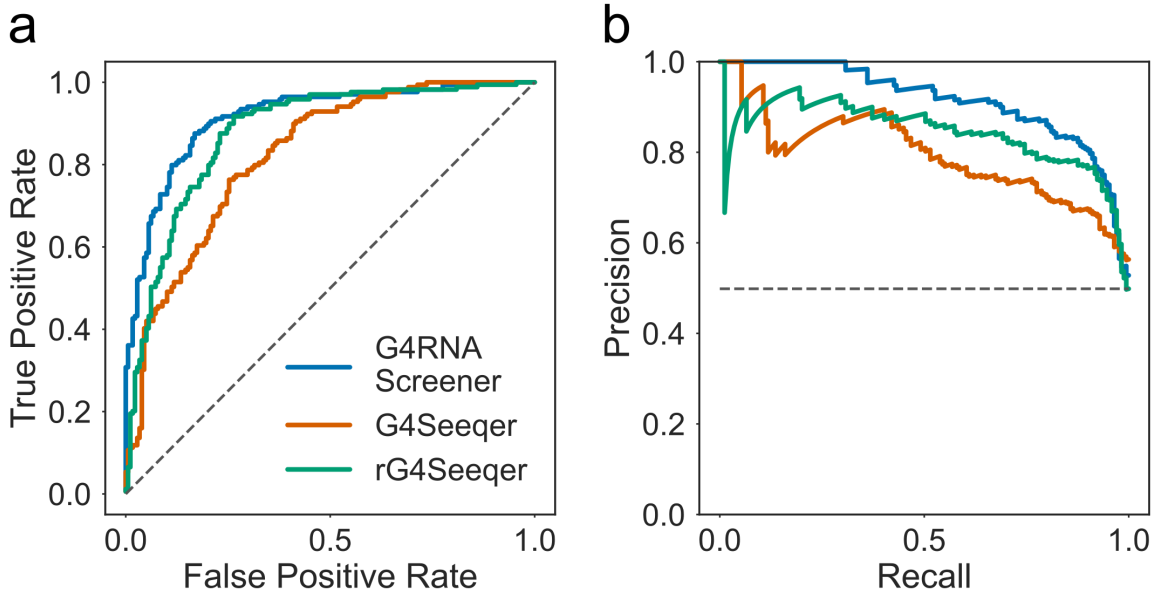

Figure 2.6: Validation of rG4Seeqer on *in vitro* experimentally categorised RNA sequences a) Receiver Operator Characteristic (ROC) curves showing the performance of rG4Seeqer, G4Seeqer and the G4RNA Screener method (Garant et al 2017), on the G4RNA dataset curated by Garant etl al. **b)** Precision-recall curves showing the performance of rG4Seeqer, G4Seeqer, and G4RNA Screener on the same dataset.

Interpreting G4Seeqer output using Mutation Mapping:

One common complaint about machine learning techniques is that the complexity of the models they produce make them "black boxes" which are impossible for humans to understand or extract useful knowledge or rules from. It is possible, however, to visualise some of the output of a neural network through various means. One commonly used method for interpretation is the "saliency" of the network, which can produce heatmaps showing the attention of the network to specific regions of the input image or sequence. This can be used to determine the important aspects of the input in classification. For biological sequences, previous studies have used similar methods, called "Mutation Maps", to analyse the importance of individual nucleotide positions on CNN model predictions. Simply, the importance of each particular nucleotide is evaluated by replacing it with each of the other three bases, and calculating the change in model score. This is then used to build a heatmap which can be used to visualise the importance of each position.

Previous studies have highlighted a possible role for G4 forming sequences in promoters, with G4 formation tending to have a positive effect on expression, particularly in proto-oncogenes (Eddy & Maizels 2006, Hänsel-Hertsch et al. 2016). We therefore decided to use the mutuation map approach to characterise PG4s in promoter regions extracted from the ENSEMBL regulatory build. Promoter sequences were screened using G4Seeqer and single base mutation maps were created for each candidate PG4, including those regions where the neural network score was low. Unsurprisingly, all of the most deleterious single base substitutions (causing a score reduction of more than 0.9) predicted by G4Seeqer mutation mapping were G->H changes.

We identified PG4 sequences scoring more than 0.9 for which a single G->H change resulted in a reduction of as much as 0.9 in score (i.e. switched the score from strongly PG4 positive to strongly PG4 negative) (Fig. 2.7a). Analysis of the mutation maps for these sequences showed that the majority of contained regions containing three G-runs with short connecting loops. These tended to have a long final loop to the next homopolymeric G-run, or no final G-run within the window size. Any G->H mutation in these G-dense regions strongly affected

the predicted G4 forming ability. Recent work by Xi et al. has shown that formation of G4 structures in human telomeric sequences occurs via a stable G-triplex structure. We believe these results suggest that the G4seq dataset is either capturing mismatches caused by G-triplex structures, or by G-quadruplexes formed from short range G-triplex interaction with more long range single G-runs. To further illustrate this we predicted all short looped (1-4bp) G-triplexes in the human genome which did not overlap with a Quadparser predicted PG4 (loop lengths 1-12bp). We found that 35% of these were associated with a %mm score greater than 20%, suggesting the formation of some secondary structure (Fig. 2.7b). To see if medium range G-run interactions might stabilise these, we then measured, for each G-triplex, the distance to the next run of at least three Gs on the same strand, and compared this to the %mm score. We found a negative correlation between distance and %mm score (Spearmans rho -0.2) for G triplexes with a G-run less than 100bp away, suggesting that these longer range interactions do occur but become weaker with distance (Fig. 2.7c). This could be due to a reduction in G4 stability with loop length, but could equally be explained by a reduction in the liklihood of the next G-run being contained in the same sequenced fragment. No clear difference was observed in the correlation of %mm score with upstream or downstream G-runs (Spearmans rho -0.17 and -0.15 respectively) suggesting there is no preference for the long loop region to be at the 5' or 3' of the G triplex.

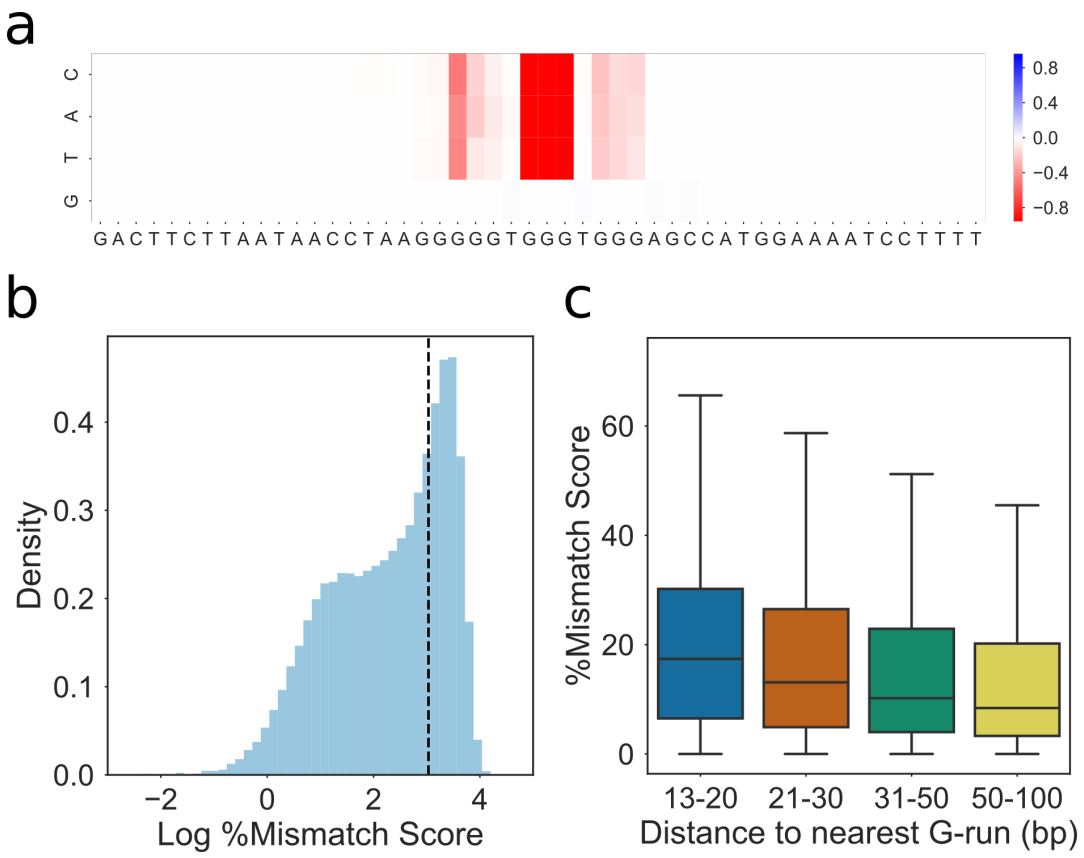

Figure 2.7: Identification of G-triplex structures by G4Seeqer Mutation Maps a) Mutation map showing the a high scoring (0.99) G4Seeqer motif which may form a G-triplex. Mutation of any base in the central G-run of the motif is sufficient to reduce the score by up to 80%. **b)** Histogram of log percentage mismatch score for motifs conforming to a G-triplex like pattern. The bimodal distribution suggests that many of these motifs form structures which disrupt polymerase in the presence of potassium. **c)** Boxplot showing the relationship between %mm score and distance to next G-run in G-triplex structures. The negative correlation suggests G-triplexes might recruit distant G-runs to form G4s.

We also noted that G4Seeqer was positively labelling certain sequences which contained only two G-runs, usually of greater than 4 bases in length (Fig. 2.8a). These scores were also very sensitive to G->H mutations. Based on the folding dynamics work by Xi et al, we hypothesised that these sequences might form G-hairpins, which could associate with other nearby hairpins to form G4 structures. To test this, G-hairpins with G-run lengths greater than four were predicted and filtered to remove any overlaps with predicted quadparser G4s (loop length 1-12) or G-triplexes (loop length 1-4). 27% of these sequences had %mm scores greater than 20% (Fig. 2.8b). For each predicted G-hairpin, the distance to the nearest G-hairpin was calculated and correlated with %mm. Again, distance was found to correlate negatively with %mm score (Spearmans rho -0.18), suggesting that these hairpins may associate with each other to form G4s (Fig. 2.8c).

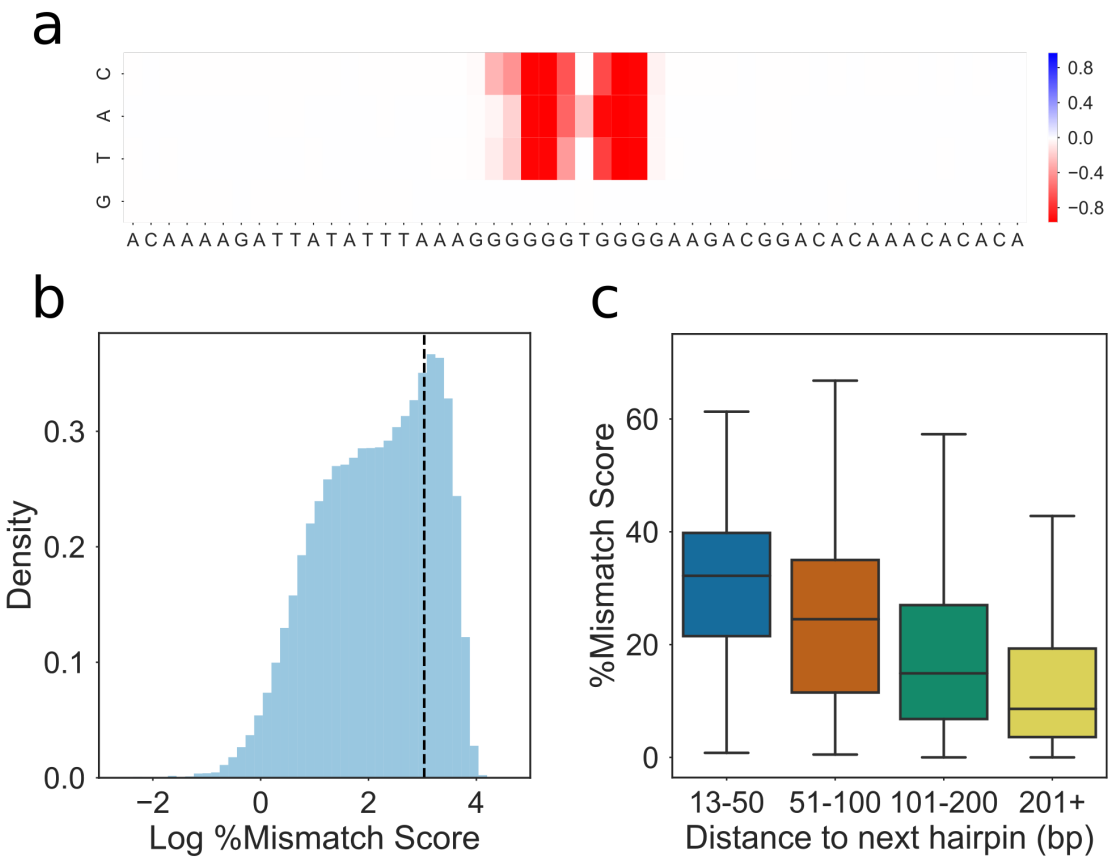

Figure 2.8: Identification of G-hairpin structures by G4Seeqer Mutation Maps a) Mutation map showing the a high scoring (0.99) G4Seeqer motif which may form a G-hairpin. Mutation of any base in the core motif is sufficient to reduce the score by up to 80%. **b)** Histogram of log percentage mismatch score for motifs conforming to a G-hairpin like pattern. The bimodal distribution suggests that many of these motifs form structures which disrupt polymerase in the presence of potassium. **c)** Boxplot showing the relationship between %mm score and distance to next G-hairpin for G-hairpin structures. The negative correlation suggests G-hairpins might interact with other relatively distant hairpins to form G4s.

In order to study how interactions between pairs mutations might affect predicted G4 stability, we developed a pairwise mutation map, in which pairs of Gs in each sequence were combinatorially mutated to Ts. We then analysed the resultant mutation maps to identify pairs of G->T transversions which interact to reduce predicted G4 forming potential more strongly than each individual mutation. Perhaps unsurprisingly, we found that in sequences which had more than four G-runs, or had G4s containing features which might form G-triplexes or G-hairpins, mutations which disrupted peripheral G-runs did not have a strong effect on predicted stability. Combinations of mutations which disrupt multiple G-runs had a much stronger effect on stability, however.

Loop Length and G4 stability.

To determine whether G4Seeqer probability scores supported previous work on the relationship between G4 loop length and stability, we first downloaded UV melting temperatures for three tetrad G4 sequences from Table 1 of Guédin et al. 2010. The majority of these G4 sequences contain only runs of Gs and Ts. In each experiment, one or two of the three T loops were held at a constant length of either 1 or 3, and the other loops varied from 1 up to 15 bp in length. For each sequence, we produced 1000 sequences padded to 128bp (the input length of G4Seeqer) using randomly generated bases, and used G4Seeqer to predict the stability. We found a very strong correlation (Spearman's rho 0.93, p = 1.1e-35) between empirically determined melting temperature in potassium, and G4Seeqer score (Fig. 2.9), suggesting that G4Seeqer is successfully capturing information about G4 structure which is transferable between conditions. The midpoint of the curve appears to suggest that a melting temperature of around 65oC is required for significant mismatches to occur in the G4Seq dataset. We also noted that G4seeqer output was more variable for sequences with lower melting temperatures, suggesting that sequence context may be more important for these G4s.

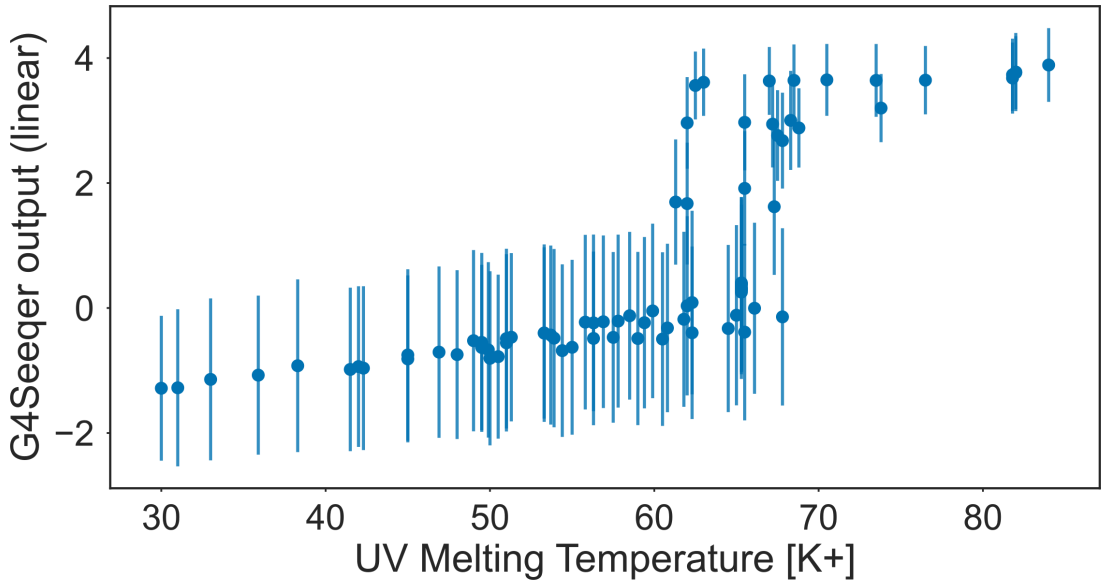

Figure 2.9: G4Seeqer scores correlate with experimentally determined melting temperatures Scatter plot showing median G4Seeqer scores vs. UV melting temperature for sequences from Guédin et al. 2010. Error bars are 68% confidence intervals generated from 1000 iterations of prediction with randomly generated flanking sequences.

We next performed a similar *in silico* experiment to that of Guédin et al., whereby we generated random QuadParser conforming G4 sequences with two loop lengths held at 1bp and the third varied from 1-60bp. Loop regions were constructed by randomly selecting from A C or T. The G4Seeqer score for these sequences was then generated. Unsurprisingly, we found that G4Seeqer score reduced with increasing loop length. This effect was strongest for the central loop of the G4, presumably because varying this loop has a greater effect on the ability to form stable G-triplex intermediates (Fig. 2.10a). We then set the length of the non-varying loops to 3bp and re-ran the analysis for third loop length 1-60bp. For these PG4s, we found that the probability of G4 prediction was much more sensitive to longer loop lengths (Fig \ref{loop_len)c). There was also no longer a strong difference in prediction when varying the central loop, compared to either loops 1 or 3, possibly suggesting that triplex formation in these sequences is less common.

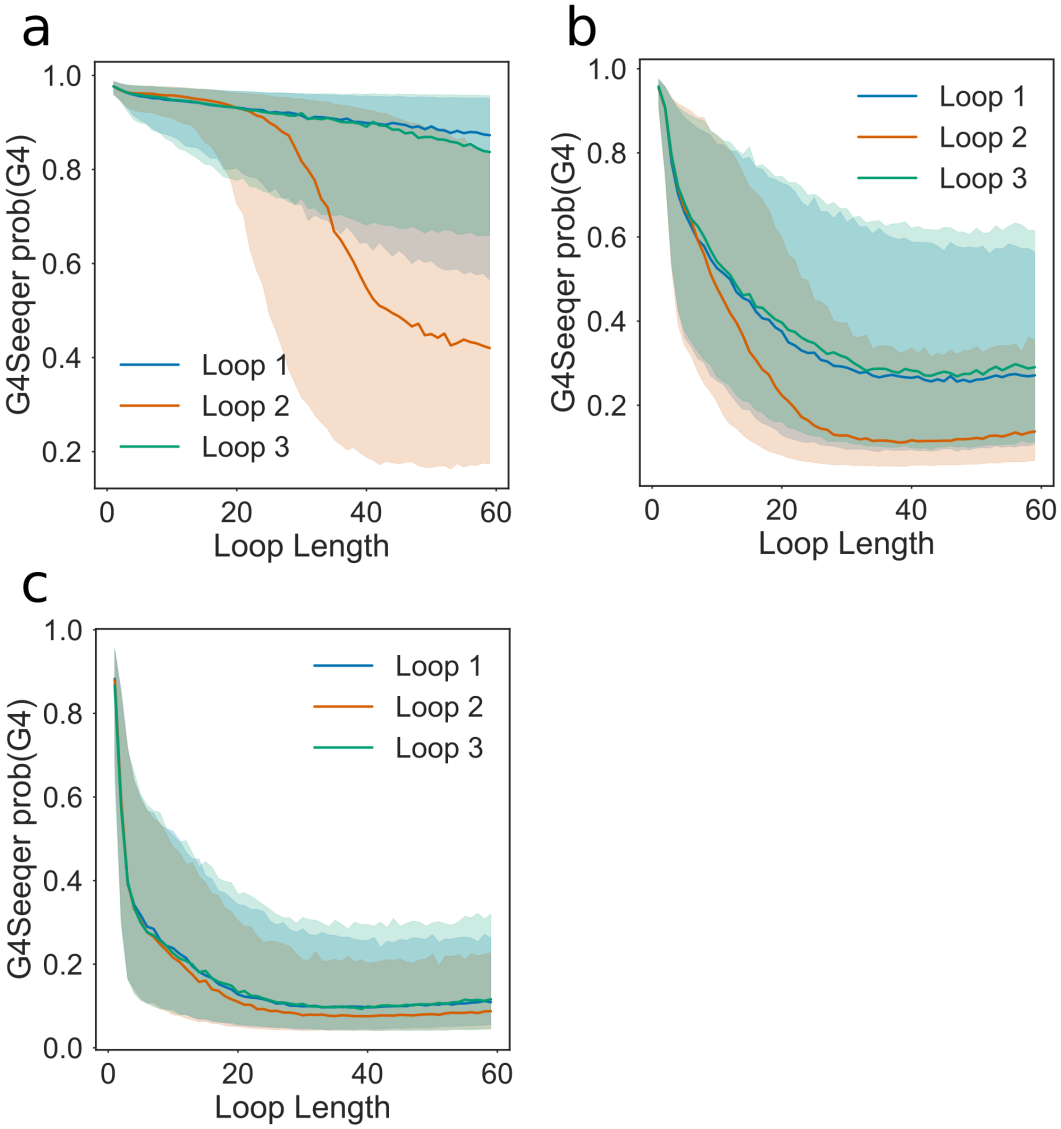

Figure 2.10: Effect of increasing loop length on G4Seeqer score a) Effect of loop length on predicted G4 stability when other loops are held at a constant length of **a)** 1bp, **b)** 2bp or **c)** 3bp. Median value and 68% confidence intervals are produced using 5000 randomly generated sequences (including flanking regions and loop contents) for each pattern analysed.

Effect of G-register on G4 stability.

Work by Harkness and Mittermaier has indicated that extra Guanines in some G-runs of a G4 forming sequence might increase the G4 forming potential of the sequence, by allowing exchange between different G4 conformations. They termed this effect G-register. We analysed the G4seq dataset and the G4Seeqer model outputs to determine whether there was evidence of a relationship between stability and G-register. Firstly, for all PG4s in the human genome conforming to the three tetrad Quadparser motif, we counted the number of G-runs which contained four Gs rather than three. These motifs were then intersected with the G4seq dataset to identify the %mm score. We found that G4seq motifs with greater G-register indeed tend to have higher mismatch scores (Fig. 2.11b). To test whether G4seeqer had successfully identified this pattern, we then randomly generated three tetrad PG4 sequences with loop lengths of 3bp, and introduced addition guanines to increase the G-register. Higher G-register strongly increased the G4Seeqer score of the sequence, showing that the model has successfully learned this feature of G4s (Fig. 2.11a).

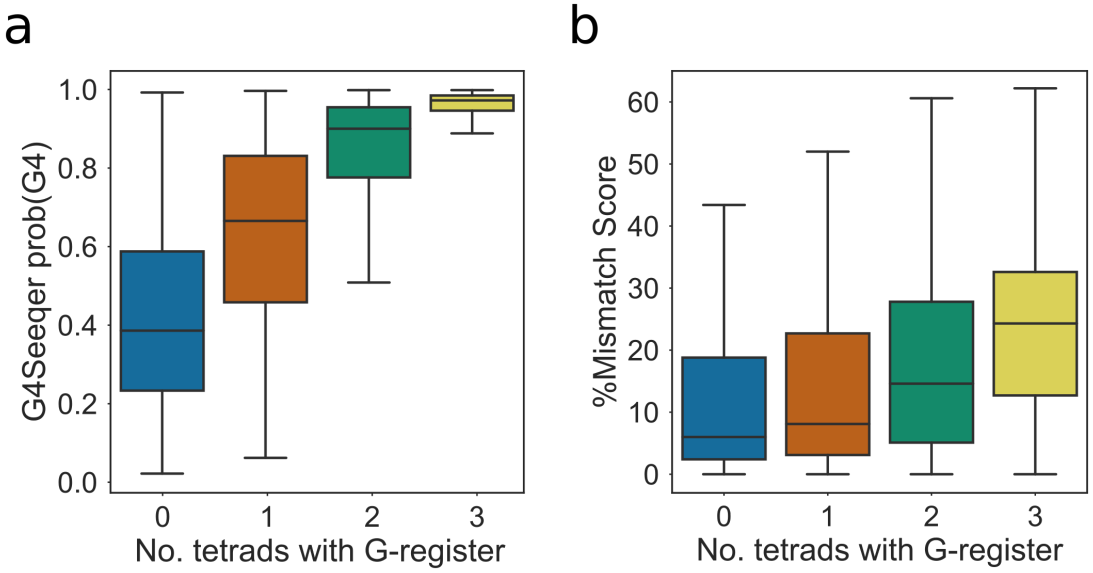

Figure 2.11: Scoring of G-register effects by G4Seeqer a) Boxplot showing G4Seeqer scores for randomly generated 3 tetrad Quadparser conforming sequences with zero to three additional Guanines per run, referred to by Harkness and Mittermaier as G-register. **b)** Boxplot showing relationship between G-register and %mm score in the G4Seq dataset for Quadparser conforming G4s.

Applicability of the model to other genomes.

Whilst the Human genome contains a large number of G4 forming sequences, this is not even close to saturating the population of all potential G4s. Indeed, simply considering the Quadparser motif with loop lengths up to 12, there are 1.1×10^{22} different conforming sequences, many orders of magnitude more than there are bases in the human genome. It is probable that the human genome contains a biased subpopulation of all G4s. This might mean that the G4Seeqer model does not generalise well to other genomes which contain different subpopulations. As an example, we analysed 3 tetrad Quadparser motifs from hg19, where all three loops were of length 3. There are 4^9 (262144) possible sequences fitting this description, of which only 1.4% (3748) appear in hg19 (Fig. 2.12a). The motifs that did appear tended to be those with lower complexity, as measured by the number of distinct dinucleotides in the sequence, than the total possible population (Fig. 2.12b). The PG4 space of the *M. musculus* genome was also measured, and found to contain 1.7% (4330) of all possible PG4s with loop length 3. There was a strong overlap of 27% ($p < 2.2e\text{-}308$) between sequences in the human and mouse genomes, however, suggesting that at least for this pattern, the PG4 populations of these genomes are comparable. The more complex the PG4 motifs become, however, the more likely it is that these subpopulations will be very different. There are clear differences in dinucleotide content between different genomes, which are often a result of differences in amino acid composition of proteins, or other environmental factors such as temperature. For genomes whose last common ancestor with *Homo sapiens* was longer ago, this divergence may be much greater. These systemic differences between may result in patterns to which the model has not previously been exposed, and reduce the performance of the model. G4Seeqer, or any other models which are trained on sequences from a single genome, should therefore be used with caution on others.

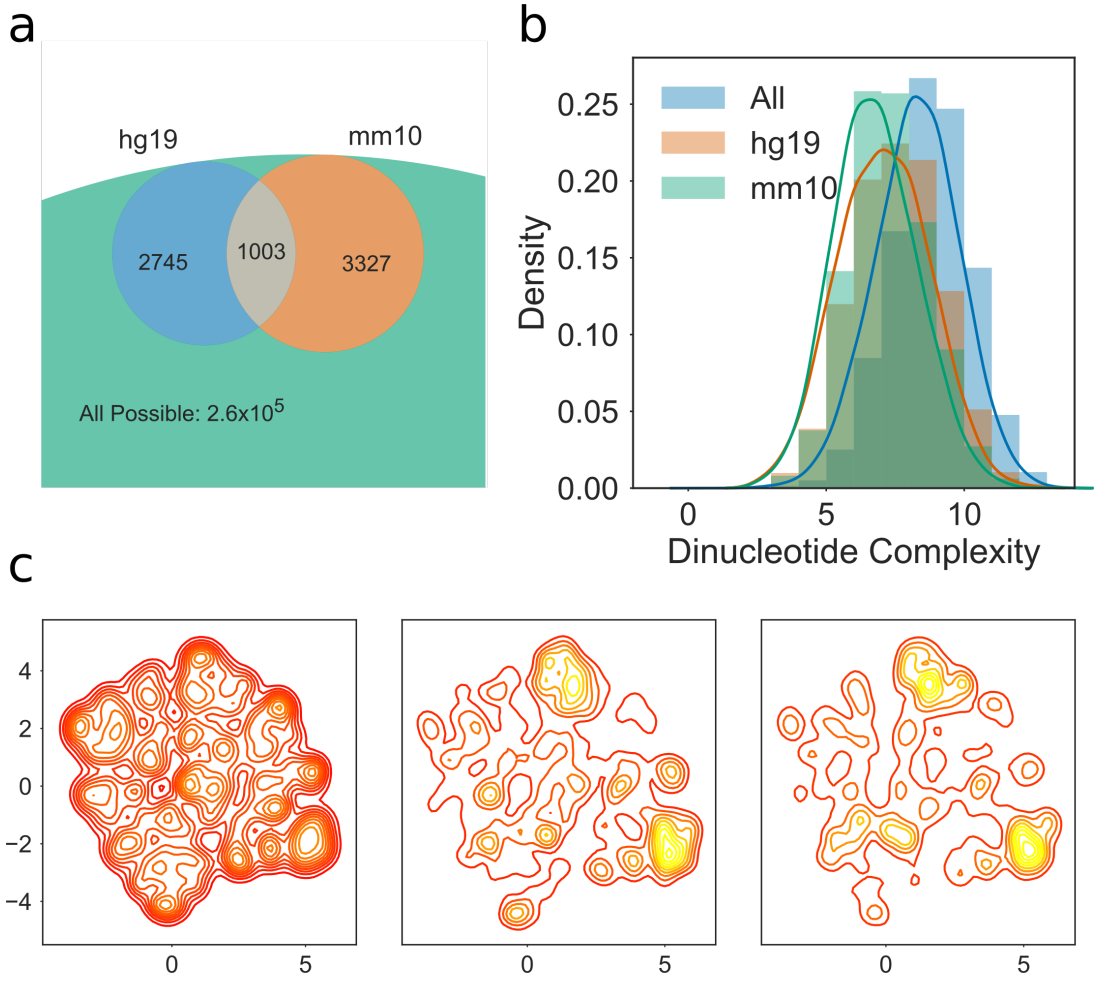

Figure 2.12: Human and mouse genomes contain different G4 subpopulations a) Venn diagram showing overlap of 3 tetrad Quadparser motifs populations with loop lengths of 3bp in the human (hg19) and mouse (mm10) genomes, compared to all possible sequences. **b)** Histogram and kernel density estimate of dinucleotide complexity for human, mouse and all possible 3 tetrad Quadparser motifs with loop length of 3bp. **c)** 2D Kernel density estimate plot showing distribution of all possible tetrad Quadparser motifs (left), those found in the human genome (centre) and those found in the mouse genome. Dimensionality reduction was conducted using UMAP with Hamming distance as the metric.

Conclusion:

We present G4Seeqer, the first Convolutional/Recurrent Neural Network model for prediction of G Quadruplex forming structures. G4Seeqer is implemented in Python, using Cython for speed-up of G4Hunter candidate region proposal, and Keras with Tensorflow backend for neural network prediction. It is able to process the whole human genome in approximately 1 hour on a 8 core i7 desktop computer with 16GB RAM. Because G4Seeqer is trained directly upon sequences from the human genome, rather than on derived sequence features, it is able to identify patterns in the G4Seq dataset that have not previously been reported, as well as removing false positive sequences which are flagged by pattern matching techniques. This greatly improves the accuracy of the model on various *in vitro* and *in vivo* datasets, from stabilities determined by UV melting to genomic regions identified by BG4 ChIP-seq.

Chapter 3

Global analysis of Predicted G-Quadruplexes in the *Arabidopsis thaliana* genome

Introduction

Arabidopsis thaliana is a member of the *Brassicaceae* family, which contains many important crop species. Whilst it is not of major agricultural importance itself, its short generation time and amenability to transformation has made it an extremely important model for genetic and genomic studies of plants. The genome of Arabidopsis is small, with various estimates putting its size between 125-150Mb (Arabidopsis Genome Initiative, 2000, Bennet et al. 2003). This is approximately 80Mb less than the closely related species *Arabidopsis lyrata*, suggesting a recent genome contraction event (Hu et al. 2011). Hu et al. identified that the majority of this genome contraction is the result of deletion of transposable element sequences, and shortening of intergenic and intronic sequences. Despite the extreme reduction in size, *A. thaliana* has only 17% fewer genes than *A. lyrata* (Hu et al. 2011). This evidence points to selection for a compact genome.

Many genomes, including most mammalian genomes, exhibit periodic GC content changes on the level of tens to hundreds of kilobases. This is referred to as the isochore structure (Eyre-Walker and Hurst, 2001). GC-rich isochores exhibit higher gene density (Mouchiroud et al. 1991), and greater G4 forming potential (Maizels 2012). In contrast, the genomes of angiosperms, including Arabidopsis, do no have a clear arrangement of GC rich regions into isochores, and the GC content of the third codon position of gene CDSs is weakly correlated with the GC content of the flanking intergenic sequence (Tatarinova, T.V. et al 2010). GC content tends to be greater in exons than in intergenc or intronic sequence (Zhu et al. 2009). Furthermore, several authors have noted a negative gradient of GC richness across exons and introns, with GC content greatest at the TSS (Wong et al. 2002, Glémin et al. 2014). This suggests there may be greater G4 forming potential at the TSS proximal end of plant genes.

Previous analyses of PG4 densities in plant genomes have been conducted by Mullen et al. 2012, and Garg et al. 2016. Mullen et al. identified only 1200 three tetrad PG4s in the Arabidopsis genome. Using a Markov chain modelled genome with a window size of 100bp, Mullen et al. demonstrated that this represents a greater than two fold depletion of three tetrad PG4 sequences. This is a much greater depletion than in the human genome, which

Huppert & Balasubramanian suggested has 1.4 times fewer three tetrad PG4s than should be expected by chance (Huppert & Balasubramanian 2005). 70% of three tetrad PG4s were found in intergenic regions, and of these, 20% corresponded to the Arabidopsis telomeric sequence. Despite genic regions having a higher GC content (38.9%) than intergenic regions (31.1%), Mullen et al. found that the PG4 density of intergenic regions was still higher (genic 4.6 PG4s/Mb, intergenic 16.7 PG4s/Mb). These are not average windowed densities, however, and so intergenic densities may be skewed by the extremely PG4 dense telomeric and centromeric regions. Mullen et al. also predicted 43000 two tetrad PG4s in the Arabidopsis genome, using a loop length of up to 4bp. This did not constitute an enrichment or depletion over the expected levels from the Markov chain modelled genome. They noted that 80% of two tetrad PG4s occur inside genic regions, however, and suggested that this might lead to their formation in mRNA (Mullen et al. 2012).

In this chapter, we briefly examine the PG4 density of Arabidopsis compared to other plant genomes, and use metagene profiles to examine the distribution of PG4s across gene models. Finally, we develop a new simulation method to test whether PG4 motifs in CDS regions are hardcoded by protein coding sequences.

Materials and Methods

Plant genome PG4 analyses

The genomes of 48 multicellular land plants were downloaded over FTP from Ensembl Plants Release 39. This included 22 Monocotyledons, 23 Dicotyledons, and 3 Non-flowering plants. The genomes of four metazoans, *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), *Mus musculus* (mouse) and *Homo sapiens* (human), were downloaded from Ensembl release 92. Bed files of non-overlapping PG4 predictions were generated using `g4predict`, an in house Quadparser matching program. PG4 densities were measured for either two tetrad PG4s, or three or more tetrad PG4s, and a maximum loop length of 7bp was used. Average PG4/Mb densities were generated by using `bedtools makewindows` to create non-overlapping windows of 1Mb in size, and then using `bedtools intersect` in count mode to count the number of PG4s per Mb. The mean density for each genome was then calculated using `awk`. Genome size for each species was calculated from the total size of the fasta file. Genomes were annotated as Monocots or Dicots using metadata downloaded from the UniProt taxonomy database. Scatter plots were generated using `matplotlib` and `seaborn`.

Metagene Profiles

Shuffled genomes were generated in Python using `ushuffle` to shuffle sequences in 20bp windows, maintaining their nucleotide and dinucleotide contents. G content on both strands was calculated in 20bp windows, using `bedtools makewindows`, `bedtools nuc` and `awk`. GC Bed files of non-overlapping PG4s were converted into BigWig format using `bedtools genomecov` and `ucsc-bedGraphToBigWig`. Gene annotations were taken from Araport11. Overlapping transcripts of the same gene were flattened into a single bed12 interval per gene, with the leftmost TSS and start codon and the rightmost TTS and stop codon (or vice versa for genes on the negative strand). GC/PG4 coverage arrays for the 500bp upstream region, 5'UTR, CDS, 3'UTR and 500bp downstream region were then extracted for each gene using `pyBigWig` and reinterpolated to sizes of 50, 20, 100, 20, and 50 respectively. These were then

averaged across all genes to produce metaprofiles. Plots were generated using `matplotlib` and `seaborn`.

Reverse Translation Method

Relative frequency of codon usage for all Arabidopsis CDS sequences was calculated in python. Reverse translation was conducted by translating CDSs into protein, then randomly selecting codons to represent the protein, weighting each codon by its usage. 100 reverse translated potential coding sequences (PCSs) were generated per CDS. G content on each strand was calculated using 20bp windows and reinterpolated to 100 bins, for both real CDSs and PCSs. PG4 content was calculated using G4Seeqer and the overlapping Quadparser method. Overlapping PG4s were flattened into a single interval. PG4s were binned into 100 equally sized bins per gene, based on the midpoint of the PG4. Resultant profiles were stored in HDF5 format using `h5py`. Averaged profiles across all iterations of reverse translation, and all genes were generated and plotted using `matplotlib`.

Hardcoded PG4 Analysis

For hardcoded PG4 analysis, all overlapping two tetrad PG4 registers in CDSs were predicted using network analysis with `networkx`. G-runs were extracted from these PG4s, and the position, frame, and resultant protein sequence coded for by each G-run was calculated. Hardcoded G-runs were identified by analysing whether it would be possible to use synonymous codons which do not change the protein sequence, which would abolish the G-run. PG4s which had G-runs which all code for the same protein motif were labelled as repetitive. For G-run frequency plots, G-runs which contribute to multiple PG4s were deduplicated to give only one G-run per position. G-runs which contributed to both repetitive and non-repetitive PG4 registers were labelled as non-repetitive. For hardcoded PG4 metagene profiles, PG4s were binned into 100 equally sized bins per CDS, based on the midpoint of the PG4. All overlapping PG4s were counted in the profile. The total number of PG4s per bin was counted, and cumulative frequency metagene profiles were plotted using `matplotlib`. Frequency plots

of hardcoded PG4s/G-runs, repetitiveness, and protein motifs were produced using `seaborn`.

Results

The genome of *Arabidopsis thaliana* poor in three tetrad PG4s, but not two tetrad PG4s.

To compare the PG4 density of the Arabidopsis genome to other organisms, we downloaded the set of 48 land plant genomes available in Ensembl Plants Release 39, which included 22 Monocotyledons, 23 Dicotyledons, and 3 Non-flowering plants. The genomes of the metazoans *Drosophila melanogaster* (fruit fly), *Danio rerio* (zebrafish), *Mus musculus* (mouse) and *Homo sapiens* (human) were also analysed. PG4s with three or more were identified using the Quadparser method and the average density per Megabase was calculated for each genome. Arabidopsis has the smallest genome of any of the sequenced plants, estimated at 135Mb (119Mb in the golden path sequence). It also has one of the lowest three tetrad PG4 densities. Only 1284 non-overlapping PG4s with three or more tetrads are predicted to form in the whole Arabidopsis genome, with an average density of 10.4 PG4s/Mb. In comparison, the human genome is extremely PG4 dense, with an average of 123 PG4s/Mb. Monocot plants also tend to have much greater PG4 densities than Dicots (median density 59 PG4s/Mb vs. 3.3 PG4s/Mb). This is likely to result from a greater GC content in Monocot genomes. Non-flowering plants such as the bryophyte *Physcomitrella patens* had PG4 densities which resembled those of the Dicots more closely. We did not find a correlation between PG4 density and genome size (Spearmans rho = -0.02).

We noted that the PG4 densities of the warm blooded mammals *M. musculus* and *H. sapiens* are much greater than those of *D. melanogaster* or *D. rerio*, or any of the plants. The PG4 density of *Mus musculus* is 227 PG4s/Mb, more than twice that of any plant analysed. We hypothesised that this greater density may be in part due to the homothermic nature of mammmals, which could mean their body temperatures are high enough that three tetrad PG4s are less stable. Since the melting temperatures of three tetrad G4s can reach up to 100 Degrees Celsius, it is feasible that at the physiological temperature ranges that plants live in, three tetrad G4s may be more difficult to resolve, leading to problems during replication or

transcription. Two tetrad G4s, which are known to form *in vitro*, but have been historically considered too unstable to be prevalent *in vivo*, might in fact be more useful as molecular switches in plants, since they melt at lower temperatures.

To determine whether plants have greater numbers of two tetrad PG4s, we again performed prediction using using the Quadparser method. Since three tetrad PG4s contain subpatterns which conform to the two tetrad Quadparser pattern, we filtered out any two tetrad PG4s which overlapped with three tetrad PG4s. We found that plants tend to contain a lot more two tetrad PG4s, with several species of Monocot in fact having higher average densities than *M. musculus* or *H. sapiens*. Arabidopsis was also more dense in two tetrad PG4s, with 959 PG4s/Mb. This was significantly higher than the median Dicotyledon density (228 PG4s/Mb), perhaps indicating a stronger role for these two tetrad G4s in Arabidopsis.

Finally, we correlated the density of two and three or more tetrad PG4s in different organisms. Unsurprisingly, we found a very strong correlation (Spearmans rho $= 0.95$), however *M. musculus* and *H. sapiens* were found to be clear outliers from the plants, and showed a much greater three tetrad density than might have been expected from the regression line. We suggest that this indicates that two tetrad PG4s may play a regulatory role in plants which could be similar to have performed by three tetrad PG4s in warm-blooded mammals.

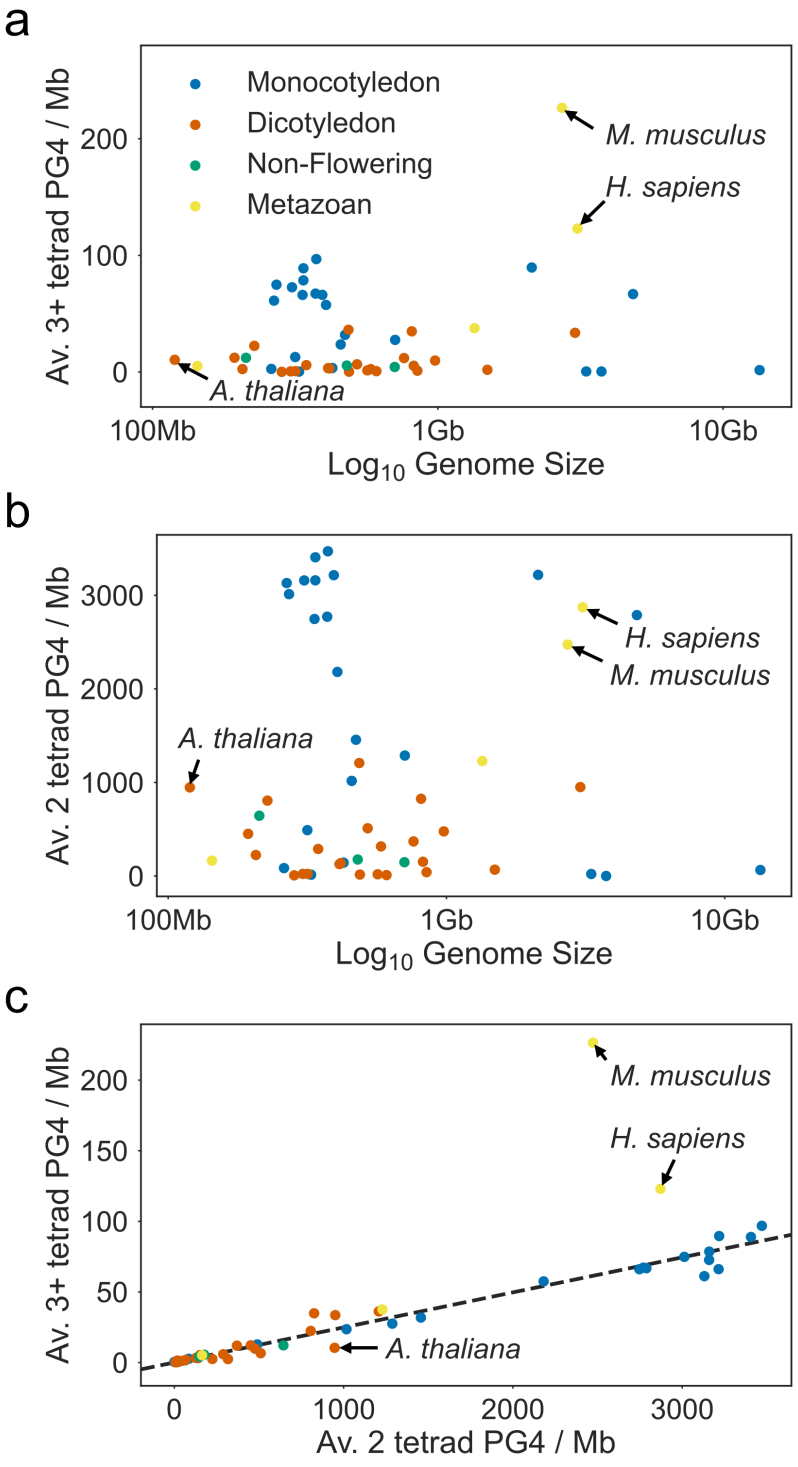

Figure 3.1: PG4 density of Plant Genomes a) and **b)** Scatter plots showing log10 genome size vs. the average **a)** three tetrad or more and **b)** two tetrad PG4 density per Megabase for 22 Monocotyledons, 21 Dicotyledons, 3 Non-flowering plants and 4 metazoans. Plant genomes are much poorer in three tetrad PG4s than the metazoans *H. sapiens* and *M. musculus*, but have more comparable densities of two tetrad PG4s. **c)** Scatter plot showing the relationship between three tetrad and two tetrad PG4 densities. *M. musculus* and *H. sapiens* do not follow the same pattern of relative PG4 densities as plants.

PG4s are non-uniformly distributed in the *Arabidopsis thaliana* genome

It is well documented that PG4s are not randomly distributed in the human genome, with an enrichment of PG4s in promoter sequences. We therefore used metagene profiles to identify the localisation of PG4s in the Arabidopsis genome. The distribution of G content (in 20bp windows), two tetrad PG4s, and three or more tetrad PG4s was calculated and averaged across all all protein coding genes. The density in UTRs and CDS regions (including introns) was reinterpolated into 20 and 100 bins, respectively, and an upstream and downstream size of 500bp was used. PG4s identified on the coding and template strands were analysed separately. The GC content of genic regions is greater than intergenic sequence, with the greatest coding strand G content towards the TSS distal end of the CDS, and the greatest template strand C content in the 5' UTR and at the TSS proximal end of the CDS (Fig **??**a). This translates into a greater density of two tetrad PG4s localised at the beginning of genes on the template strand and towards the end of genes on the coding strand (Fig **??**c). The number of three or more tetrad PG4s is extremely low in the Arabidopsis genome, however there is a slight increase around the TSS and in the 5' UTR on the template strand, and at the TSS distal end in the coding strand (Fig **??**b).

In order to identify whether these PG4s are simply a product of higher local GC content in different genic regions, we performed simulation of genomic sequence by shuffling in 20bp windows, maintaining the nucleotide and dinucleotide contents. This does not change the GC content distribution of the metagene profile. It does, however, indicate a strong enrichment of two tetrad PG4s over expected levels throughout the CDS, particularly on the template strand at the TSS proximal end (Fig **??**c). Furthermore, we see an enrichment of three or more tetrad PG4s over expected levels around the TSS on the template strand (Fig **??**b), and a depletion inside coding regions on both strands.

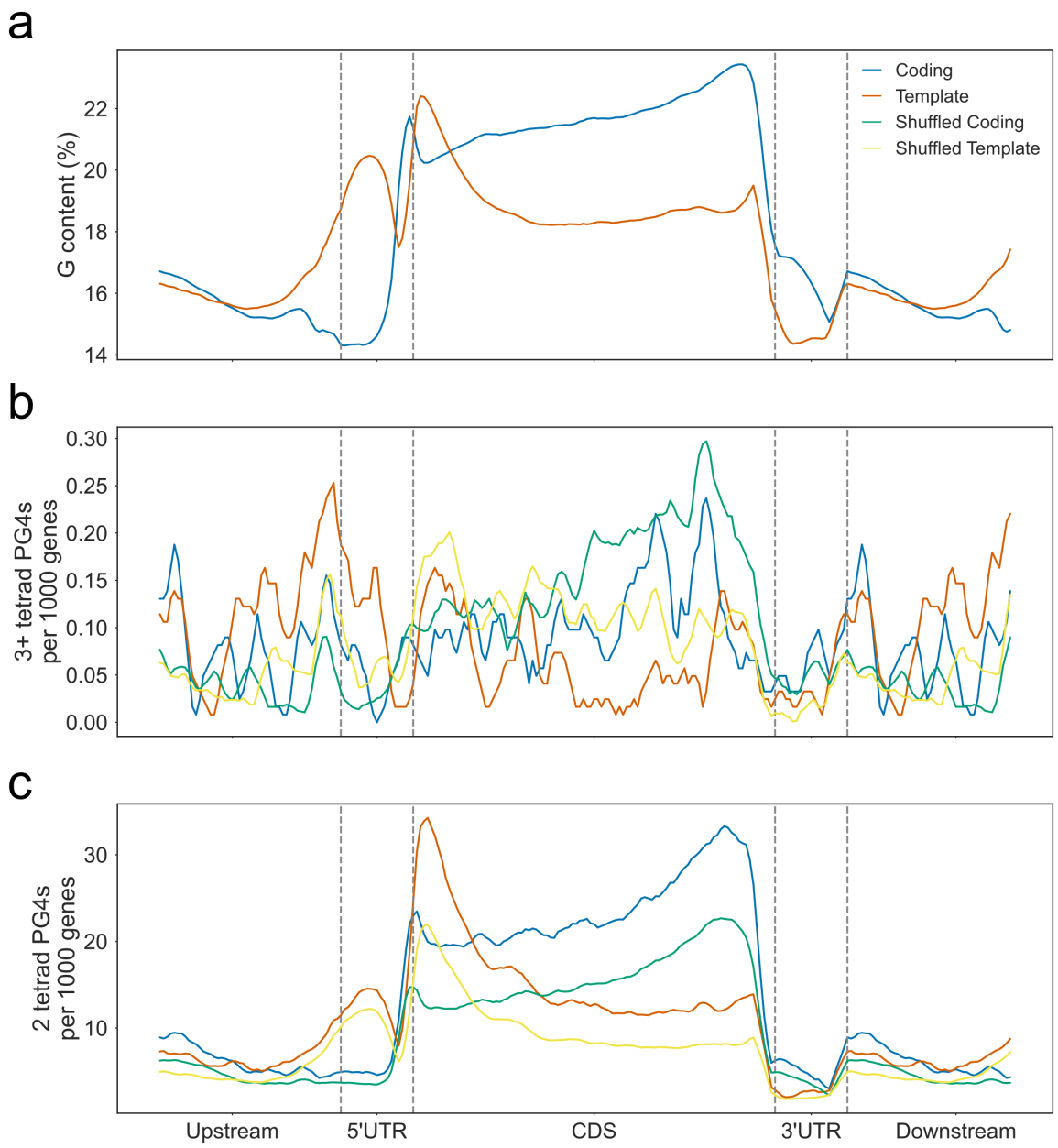

Figure 3.2: Metagene Profile of GC content and PG4 density Metagene profiles showing the **a)** G content, **b)** three or more tetrad PG4 content and **c)** two tetrad PG4 content of protein coding genes on the coding (blue) and template (orange) strands. Green and yellow lines show the average coding and template PG4 contents for genes where the sequence has been shuffled in 20bp windows, maintaining mononucleotide and dinucleotide frequencies. G content metaprofiles are identical to the real metaprofiles in these shuffled controls.

Reverse translation permutation testing reveals codon usage bias towards G4 formation in template strand 5' UTRs.

The enrichment of two tetrad PG4s inside coding regions in Arabidopsis could be indicative of a function for G4s in transcriptional or translational regulation, however, it is also possible that these sequences could be a byproduct of specific protein motifs which are encoded by GC rich codons. In order to explore this, we developed a novel sequence permutation method which we call "reverse translation". First, we calculated the codon usage table (the quantifiable bias in usage of synonymous codons) from all CDS sequences. The sequences were then translated into their protein product sequence, then reverse translated back into potential coding sequences (PCSs) with randomly selected codons, but using the codon usage table as weights. These PCSs are therefore sequences which might be expected to code for the given protein, assuming that the codon bias is identical across all genes and all positions in genes. We performed 100 reverse translation shufflings for each CDS and then calculated the GC and PG4 content of each PCS using the Quadparser method and G4Seeqer.

The GC content of CDSs is greatest towards the start and end of the interval, and dips in the middle (Fig 3.3a). This is due to a greater G content at the start codon proximal end on the template strand, and a greater G content at the start codon distal end on the coding strand. When we performed reverse translation using a single codon usage table, some of this bias was abolished. This indicates that most of the GC content of the CDS is not hardcoded into the sequence by the protein content. Codon usage is therefore presumably different at the start and end of the gene.

As shown in Fig **??**c, there is a higher density of PG4s at the start of the CDS on the template strand, and a higher density towards the end of the CDS on the coding strand. This is also seen in Fig 3.3b & c. PCS sequences demonstrate the same biases in PG4 distribution using both Quadparser and G4Seeqer predictions. This demonstrates that unlike GC content, the PG4 content of some genes is hardcoded by protein sequence (Fig 3.3b & c). This may be due to the repetitive nature of some protein motifs. PCS PG4 content is higher than the real PG4 levels across the coding strand however, suggesting that codons which reduce PG4

forming potential on this strand may be selected for. On the template strand, we see strong enrichment of PG4s in real sequences over expected levels from PCSs in the first 50% of the CDS (Fig 3.3b & c). This suggests that C rich codons may be selected for at the start of genes to increase the G4 forming potential of the template strand.

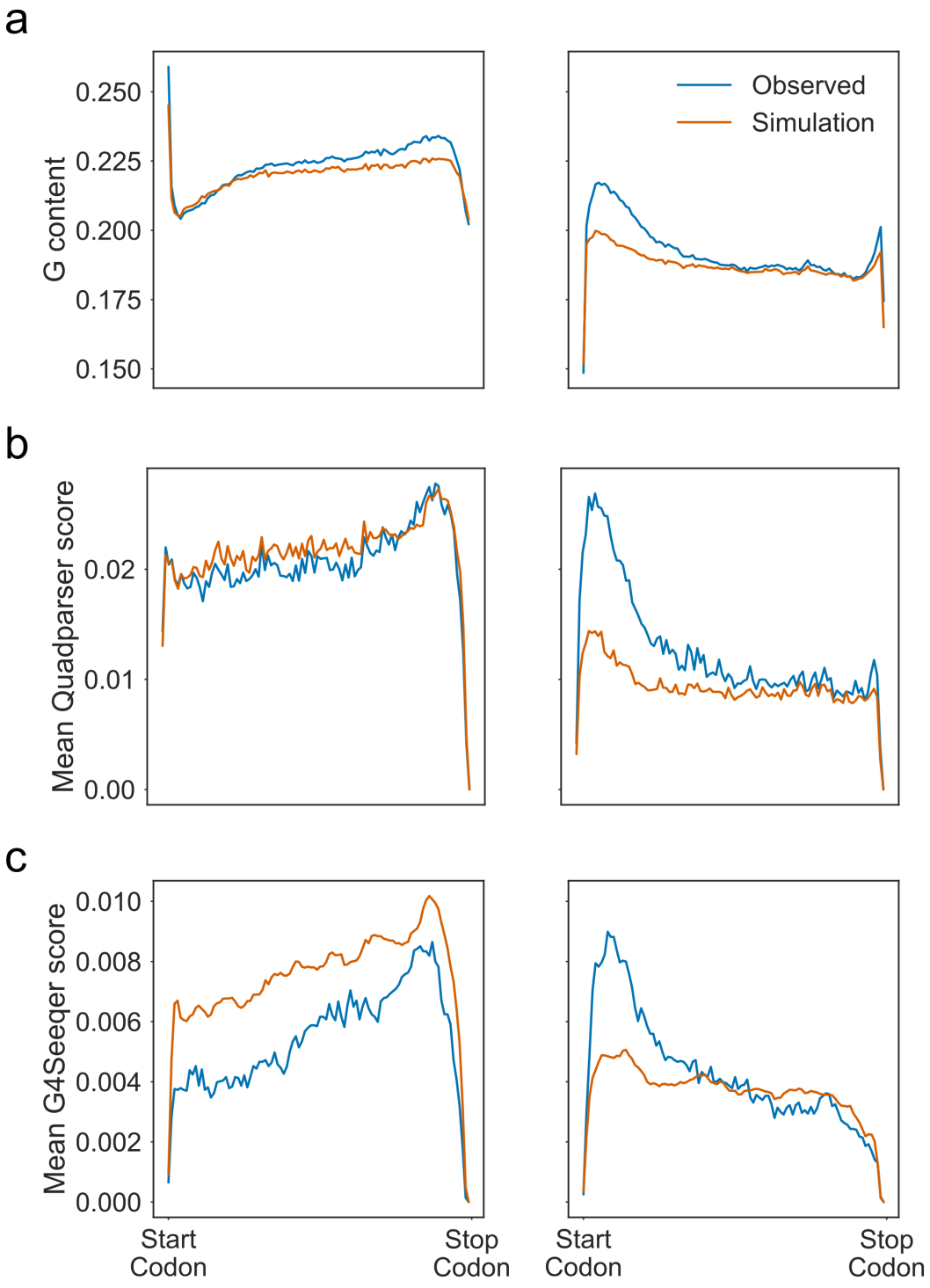

Figure 3.3: Reverse Translation Simulation shows that PG4s are enriched at the Start Codon proximal region of the template strand. Metagene profiles showing **a)** G content **b)** Quadparser PG4s and **c)** G4Seeqer PG4s for real CDS regions (blue) vs. reverse translation simulated potential coding sequences (PCS) (orange).

Protein motifs hardcode G4 forming potential into coding regions

Our reverse translation method indicates that the PG4 forming potential of may arise from protein sequence, i.e. if protein sequence is evolutionarily constrained, then PG4s are hardcoded into the CDS. We performed an analysis of which protein motifs most often lead to two tetrad PG4s. All overlapping PG4s and the G-runs that form them were predicted, and the amino acids which are encoded by each G-run was identified. The G-run was then classified as either hardcoded or not, depending on whether or not the same amino acids could be encoded differently without introducing a G-run in the same position. We also identified repetitive and non-repetitive G-runs. G-runs were considered repetitive if the protein motifs that they encode were the same for all G-runs in the G4.

Our analysis shows that 58% of PG4 G-runs on the coding strand, and 48% on the template strand, are hardcoded. Of these hardcoded PG4s, around 51% and 60% are found in repetitive PG4s on the coding and template strand, respectively. On the other hand, most non-hardcoded PG4s on both strands are also non-repetitive (Fig 3.4a). This suggests the presence of a number of entirely hardcoded PG4s which are encoded by repetitive protein motifs.

We counted the total number of hardcoded G-runs in each overlapping PG4 register on both strands (Fig <ref{hardcoded}b). We found that on both strands, the greatest number of PG4s were completely hardcoded, again suggesting a large number of PG4s encoded by repetitive protein motifs. On the template strand, however, we also found that 19% of PG4s had no G-runs hardcoded, suggesting the presence of template PG4s which are selected for specifically.

Analysis of the frame of the first G in G-runs vs. their hardcoded status identifies that 46% and 48% of coding and template G-runs in PG4s are frame 0, i.e. are made up of the first two bases of a codon. These are all hardcoded. This is intuitive since the third position of codons is the "wobble" position which is most often degenerate amongst synonymous codons. Approximately one third and one fifth of coding strand G-runs in frames 1 and 2 are hardcoded, whilst all template strand G-runs in frames 1 and 2 are not hardcoded. Interestingly, 36% of G-runs on the coding strand are frame 2 whilst only 24% on the template strand are. This is most likely due to the relative frequencies of different amino acids whose codons may form

G-runs.

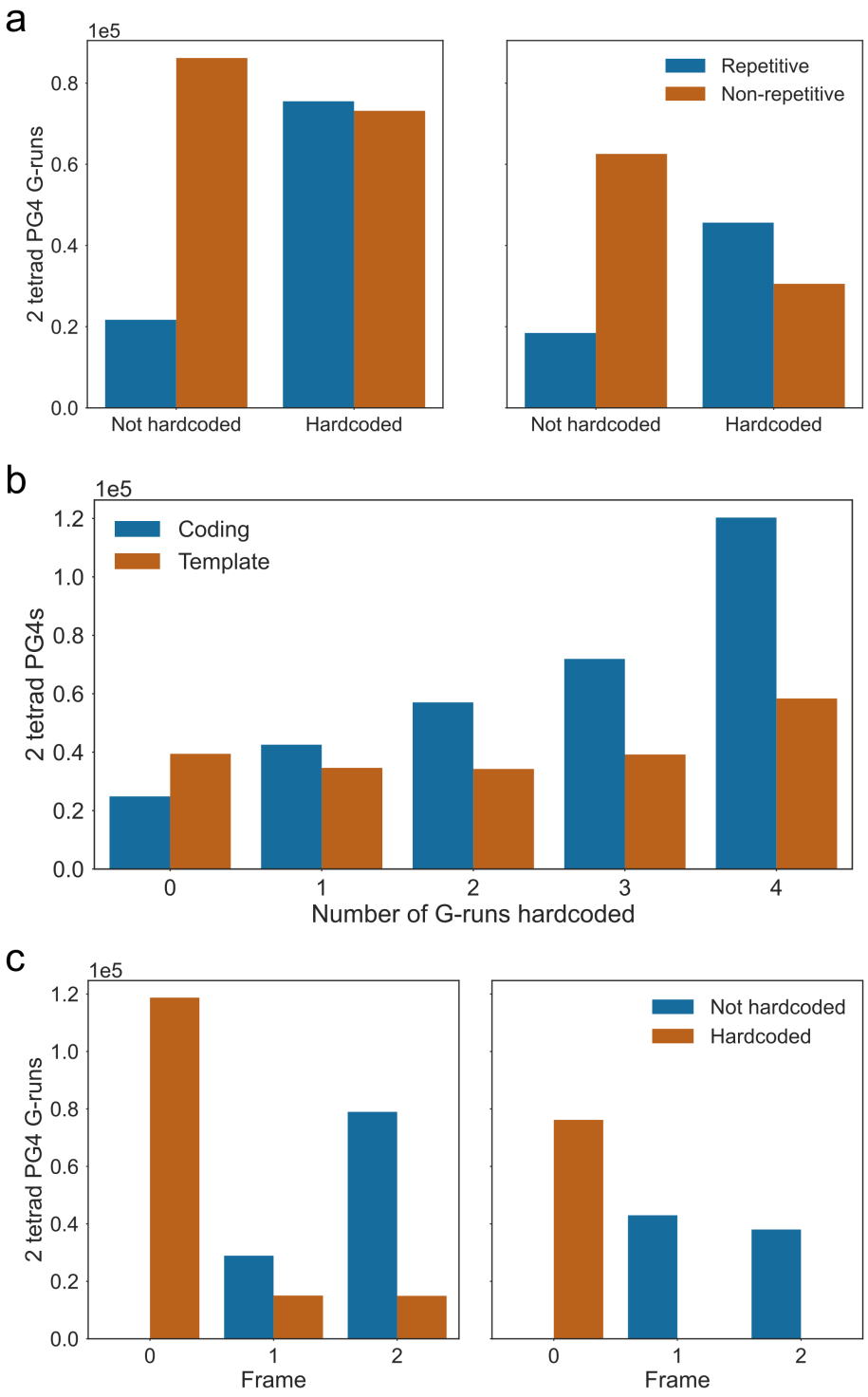

Figure 3.4: 54% of CDS PG4 G-runs are hardcoded by protein sequence. a) Frequency plot showing the total number of G-runs contributing to PG4s which are hardcoded and repetitive. Left and right panels show frequencies on coding and template strands, respectively. Hardcoded G-runs are defined as those GG dinucleotides that cannot be removed from the sequence without changing the amino acid sequence which is coded for. Repetitive G-runs are defined as those which contribute to PG4s where all G-runs are part of codons which encode the same sequence. **b)** Frequency plot showing the total number of hardcoded G-runs for each overlapping PG4 register on the coding (blue) and template (orange) strands, respectively. **c)** Frequency plots showing start frame of CDS G-runs vs. hardcoded status. Left and right panels show frequencies on coding and template strands, respectively.

To identify the location of these non-hardcoded template PG4s in the CDS, we plotted metagene profiles (Fig 3.5). We found that on the coding strand, non-hardcoded PG4 levels were approximately the same throughout CDSs (mean 7.7%, standard deviation 1.84%) (Fig 3.5a). On the template strand however, the average non-hardcoded PG4 levels on the template strand is 17%, with a standard deviation of 7%. We found that 27% of PG4s in first 10% of the metagene profile downstream of the start codon were completely non-hardcoded (Fig 3.5b). This shows that PG4s are selected by codon usage in the start codon proximal region of the template strand.

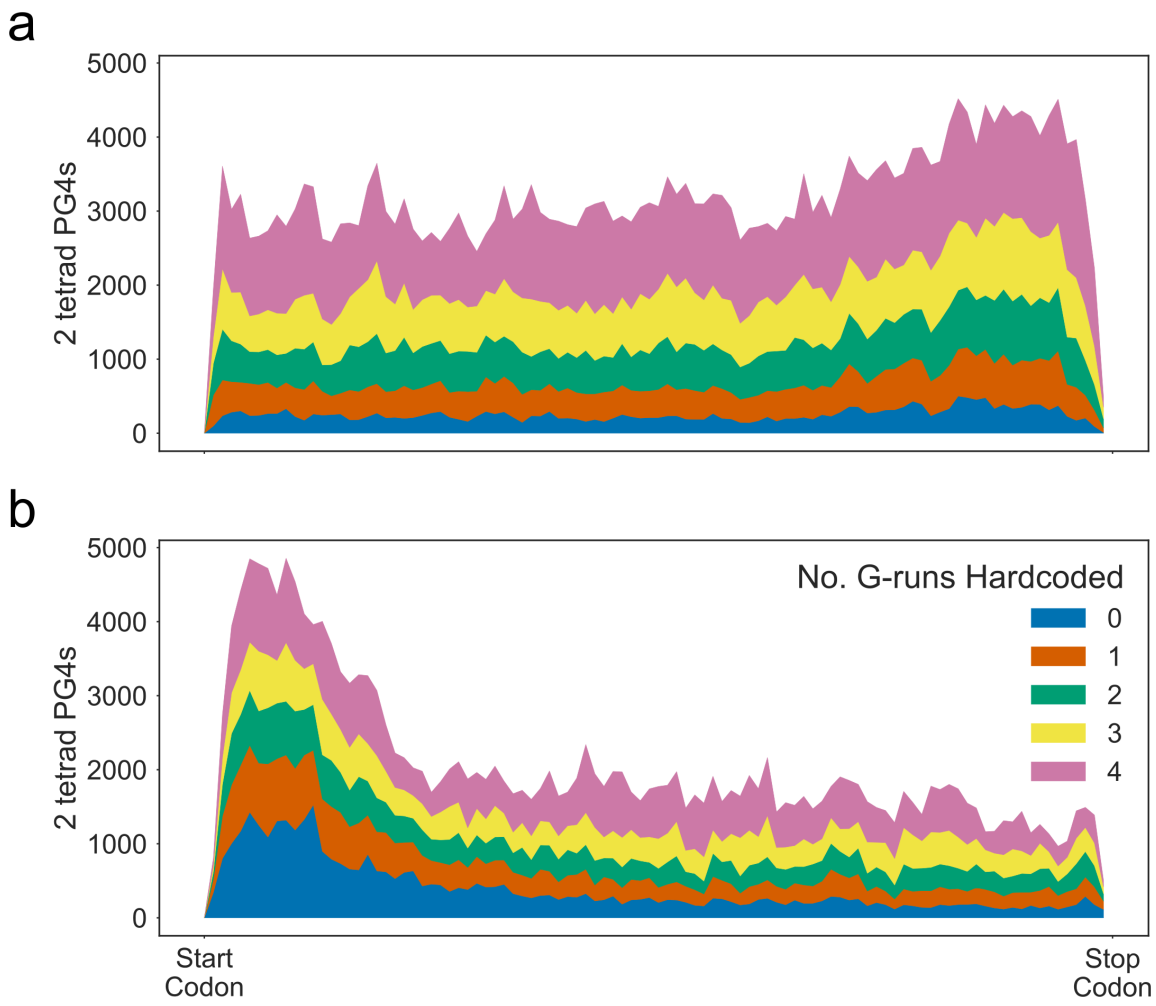

Figure 3.5: Non-hardcoded PG4s levels are greater at the start codon proximal region of CDSs, on the template strand. a) Cumulative metagene profiles showing the distribution of PG4s with different numbers of hardcoded G-runs on the **a)** coding and **b)** template strands, respectively.

Harkness & Mittermaier showed that sequences which are able to form a variety of G4s through combinatorial interactions of greater than four G-runs are more entropically favourable, and therefore higher melting temperatures (Harkness & Mittermaier 2016). Different PG4 registers can be formed in sequences where there are extra Guanines in some G-runs, allowing "sliding" (Harkness & Mittermaier 2016), or where there are five or more G-runs in close proximity, where G4s could be formed from any four G-runs. We were interested to see whether template strand PG4s at the start codon proximal end of CDSs tended to have more registers, which might make them more stable. We therefore identified all overlapping PG4 register clusters using network analysis (Figure 3.6a), and produced metagene profiles for clusters with only 1, 2-5, 6-20, or greater than 20 registers. We found that 64% of PG4 clusters on both the coding and template strands of CDSs had greater than one register. The largest number of registers for a single cluster was 3096. This cluster was made up of 329 two base G-runs and was able to form a maximum of 75 non-overlapping PG4s. To identify whether the number of G-registers formed by PG4 clusters varied by CDS position, we binned PG4s into metagene profiles (Figure 3.6b). We did not find much variation in the percentage of PG4s with G-register number greater than one, on either strand, however (mean 64%, standard deviation 3.5% for both strands).

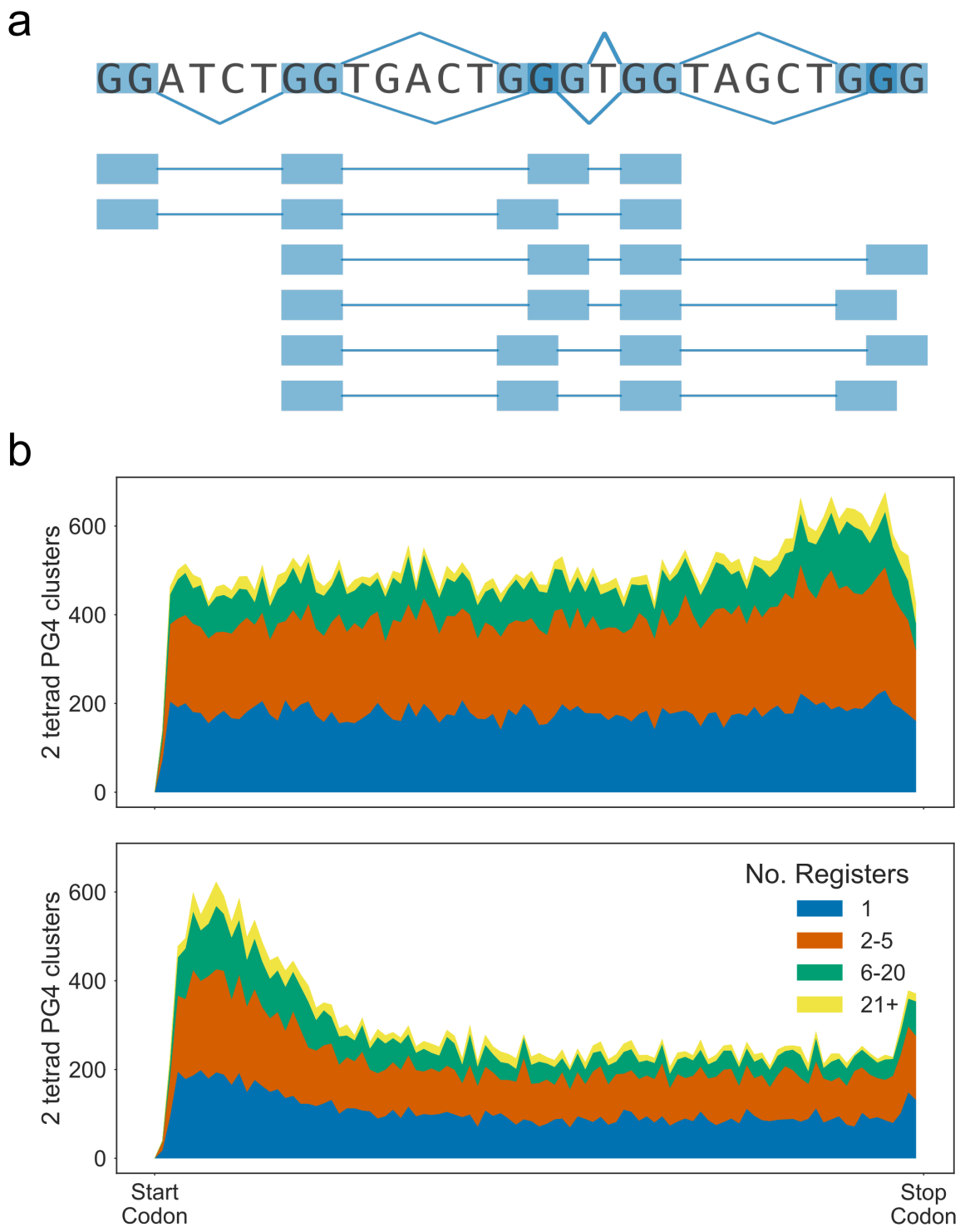

Figure 3.6: PG4 Register does not vary over CDSs a) Diagram showing how a sequence with multiple G-runs can form multiple overlapping PG4 registers and topologies. **b)** Cumulative metagene profiles showing distribution of PG4 clusters with different numbers of PG4 registers on the coding strand (top panel) and template strand (bottom panel).

Finally, we examined which amino acid motifs were most common in PG4 G-runs. We found that on the coding strand, by far the most common G-run motif was glycine (Fig 3.7). The majority of these G-runs are hardcoded, because the codon for glycine is GGN (Fig 3.7a, left panel). Furthermore, more than 50% of glycine G-runs are found in repetitive PG4s, suggesting that poly-glycine is a common PG4 forming motif (Fig 3.7b). There was not a clear majority motif for non-hardcoded PG4 G-runs.

On the template strand, we found that the most common PG4 G-run motif was proline. This is again mostly hardcoded, since the codon for proline is CCN (Fig 3.7a, right panel). More than 50% of proline G-runs were also found in repetitive PG4s, suggesting that like glycine on the coding strand, proline homopolymers are the most common PG4 forming motif on the template strand.

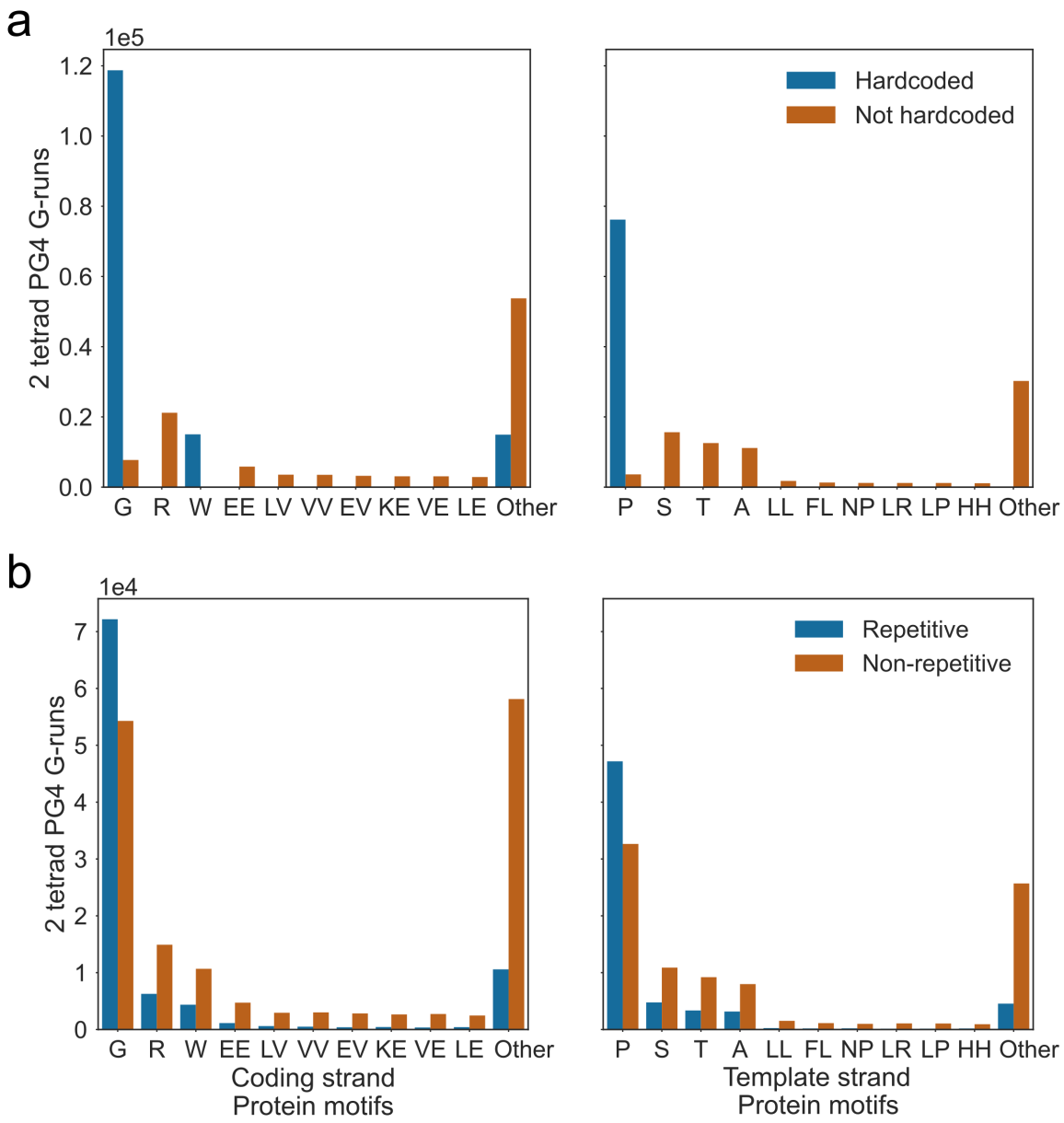

Figure 3.7: Protein motifs that are coded by PG4 G-runs. Frequency plots showing the 10 most common amino acids motifs which PG4 G-runs contribute to the coding of. Left and right panels are for coding and template strands, respectively. Bars are coloured by the frequency of **a)** hardcoded and **b)** repetitive G-runs, respectively.

Discussion

The majority of analysis on the effects of G4s on biological processes has thus far been conducted in mammalian systems, particularly in human cells. The genomes of the mammals *M. musculus* and *H. sapiens* are extremely rich in three tetrad PG4s compared to those of most plants. Many plant genomes, particularly those of Monocots, have comparable levels of two tetrad PG4s to the mammalian genomes, however. Furthermore, the ratio of two tetrad to three tetrad PG4s in plant genomes is much greater than in humans or mice, suggesting that in plant systems, two tetrad PG4s may play a greater role in regulation that in mammals. This is intuitive because plants tend to exist at more ambient temperatures than warm-blooded mammals, and therefore the lower melting temperatures of two tetrad PG4s might make them more favourable choices for molecular switches than three tetrad PG4s.

The dicotyledon *Arabidopsis thaliana* has a low three tetrad PG4 density but a relatively high two tetrad PG4 density. Previous analyses by Mullen et al. have identified that the majority of Arabidopsis two tetrad PG4s are located inside genic regions, whilst the majority of three tetrad PG4s are located in intergenic regions (Mullen et al. 2010). We performed metagene profile analyses and showed that the levels of two tetrad PG4s are greatest inside CDS regions on both the coding and template strand. These levels were greatest at the start codon proximal end on the template strand, and the distal end on the coding strand. We also identified a peak of PG4s on the template strand in the 5' UTR. These levels are higher than would be expected from random sequence with the same mononucleotide and dinucleotide frequencies, suggesting that the sequence is ordered specifically to allow G4 formation. Since the template strand is scanned by RNA polymerase II (Pol II) using transcription, it is possible that G4s which form in this strand may cause blockages that slow or stall the progress of Pol II using transcription. Since the levels of PG4s are greatest at the TSS proximal end on this strand, these might perhaps be involved in proximal pausing or slowing elongation initially to ensure modifications and co-factors of the transcriptional complex are correct. Proximal pausing is a common checkpoint of transcription in mammalian systems, but was not identified by Hetzel et al. in Arabidopsis or maize GRO-seq data (Hetzel et al. 2016). Another possibility

is that TSS proximal template strand G4s might act as molecular switches which could cause premature termination when folded, thereby regulating the expression of genes.

Since PG4 forming sequences in CDS regions must also code for protein sequence, we were interested in identifying the degree to which PG4s are determined by coding sequence. Some PG4 motifs cannot be removed from the CDS without changing the protein sequence. We refer to these as hardcoded PG4s. To explore this idea, we developed the reverse translation simulation, where codon usage across the whole genome was used to simulate potential coding sequences (PCSs) for each CDS, and the number of PG4s in the real CDS vs the average PCS was compared. This method identified that the G-content skew on both strands is heightened by codon choice, however some G-content skew is also hardcoded by protein sequence. This suggests that protein sequence may in fact be under selection to increase template strand G-content at the start of the CDS. We also found that the levels of PG4s on the coding strand of real CDSs were lower than was expected from PCSs, i.e. codon usage selectively removes non-hardcoded PG4s on the coding strand. This is possibly to remove obstacles to the ribosome during translation, since coding strand PG4s may also form in the mRNA, and RNA G4s are more stable than DNA G4s. The levels of coding strand PG4s were greater at the distal end of the CDS in both real CDSs and simulated PCSs, suggesting a greater level of hardcoded PG4s occur towards the end of CDSs.

Reverse translation identifies that the levels of template strand PG4s are greater in real CDSs than expected from PCSs at the start codon proximal end. This enrichment falls through the CDS, and the second half of the template strand CDS is depleted in PG4s. This suggests that PG4s serve some purpose at the start of CDSs, perhaps in regulating Pol II speed. Furthermore, we see a peak of template strand PG4s at the proximal end of PCSs, suggesting that there are more hardcoded or partially hardcoded PG4s at the proximal end, and that N-terminal protein sequence may in fact be selected to allow PG4 formation in the DNA.

To further explore the levels of hardcoded vs. non-hardcoded PG4s in Arabidopsis CDSs, we identified, for each overlapping PG4 register, whether each G-run was hardcoded or not. The one or two amino acid motif which was contributed to by each G-run was also determined. We found that greater than 50% of all PG4 G-runs are hardcoded, and 34% of all PG4s are

totally hardcoded. The start codon proximal end of the template strand contains the greatest number of non-hardcoded PG4s, explaining the strong enrichment in this region compared to PCSs.

The most common amino acids which contribute to hardcoded PG4s are glycine (codon GGN) on the coding strand, and proline (codon CCN) on the template strand. G-runs encoding these amino acids also tend to be repetitive, i.e. contribute to PG4s in which all G-runs encode the same amino acid motif. Polyproline and polyglycine rich motifs are common in the Arabidopsis genome. Polyglycine rich proteins (GRPs) are involved in a number of processes, including cell elongation, plant defense, and osmotic or salt stress (Mangeon et al. 2010). A number of RNA-binding GRPs which have RNA chaperone activity are regulated by osmotic stresses and by abscisic acid (Mangeon et al. 2010). Interestingly, the cellular concentration of G4 stabilising potassium cations is increased during these stresses, suggesting that G4s may be more favourable. Mullen et al. have previously suggested that intracellular potassium concentrations might regulate two tetrad G4 formation in Arabidopsis mRNAs, causing conformational changes in the RNA (Mullen et al. 2012). Furthermore, Kim et al. used SELEX to identify that the stress responsive RNA chaperone GRP7 binds preferentially to G-rich single stranded DNA or RNA (Kim et al. 2007), though they did not test whether these formed G4s. It is possible that GRPs are involved in a feedback mechanism, stabilising mRNAs (including their own mRNAs) during stress by either binding to or resolving G4s in the mRNA.

Polyproline rich proteins are often structural proteins, and are a major constituent of the plant cell wall. Proline rich motifs form PG4s in the template strand of DNA. These will not form in the mRNA, but may cause issues for Pol II using transcription. This will be discussed further in the following chapters.

Chapter 4

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 4.1a, 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. 4.1b, Rhodes & Lipps 2015).

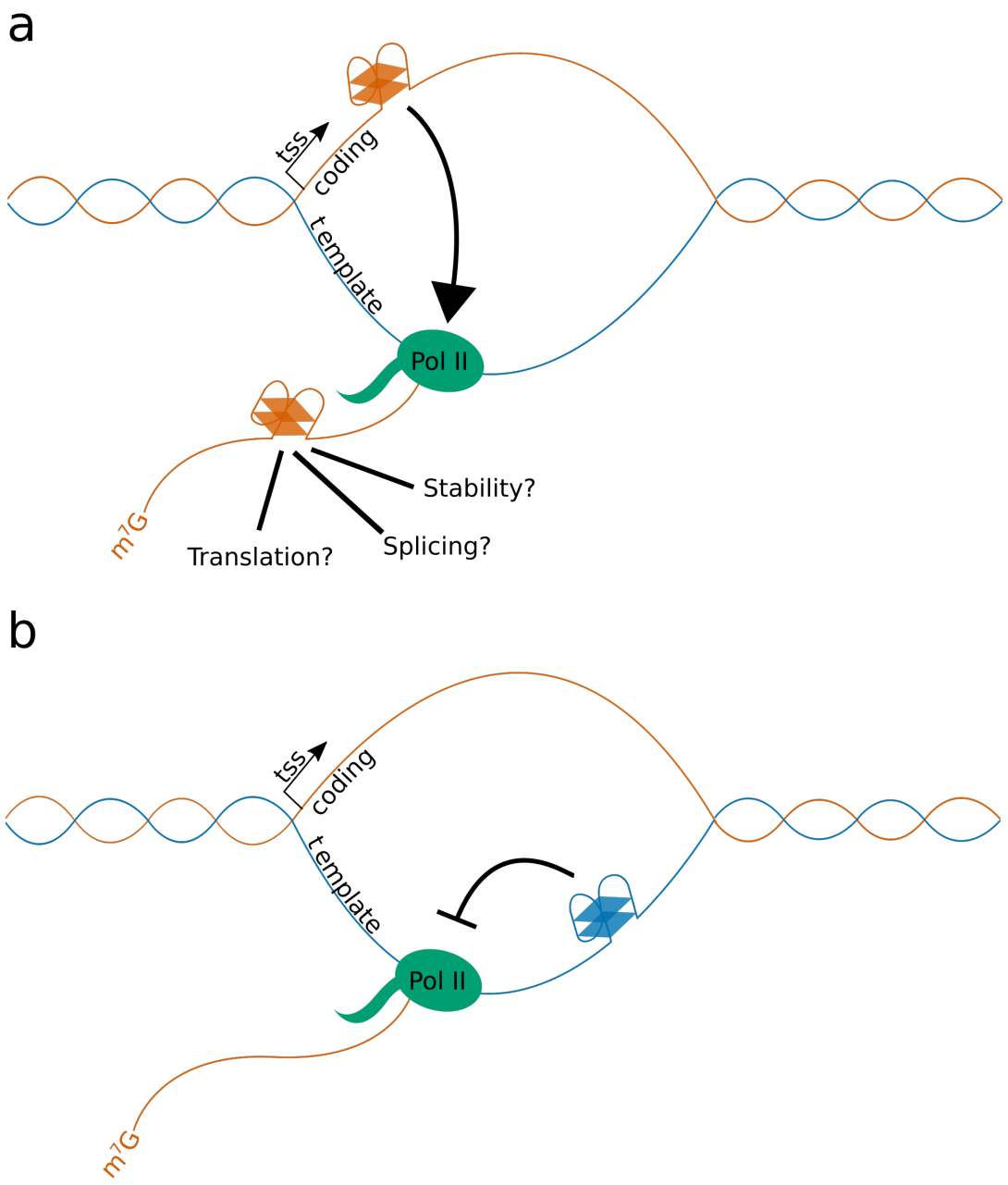

Figure 4.1: Possible Mechanisms for DNA and RNA G4 function in transcription and gene expression. Adapted from Figure 4. Rhodes & Lipps 2015. **a)** Possible mechanism for function of G4s located in the coding strand. Since G4s form from single stranded DNA, G4s in the coding strand may promote melting 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 downregulated 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 scanning 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, downloaded 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 Araport11 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 4.2a). To more clearly visualise this skew we binned genes by average expression quartile (Fig 4.2b). 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.

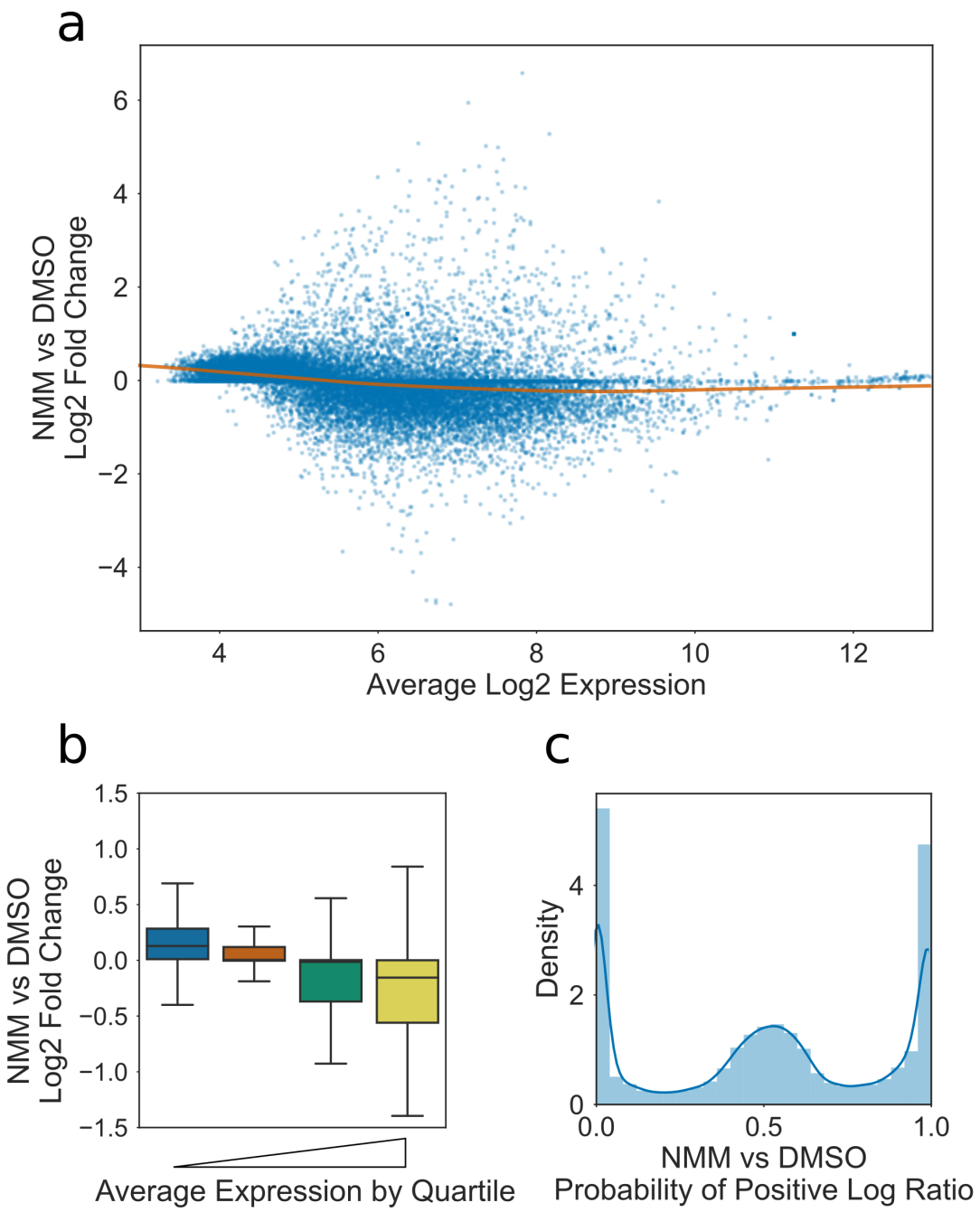

Figure 4.2: Global effect of NMM on Gene Expression. a) MA plot showing relationship between average gene expression and Log2 fold change in expression upon treatment with NMM. Orange line is lowess curve fit showing slight negative correlation between expression and fold change for genes in the expression 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 4.3a) (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 4.3b, 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.

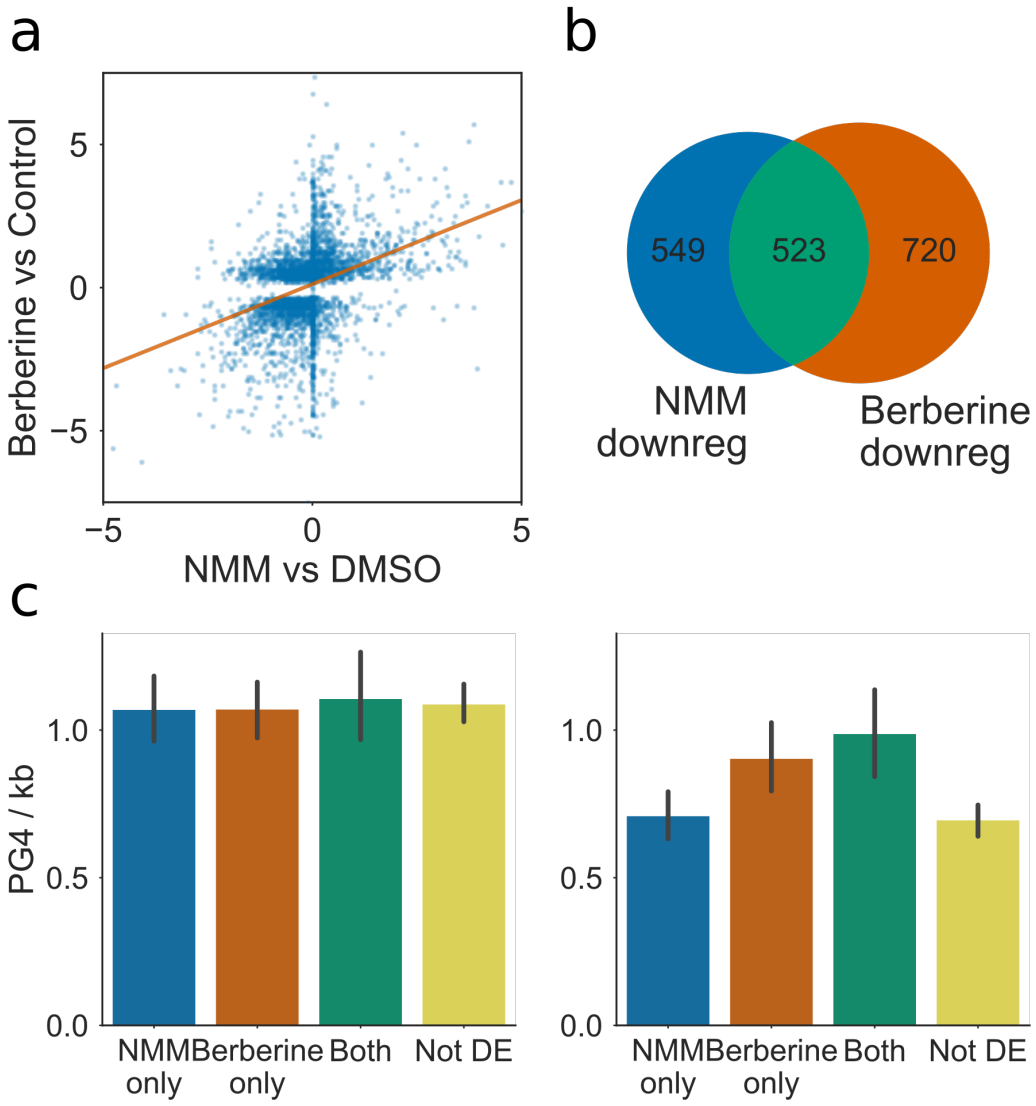

Figure 4.3: Comparison of gene expression during NMM treatment with expression during Berberine treatment. a) Scatter plot with regression line showing the correlation in expression change for NMM vs DMSO and Berberine vs Control. Processed Berberine data was taken from supplementary information of Nakagawa 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 10% more genes containing template PG4s than expected for a gene set of that size (p=2e-48) (Fig 4.4). This enrichment occured most specifically in the CDS and 5' UTR regions of genes (Fig 4.5). 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.

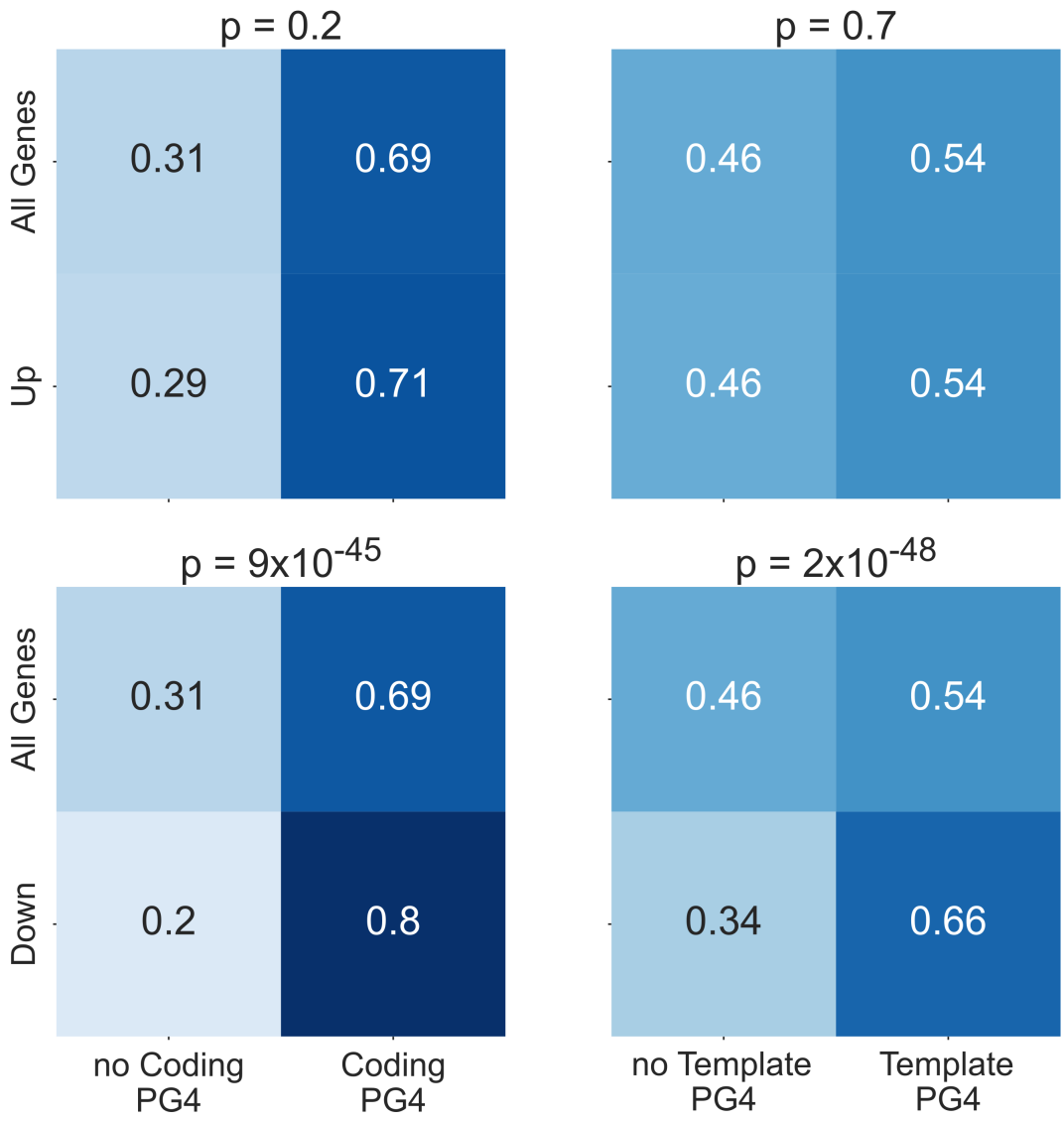

Figure 4.4: NMM downregulated genes are enriched in PG4s Heatmaps showing fractions of genes containing at least one predicted G4 in their gene body for upregulated genes vs all genes (top row) and downregulated genes vs all genes (bottom row) (For down and upregulated genes, FDR < 0.05 and absolute logFC > 0.5). PG4 predictions for the coding strand are in the left hand column whilst PG4 predictions for template strand are on the right. P values for each heatmap are calculated using Chi-squared tests. Genes downregulated by NMM show a particularly strong enrichment of PG4s, and particularly on the template strand.

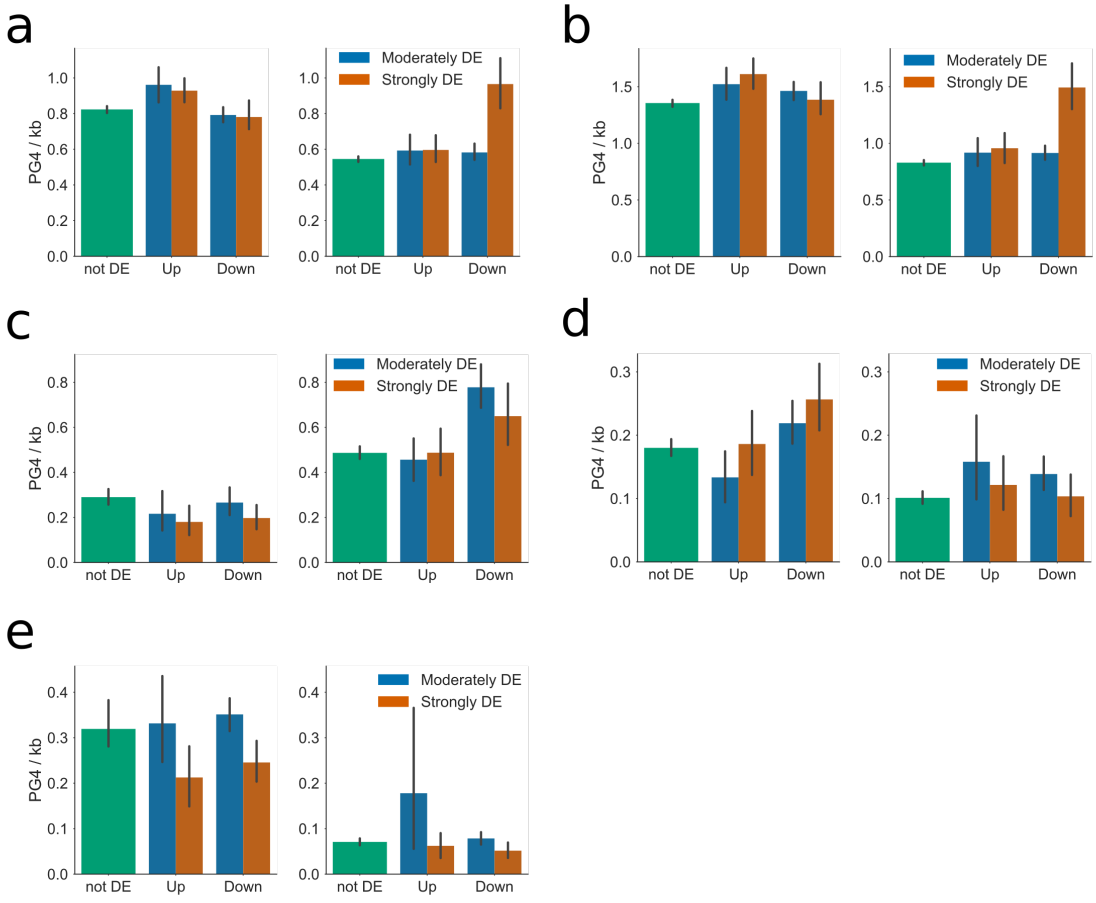

Figure 4.5: Distribution of PG4s in genes differentially regulated by NMM. Bar plots showing the average PG4 densities of genes up or downregulated by NMM, for **a)** full gene body (exons and introns), **b)** coding regions, **c)** 5' UTR, **d)** 3' UTR, and **e)** introns, respectively. In each figure, 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 4.3c). 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 4.6b). Each cluster contained between approximately 1000 and 8000 PG4s (Fig 4.6a). 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 down-regulated genes (Fig 4.6c). One cluster, however, was strongly enriched on the template strand (Fig 4.6d). 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).

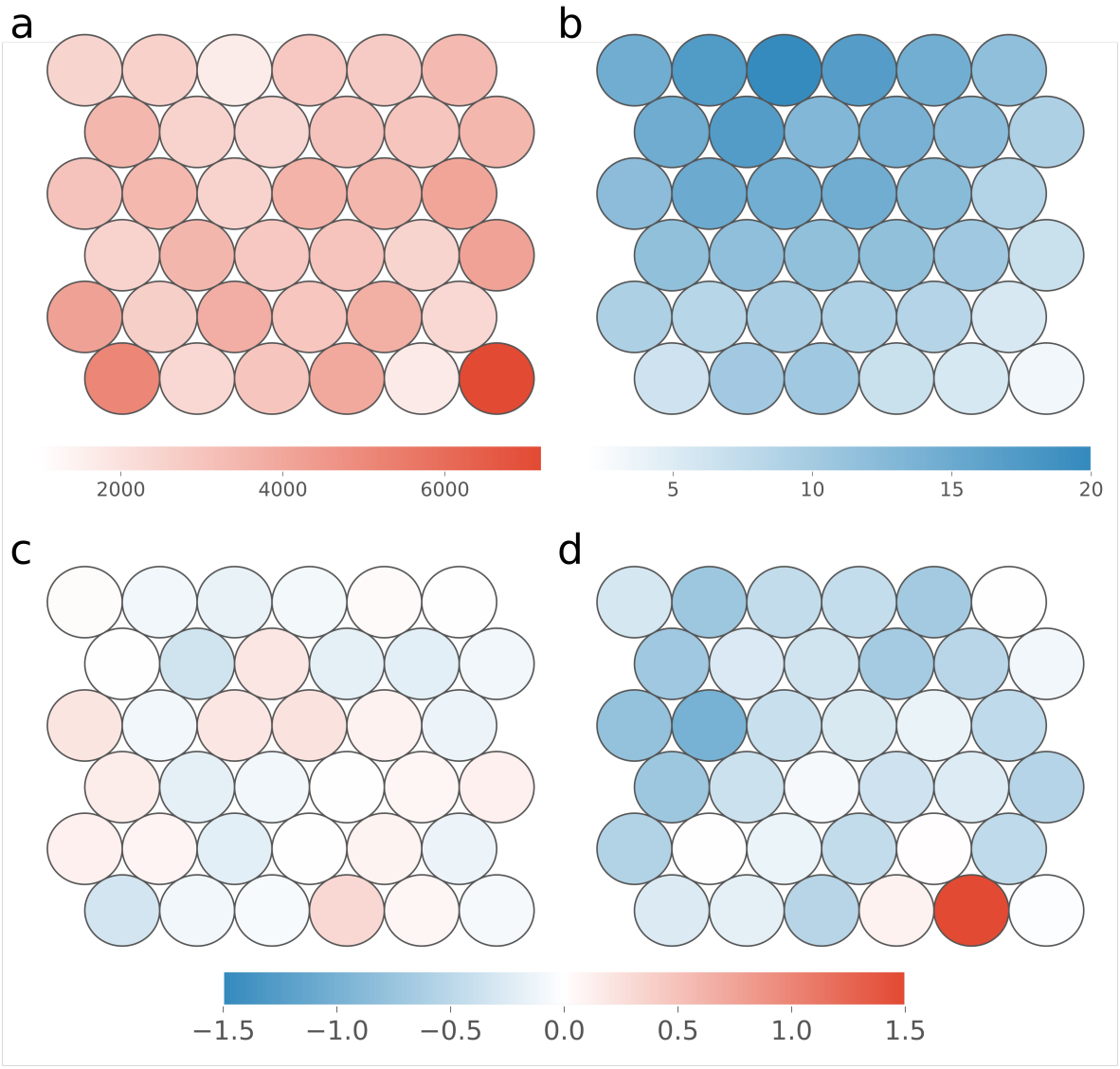

Figure 4.6: Self Organising Maps demonstrate NMM downregulated genes contain specific PG4 types. Self Organising Map (SOM) plots for clustering of Quadparser predicted two tetrad PG4s by loop length. In each figure, each circle represents a cluster of similar PG4s. **a)** SOM plot coloured by cluster 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. 4.7). 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.

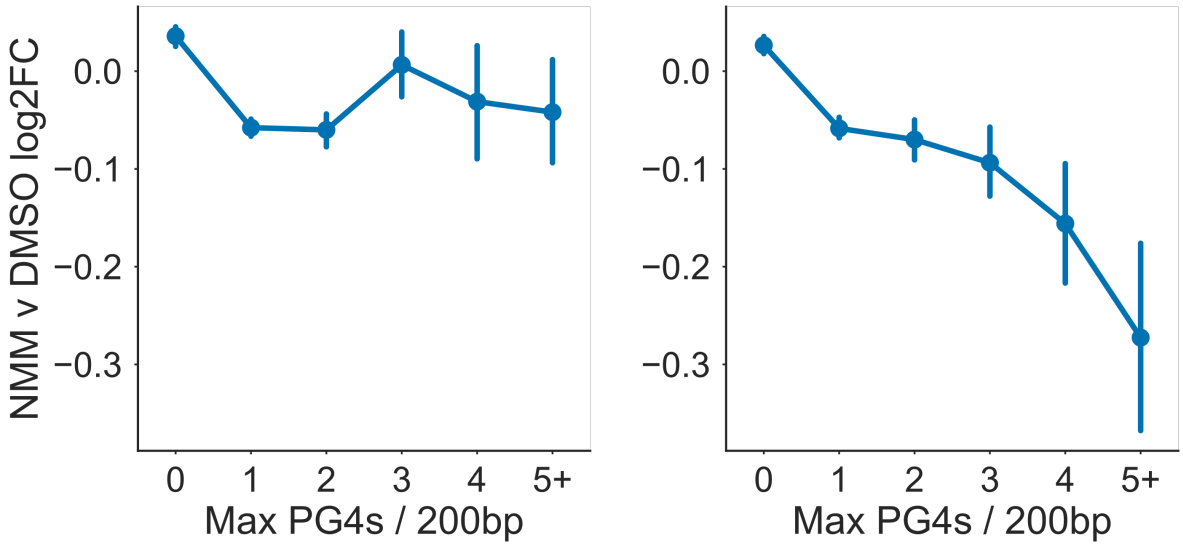

Figure 4.7: NMM regulates genes with large maximal PG4 density Point plot showing mean log fold change in gene expression during NMM treatment for genes binned by "maximal PG4 density", defined as the greatest concentration of PG4 motifs within a 200bp sliding window anywhere in the body of the gene (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 4.8, 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 4.8, 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.

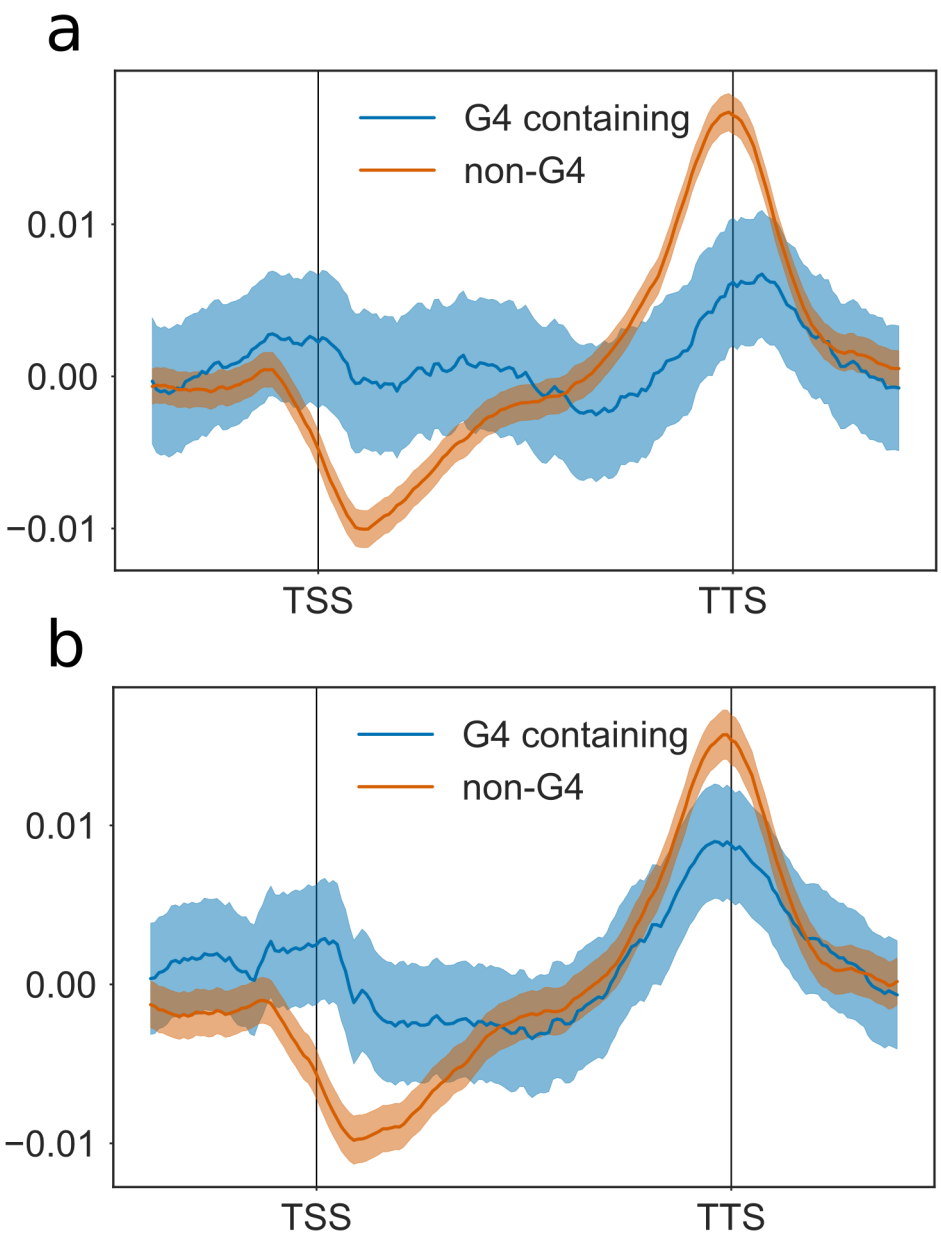

Figure 4.8: PG4 dense genes have altered RNA Polymerase II occupancy Metagene profiles showing RNA Polymerase II (Pol II) occupancy across binned exonic gene bodies. Profiles are made up of 500bp upstream region (in 10bp bins), 100 gene body bins, and 500bp downstream region (in 10bp bins). **a)** Metagene 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. 4.9a). 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. 4.9b, 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. 4.9b, 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. 4.9c). 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.

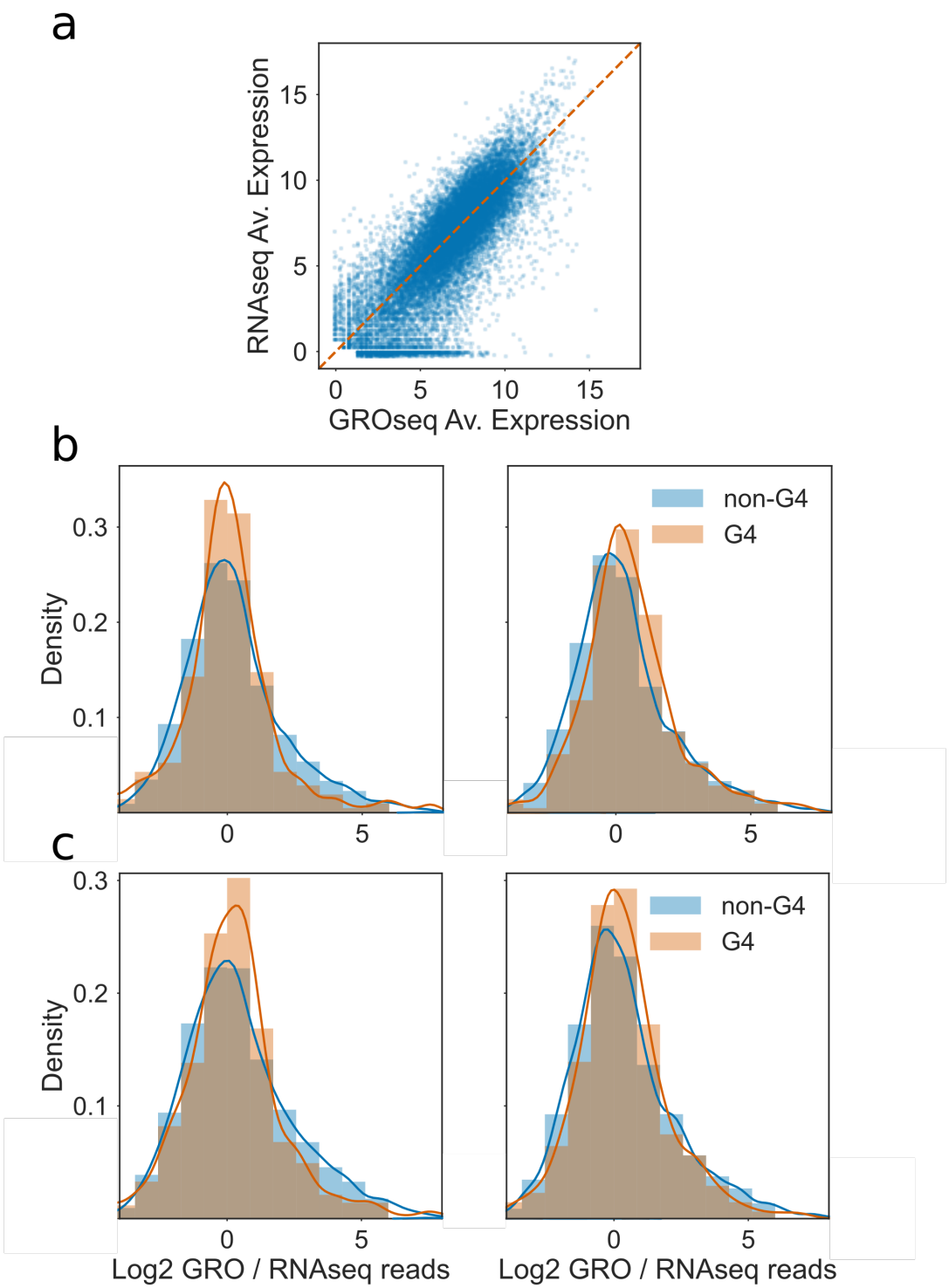

Figure 4.9: Log Ratio of GRO / RNA seq counts per million detects abortive transcription of PG4 dense genes. a) Scatter plot showing measured expression in log2 counts per million (logCPM) for each gene from GRO-seq vs. RNAseq datasets. Genes which fall below and to the right of the orange line have 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.

NMM treatment does not activate DNA damage response pathways.

Rodriguez et al. showed that treatment of human cells with Pyridostatin (PDS), a G4 binding agent, caused an increase in the histone marker H2AX, which is associated with double stranded DNA breaks. Cells in gap phases 1 & 2, during which DNA is not being replicated, showed a reduction in double strand breaks when also treated with 5,6-dichloro-1- -D-ribofuranosylbenzimidazole, an inhibitor of transcription. This suggests that PDS causes transcription and replication dependent DNA damage, presumably resulting from Pol II and replication fork stalling at G4 loci. To see if NMM causes similar DNA damage in *Arabidopsis thaliana*, we looked for transcriptional signatures of damage response. A set of 7 genes: AGO2, PARP1, RPA1E, BRCA1, GRG, RAD51, and RAD17, which have been reported as strong transcriptional markers of DNA damage (Ryu et al. 2018) were tested. Only AGO2 showed upregulation (logFC=1.23, PPLR=1) upon NMM treatment (Fig. 4.10a). We next compared data from a microarray in which seedlings were treated with Gamma Irradiation so see if there was any overlap with genes regulated by NMM. We found a small but significant overlap in upregulated genes (P = 2.9e-29). Gene ontology analysis of the overlapping gene set yielded no clear results, however. In comparison, the total gamma irradiation upregulated set was strongly enriched for genes involved in double strand break repair (FDR = 9.68e-4). Taken together, this data suggests that NMM treatment does not cause major DNA instability or DNA damage response.

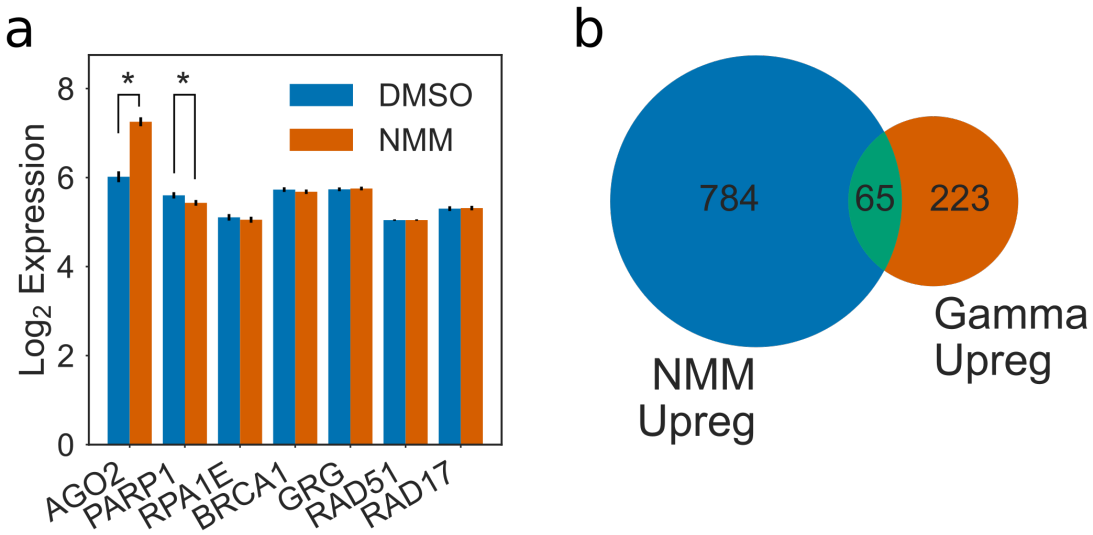

Figure 4.10: NMM does not activate DNA damage response a) Barplot showing normalised log2 expression of seven DNA damage responsive genes: AGO2, PARP1, RPA1E, BRCA1, GRG, RAD51, and RAD17, in the presence and absence of NMM. Errorbars are standard errors calculated by `puma`. Asterisks denote significant results. **b)** Venn diagram showing the overlap between NMM upregulated genes and genes upregulated in response to gamma irradiation.

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. Around 8% more genes than expected contained at least one template PG4 in the gene body (p=1e-11) (Fig 4.11). This enrichment appeared specifically in CDS and 5' UTR regions (Fig. 4.12b-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. 4.12d).

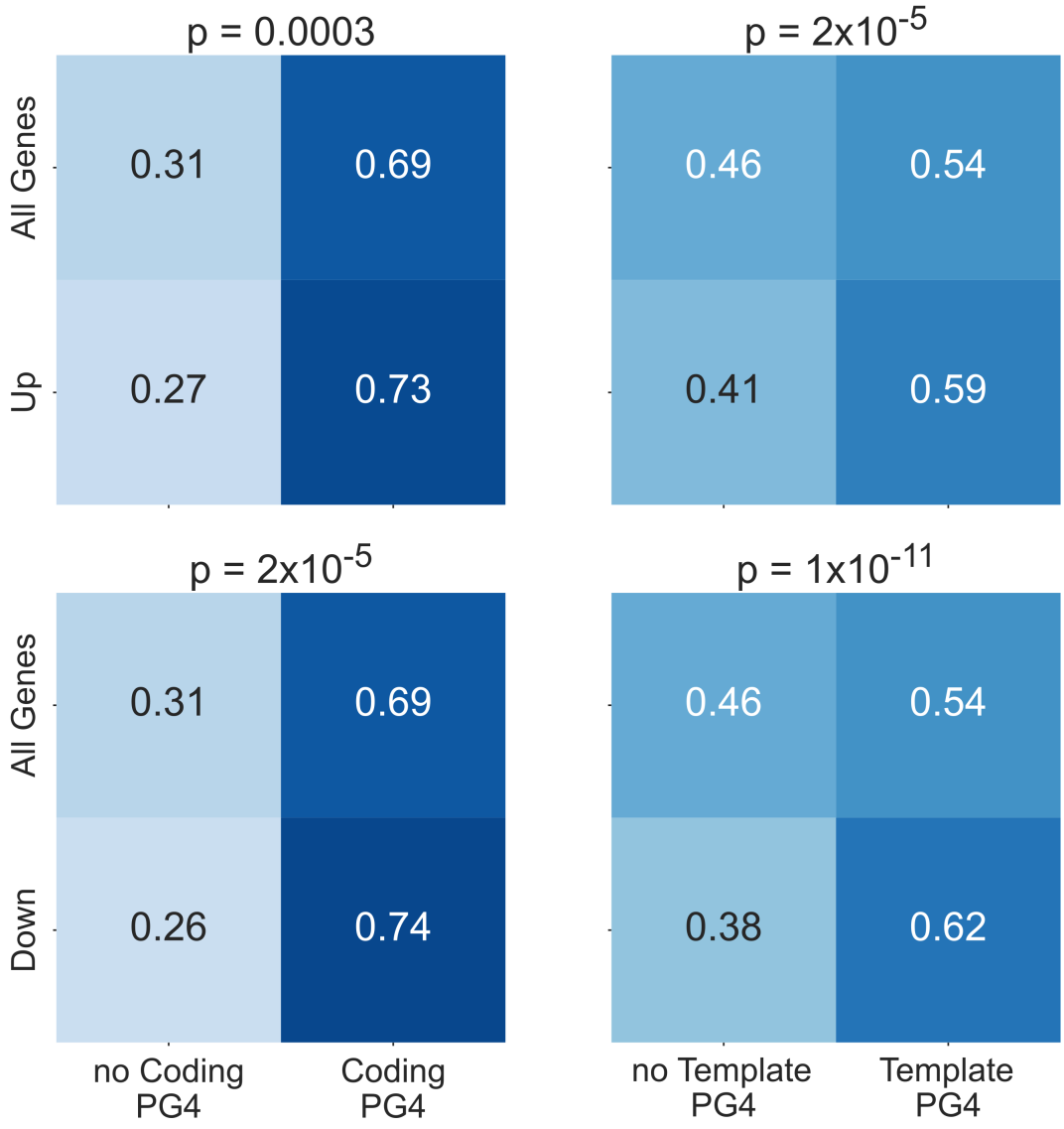

Figure 4.11: Drought downregulated genes are enriched in PG4s Heatmaps showing fractions of genes containing at least one predicted G4 in their gene body for drought upregulated genes vs all genes (top row) and drought downregulated genes vs all genes (bottom row) (For down and upregulated genes, FDR < 0.05 and absolute logFC > 0.5). PG4 predictions for the coding strand are in the left hand column whilst PG4 predictions for template strand are on the right. P values for each heatmap are calculated using Chi-squared tests. Genes downregulated by drought stress show a particularly strong enrichment of PG4s, and particularly on the template strand.

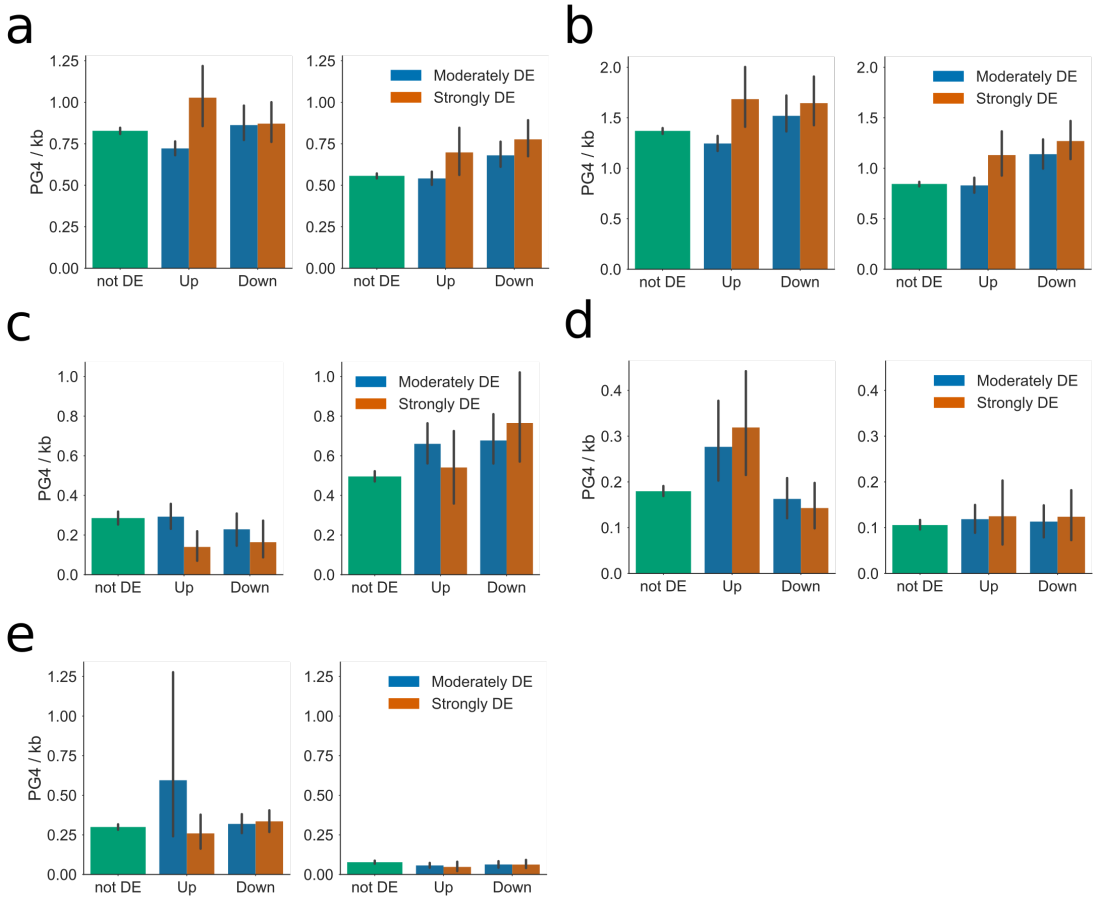

Figure 4.12: Distribution of PG4s in genes differentially regulated by Drought stress. Bar plots showing the average PG4 densities of genes up or downregulated by Drought stress, for **a)** full gene body (exons and introns), **b)** coding regions, **c)** 5' UTR, **d)** 3' UTR, and **e)** introns, respectively. 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 ($p = 1.7e\text{-}196$) (Fig 4.13a). 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 4.13b). This suggests that these genes could indeed be regulated through the same mechanism of G4 stabilisation.

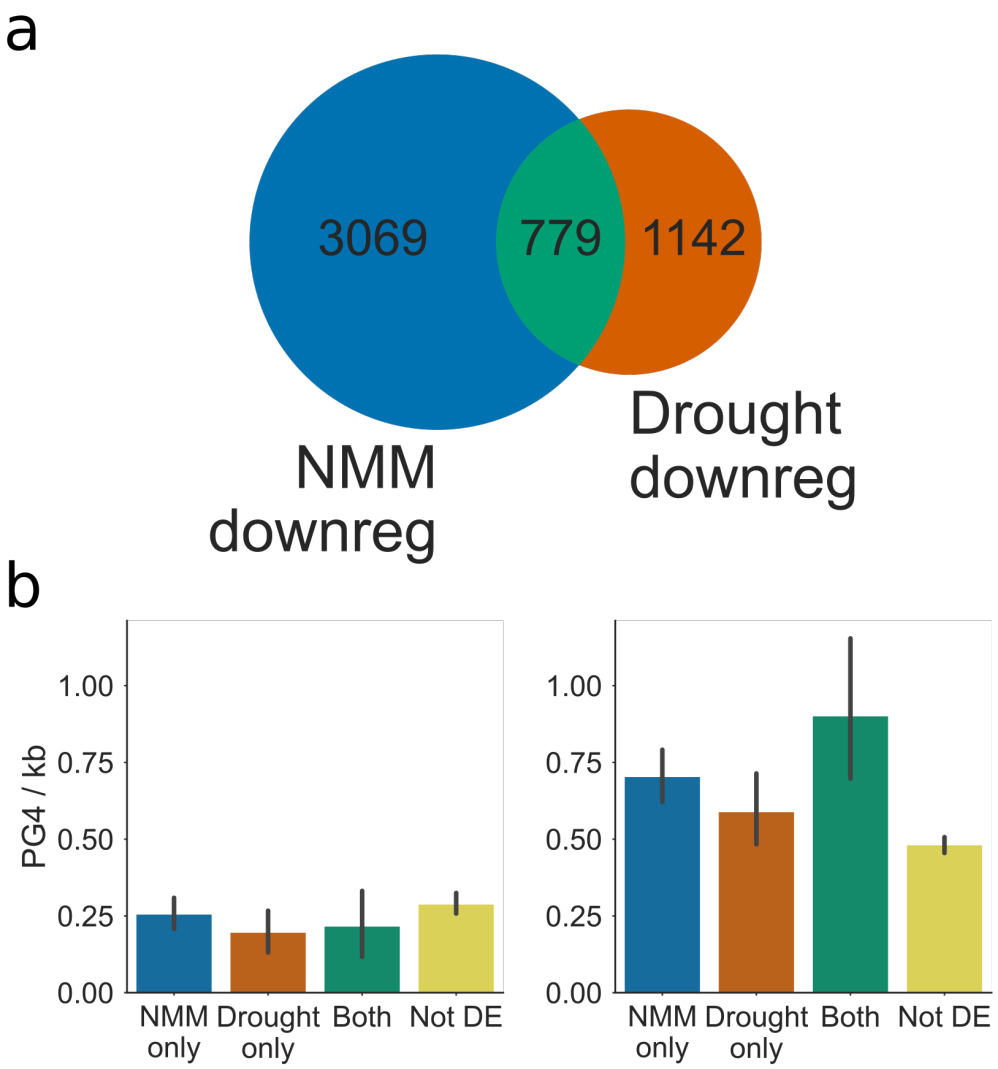

Figure 4.13: Overlap of genes downregulated by NMM with those downregulated by Drought stress. a) Venn diagram reporting the overlap of genes downregulated by NMM with those downregulated by drought stress. **c)** Bar plot showing the average PG4 densities in 5' UTRs of NMM and drought stress downregulated 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.

Chapter 5

Effect of G-Quadruplexes on expression and splicing of the Extensin gene family

Introduction

It is well characterised that G4s have the ability to stall polymerases, including both DNA and RNA polymerases. This has been demonstrated through *in vitro* Polymerase stop assays as well as by identifying transcriptionally dependent DNA breaks during G4 ligand treatment. Furthermore a number of helicases which are associated with the transcription initiation factor (TFIIH) and elongation, including BLM, WRN and XPD, have been shown to preferentially bind and unwind G4 DNA *in vitro*. XPD has also been associated with human TSSs containing PG4s by ChIP-seq. This evidence points to an effect of G4s on elongation by Pol II. We showed in **??** that PG4 dense genes appeared to have slower elongation of Pol II and hence increased Pol II occupancy at the start of the gene. This reduction in Pol II speed may have knock-on effects on co-transcriptional modifications, such as splicing of mRNA. Furthermore, if G4 stabilisation can be regulated, this could constitute a new mechanism for the regulation of splicing.

It has been estimated that around 80% of all splicing occurs co-transcriptionally in higher eukaryotes (Girard et al. 2012). This is due to the coupling of splicing to export and quality control mechanisms such as nonsense mediated decay. How splicing occurs is highly dependent on Pol II speed, since changes in speed can alter how strong and weak splice junctions compete for assembly of spliceosomes. The classic mechanism involving differential acceptor site usage is shown in (Fig 5.1a). When Pol II is elongating at high speed, acceptor sites which are more strongly canonical but further from the donor in sequence space are favoured. This is because on average, the greater strength of the canonical junctions outweigh the extra time that weaker junctions which are closer in sequence space have to be spliced. When Pol II elongates more slowly, however, weak acceptor sites which are more proximal may have much more time to be utilised before stronger distal acceptors are transcribed into the nascent RNA. This tips the balance towards utilisation of weak splice junctions, and can result in alternative splicing.

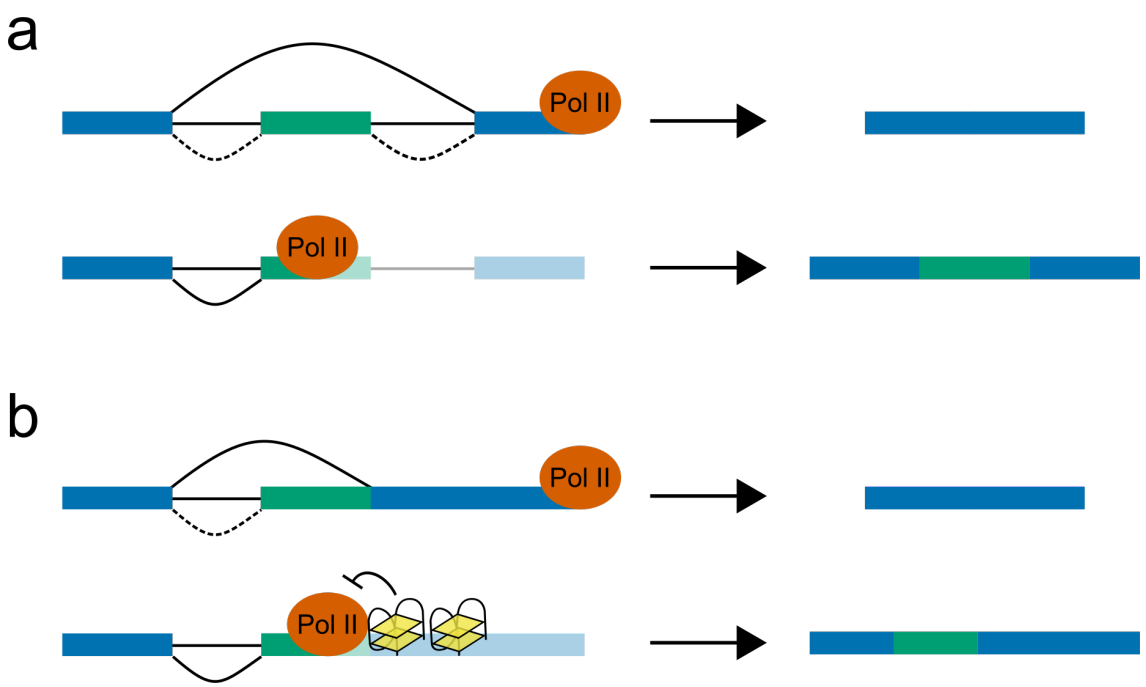

Figure 5.1: Pol II elongation speed affects co-transcriptional splicing a) Mechanism for effect of Pol II speed on co-transcriptional splicing: when elongation occurs rapidly (first row), the more canonical but distal acceptor site is used, resulting in the exclusion of the alternate exon (shown in green). When Pol II elongates more slowly (second row), there is more time for the weaker proximal site to be utilised, resulting in the inclusion of the alternate exon. **b)** Example mechanism for how G4s could affect splicing. When G4s are not present (top row), the constitutive splice acceptor is used and the green exonic chunk is excluded. When G4s are formed in the template strand of the DNA (second row), these slow down Pol II elongation, allowing the weaker proximal splice junction to be utilised, and including the alternate exon chunk. (

An unusual and only recently characterised form of alternative splicing is the exitron, which was named such because it involves the splicing out of constitutively exonic sequence. This often occurs in coding regions of the mRNA, resulting theoretically in an internal deletion from the protein. A study of exitrons in *Arabidopsis* has found them to be relatively common in long exons (Marquez et al. 2015). The study also found evidence of protein products from these exitrons, indicating that unlike intron retention events, which are thought to be regulatory or caused by splicing errors, exitronic transcripts are functional at the protein level.

Here we identify a novel family of PG4 rich genes, which are downregulated by NMM, and appear to be heavily exitronically spliced. We characterise this splicing and identify whether it can be altered through G4 stabilisation by NMM.

Methods

Plant growth conditions and drug treatments

All experiments used Arabidopsis thaliana Columbia (Col-0) ecotype and all mutant plants used were of Col-0 background. For all experiments seeds were surface sterilised, stratified for 2-3 days at 4°C then sown on vertical plates containing Murashige & Skoog (MS) agar with 1% sucrose and 0.8% agar and transferred to growth cabinets at constant light at 23°C (growth time refers to the time after transfer to growth cabinet). Seedlings used for qPCR and RNAseq were grown for 7 days on MS plates, treated for 6 hours by flooding the plate with MS liquid media containing the drug after which roots and shoots are harvested separately. Drugs used were N-Methyl Mesoporphyrin (Frontier Scientific, NMM580), Berberine (Sigma B3251) and Cyclohexamide (Sigma C7698). For combined Cyclohexamide and NMM treatment, plants were pretreated for 2 hours with Cyclohexamide before adding NMM and treating for a further 6 hours.

RNA extraction for qPCR and RNAseq

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). DNA contamination was removed using DNase I (Sigma AMPD1).

RNAseq analysis

RNA was sent to Sheffield Children's Hospital Genomics Facility for library preparation and sequencing. Polyadenylated RNA was enriched using NEBNext Poly(A) mRNA Magnetic Isolation, and libraries were produced using the NEBNext Ultra II Directional RNA kit for Illumina. The chemical fragmentation step was adjusted to increase estimated insert size to

the 400-450bp range. Paired end sequencing was conducted across two lanes of an Illumina HiSeq 2500 in rapid run mode, with 220bp read length. The run returned 157 million pass filter reads in lane 1, and 154 million in lane 2. Initial read preprocessing and adaptor trimming was conducted by the Genomics Facility.

Read quality was assessed locally using `FastQC`. Differential expression analysis was conducted by using `salmon` pseudoalignment to estimate transcript abundance against Araport11 cDNA and ncRNA. Mean insert size for each sample was assessed to be in the 400-500bp range. Gene level abundance was then aggregated from this using `tximport` and differential expression testing was conducted using `edgeR` linear modelling. Normalised log2 counts per million (CPM) were calculated and used for plotting. P values were adjusted using the Benjamini Hochberg multiple testing correction. Barplots of NMM and DMSO expression were produced using `seaborn`.

Spliced reads were mapped to the TAIR10 genome using `STAR`. The parameters for spliced mapping were adjusted to increase precision. A minimum of 8bp overhang was used for unannotated splice junctions, and 5bp for annotated splice junctions (following ENCODE guidelines). Minimum intron size was set to 60bp and maximum intron size to 10000bp. Output BAM files were sorted and indexed using `samtools`.

Gene Ontology analysis

For PG4 enrichment Ontology analysis, two tetrad Quadparser PG4s were predicted in the TAIR10 genome using `g4predict`. The number of PG4s overlapping each strand of the flattened exon models for each gene in Araport11 was calculated using `bedtools intersect`. To calculate enrichment, a permutation test experiment was conducted, where each gene was assigned a weighting proportional to the total length of its exonic sequence. PG4s were then shuffled randomly amongst all genes. For each Ontology group, the number of PG4s observed in genes of that group was compared to the expected numbers if PG4s were distributed randomly across the transcriptome. 10000 permuations were used for testing, and two tailed P values were calculated as $max(min(\frac{\sum_{i=0}^{n} exp_i<obs}{n}, \frac{\sum_{i=0}^{n} exp_i>obs}{n}), \frac{1}{n})$, where n is the total

number of permutations and exp_i is the expected value from the ith permuation. P values were adjusted using the Benjamini Hochberg multiple testing correction.

For Gene Ontology analysis of differentially expressed genes, `GOseq` was used. Up and down-regulated genesets were produced using a log2 fold change threshold of 1 and an FDR of 0.05. Weighting factors used were the median transcript length for each gene. P values for enrichment were produced by `GOseq` using the Wallenius approximation method and were corrected for multiple testing using the Benjamini Hochberg method.

Tables of enriched GO terms were generated using `pandas` and formatted with `inkscape`.

Extensin gene family total PG4 estimation

For estimated PG4 numbers in the table in Fig 5.4, PG4s were predicted using three different methods. All instances of the dinucleotide `GG` were identified in each Extensin gene, and then a graph was built (using `networkx`) where each `GG` was a node and nodes were connected by an edge if dinucleotides were less than 7bp apart from each other. The number of overlapping PG4 conformations was then calculated as the number of subgraphs in the graph with exactly four members, whilst the number of merged PG4s was calculated as the number of unconnected subgraphs with four or more members. To identify the number of non-overlapping PG4s, a dynamic programming method was used. Overlapping PG4s were grouped and scored by inverse length, then filtered for the maximum number of high scoring, non-overlapping PG4s. Code for these calculations is available in Appendix XXXX.

Quantitative PCR experiments

Total RNA was reverse transcribed into cDNA using the High Capacity cDNA Reverse Transcription Kit with a PolyT primer (Invitrogen Cat. No. 4368813). qPCR was carried out using SYBR® Green JumpStart™ Taq ReadyMix™ (Sigma-Aldrich Cat. No. S4438) using the Mx3000P qPCR System (Agilent Technologies). Thermal cycling conditions consist of a denaturation step at 94 °C for 2 minutes, 40 cycles of 15-second denaturation at 94 °C and

1-minute extension at 60 °C, then a final dissociation step of 2 minutes at 94°C, 1 minute at 60 °C and 2 minutes at 94 °C.

Analysis of public RNAseq data

Root RNAseq from Li et al 2016 was downloaded in FASTQ format from ENA. Quality assessment was performed with `FastQC` and `fastq-screen`, and adapter contamination was removed using `Cutadapt`. The data was mapped using `STAR` with default settings, except than max intron length was set to 10000bp. Output BAM files were sorted and indexed using `samtools`.

Normalised gene expression estimates were generated using `featureCounts` to get raw counts of alignments overlapping each gene in Araport11, and `edgeR` to perform estimation of log2 counts per million.

Splice junction analyses

Splice junction sites were extracted from aligned reads using `pysam`. For all analyses, reads were filtered to produce a set of unique donor/acceptor site pairs. Scatter plots of spliced read percentages and frequency plots of in frame exitrons were produced using `matplotlib` and `seaborn`.

Sequence logo generation

Consensus sequence logos were generated using an in-house Python module `matplotlib_logo`. Unique splice junction pairs from spliced reads were identified, and the corresponding sequence information (8bp up and downstream of donor and acceptor) was extracted from the TAIR10 genome using `pysam`. Position frequency matrices were generated from these sequences, and entropy score in bits was calculated and plotted.

RNAseq read simulation and bootstrapping experiment

For RNAseq read simulation, read counts generated using `featureCounts` for each file in the Li et al. 2016 dataset were used to inform a `polyester` simulation of Illumina 125bp paired end reads, using the Araport11 annotation. Fragment length was simulated using a normal distribution with mean 250bp and standard deviation of 25bp. Error rate was simulated at a uniform 0.5% across samples, reads, and nucleotides.

For the bootstrapping experiment, one or more pairs of real and simulated samples were randomly sampled from the dataset and unique splice donor/acceptor pairs in EXT9 were identified. Only splice sites from primary alignments, with overhangs greater than 20bp, and with more than 2 suppporting reads were kept. For each splice site, the exonic overhang sequence 20bp upstream of the donor site and 20bp downstream of the acceptor site were extracted from the read and concatenated. Any splice site whose concatenated sequence was present as a contiguous kmer in the reference EXT9 sequence was assumed to be a mapping error and filtered out. Finally, splice sites were deduplicated by directional clustering of sequences with edit distance of one or less using `umi_tools` directional clusterer. 500 iterations were used for bootstrapping. 67% condifence intervals were produced and plotted using `seaborn`.

Mappability analyses

Mappability scores were generated for the TAIR10 genome using `gem-mappability` with a kmer size of 75bp, and converted to `BigWig` format using `gem-2-wig` and `wigToBigWig`. Minimum mappability scores for each Extensin gene were extracted using `pyBigWig` and plotted against spliced fraction using `matplotlib`.

Sanger sequencing analysis

For sanger sequencing, cDNA was produced using the Invitrogen High Capacity cDNA Reverse Transcription Kit with a PolyT primer. Gene specific primers were used to amplify transcripts

of interest by PCR. Products were then cloned using the Thermo Scientific CloneJET PCR cloning kit, and transformed into DH5alpha competent cells. Colonies were grown overnight on ampicillin and tested by colony PCR. Picked colonies were then grown up for a further 24 hours in liquid media before miniprepping with Qiagen QIAprep Spin Miniprep Kit. Plasmids were sent, with pJet1.2 forward and reverse sequencing primers, for Sanger sequencing at the Sheffield Hallamshire Hospital Core Genomics Facility.

Differential Splicing Analysis

For differential splicing analysis, we identified novel splice junctions in our RNAseq dataset to augment the existing Araport11 annotation. Splice junctions for each gene were extracted from reads mapped to correct strand of the annotated gene body using Python and `pysam`. Junctions were kept if they were supported by at least 20 reads across all samples. Differential splice junction usage between NMM and DMSO treatments was then conducted in R using `limma-voom` and `DiffSplice`, and P values were adjusted using Benjamini Hochberg correction. Differentially utilised junctions were identified using an absolute log2 fold change in expression of 0.5 and an FDR of 0.2.

To produce junction categories based on relation to reference annotation, Araport11 GTF annotation was flattened using the python module `CGAT.GTF` to produce a single model for each gene, which was converted to bed12 format. This was then used to identify junctions which shared one or both of their donor and acceptor sites with reference introns. These were labelled "alternate" and "constitutive" junctions, respectively. Constitutive junctions which spanned an internal exon where labelled "skipping" junctions. Junctions which were contained wholly within a single contiguous exonic region were labelled as "retained intronic / exitronic" (the distinction between these is whether retention or splicing of the region is more common). Finally, junctions which do not contain a donor or acceptor present in the reference, and which span a mixture of exonic and intronic regions, were labelled "other" junctions. Violin plots of distribution of junction type expression, and stacked barplots of class distribution amongst differentially utilised junctions, were produced using `seaborn` and `matplotlib`.

For barplots of spliced read percentages in Extensin genes, `pysam` was used to extract all uniquely mapped and properly paired reads covering each gene. Each read was counted separately (i.e. pairs were not counted as one fragment). Reads that were mapped across an exitronic splice junction were counted and divided by the total number of reads to get a percentage of exitronic reads. These percentages were compared between NMM and DMSO treatments.

Results

Gene Ontology shows plant cell wall specific genes are enriched in PG4s and downregulated by NMM

To identify gene ontology groups which are specifically enriched with exonic PG4s, we compared the observed levels of PG4s per gene to expected levels if PG4s were randomly distributed across all genes (weighted by gene length). These observed and expected levels were summarised for each gene ontology group. Sorting the results for groups with the greatest positive observed/expected ratio of PG4s on the template strand, we discovered that gene ontology groups involved with functions at the cell periphery, particularly in the plasma membrane and cell wall, had strong enrichments (Fig 5.2). The log2 fold enrichment in `GO:0005199`, which contains structural cell wall genes, was $+4.4$ (FDR $<$ 4.8e-4). This corresponded to an observed number of 992 PG4s in only 32 genes (the average expection under the null hypothesis was 46 PG4s). These PG4 dense gene ontology groups were also strongly enriched in the set of genes which are significantly downregulated by NMM in our RNAseq dataset (Fig5.2, 50% of expressed genes in `GO:0005199` were downregulated by NMM, FDR $=$ 9.6e-7).

			Coding Strand PG4 Log2 Enrichment	Coding Strand FDR	Template Strand PG4 Log2 Enrichment	Template Strand FDR	NMM vs DMSO Upregulated Geneset Enrichment	NMM vs DMSO Upregulated Geneset FDR	NMM vs DMSO Downregulated Geneset Enrichment	NMM vs DMSO Downregulated Geneset FDR
GO:0005199	*functions as*	**structural constituent of cell wall**	-0.77	8.7e-04	+4.40	4.8e-04	-	0.95	+	9.6e-07
GO:0009664	*involved in*	**plant-type cell wall organization**	-0.22	0.17	+3.28	4.8e-04	-	0.95	+	3.1e-10
GO:0016722	*has*	**oxidoreductase activity**	+0.15	0.26	+1.40	4.8e-04	+	0.29	+	0.97
GO:0031225	*located in*	**anchored component of membrane**	+0.01	0.44	+1.35	4.8e-04	+	0.053	+	0.11
GO:0046658	*located in*	**anchored component of plasma membrane**	-0.05	0.37	+1.17	4.8e-04	+	0.95	+	0.97
GO:0000977	*has*	**RNA polymerase II regulatory region sequence-specific DNA binding**	+0.44	0.033	+1.09	4.8e-04	-	0.95	-	0.97
GO:0006869	*involved in*	**lipid transport**	-0.37	0.0098	+1.07	4.8e-04	+	0.68	+	0.0028
GO:0004721	*has*	**phosphoprotein phosphatase activity**	+0.01	0.45	+0.98	4.8e-04	-	0.95	-	0.92
GO:0001228	*has*	**transcriptional activator activity**	+0.52	0.0032	+0.95	4.8e-04	-	0.91	+	0.52
GO:0008233	*has*	**peptidase activity**	-0.26	0.033	+0.91	4.8e-04	-	0.95	-	0.97

Figure 5.2: Gene Ontology groups enriched in template stranded PG4s Table showing the top ten Gene Ontology groups most enriched for exonic PG4s compared to null distribution. The top two groups, both containing genes involved in cell wall structure and organisation, are also enriched for genes downregulated by NMM.

The proline rich Extensin gene family contain large numbers of hard-coded PG4s

We discovered that the GO:0005199 geneset was primarily made up of genes from the Extensin cell family (29/32 genes, 90.6%), including classical SP4/SP5 Extensin genes and chimeric Leucine Rich Repeat/Extensin (LRX) genes. These genes were found to be extremely PG4 rich on the template strand, with many genes containing greater than 10 PG4s per kilobase of exon (Fig 5.3a). Upon visualisation of these genes, we noted that in the majority of cases the PG4s were regularly spaced along the gene, and were contained solely within the coding region (CDS) of the gene (Fig 5.3b).

Figure 5.3: Expression and PG4 density of genes in the Cell Wall Structural Ontology group `GO:0005199` a) Panels showing gene expression (top panel) and PG4 density (bottom panel) for genes in the `GO:0005199` group. Expression in DMSO (blue) and NMM (orange) conditions is shown at log2 counts per million (from root RNAseq dataset). Errorbars are standard deviation of three biological replicates. Genes which are differentially expressed with FDR < 0.05 are labelled with asterisks. In PG4 panel, the exonic PG4 density per kilobase is shown separately for coding (blue) and template (orange) strands of the gene. **b)** Gene tracks showing the location of predicted two tetrad PG4s in orange for (from top to bottom) LRX1, EXT13, and EXT9. Gene models from Araport11 are shown in blue. In gene models, thin boxes represent untranslated regions (UTRs), fat boxes represent coding regions (CDS), and connecting lines represent intronic regions.

From a search of the literature, we discovered that Extensin genes are highly repetitive, proline rich proteins. These proteins polymerise to function as a structural matrix in the protein component of the plant cell wall. In particular, we noted that these proteins are characterised by large numbers of the SP3-5 repeat, which is made up of the sequence $Ser(Hyp)_{3-5}$, where Hyp is Hydroxyproline, a proline derivative. Since the codon for proline is `CCN`, the DNA which encodes SP4 and SP5 motifs will conform to the two tetrad Quadparser motif on the template strand of the gene (Fig 5.5a). This is the source of the PG4 density of Extensin genes, and PG4 counts in these genes are well correlated with SP3-5 repeats (Fig 5.4). Since the SP4 motif is required for the function of the protein, and is restricted by the codon for proline, these PG4s are "hardcoded" into the body of the gene.

AGI Identifier	Gene Symbol	Class	Organ Specific Expression	SP3 motif	SP4 motif	SP5 motif	Overlapping PG4	Merged PG4	Gene Length	NMM vs DMSO logFC	logCPM	FDR
AT1G26240	EXT20	SP5	Roots	2	1	40	186	84	1773	3.72	6.80	1.1e-168
AT1G26250	EXT21	SP5	Roots	7	0	28	139	69	1627	-3.40	5.57	5.1e-10
AT4G08370	EXT22	SP5		3	1	13	269	42	1353			
AT4G13390	EXT18	SP5/SP4	Roots	0	14	8	207	47	1576	-2.63	5.47	8.4e-28
AT5G19810	EXT19	SP5/SP4	Roots	0	4	13	185	45	1150			
AT1G23720	EXT6	SP4	Roots	2	61	3	234	94	3327	-2.27	8.84	2.4e-30
AT2G24980	EXT7	SP4	Roots	3	37	0	246	64	1914	-2.64	5.51	1.6e-30
AT2G43150	EXT8	SP4	Roots	0	22	0	160	32	1494	-0.87	11.48	1e-09
AT3G28550	EXT9	SP4	Roots	3	70	0	421	100	3457	-2.12	8.79	3.9e-26
AT3G54580	EXT10	SP4	Roots	2	68	0	361	117	3347	-2.21	8.61	2.6e-27
AT3G54590	EXT2	SP4	Roots	2	51	0	204	88	2656	-2.60	7.09	4e-42
AT4G08400	EXT11	SP4	Pollen, Roots	2	31	0	612	60	1769	-3.08	4.79	1.5e-22
AT4G08410	EXT12	SP4	Roots	2	41	0	823	81	2391	-3.01	5.60	2.6e-44
AT5G06630	EXT13	SP4	Roots	1	29	0	258	53	1544	-2.98	4.64	9.2e-32
AT5G06640	EXT14	SP4	Roots	2	42	0	762	68	2331	-2.71	6.23	2.6e-51
AT5G35190	EXT15	SP4	Roots	2	12	2	24	20	1274	-3.15	5.41	1.8e-34
AT5G49080	EXT16	SP4	Roots	0	41	0	151	59	2026			
AT1G21310	EXT3/5	SP4/SP3	Radicle, Roots	13	27	1	297	56	1765	-0.72	12.50	4.3e-06
AT1G76930	EXT1/4	SP4/SP3	Roots	8	9	0	205	49	1820	0.67	10.01	8.2e-05
AT4G08380	EXT17	SP3	Roots	34	2	0	41	5	1314			
AT1G02405	EXT30	Short	Siliques	0	3	0	45	14	754			
AT1G23040	EXT31	Short		0	2	0	25	11	1935	-0.04	4.93	0.75
AT1G54215	EXT32	Short		0	1	1	26	26	853			
AT1G70990	EXT33	Short	Roots	0	2	0	6	6	966	-0.80	3.21	1.2e-12
AT3G06750	EXT34	Short		0	1	1	6	6	901	0.13	3.14	0.34
AT3G20850	EXT35	Short	Roots	1	0	1	5	5	588			
AT3G49270	EXT36	Short	Siliques	0	2	0	12	8	1123			
AT4G16140	EXT37	Short		0	1	1	9	9	1121	-0.61	4.12	8.7e-10
AT5G11990	EXT38	Short		4	0	1	9	9	787	0.57	1.35	0.0081
AT5G19800	EXT39	Short	Roots	0	0	3	38	15	657	-3.04	2.14	5.2e-35
AT5G26080	EXT40	Short	Roots	2	1	3	59	13	674			
AT5G49280	EXT41	Short		0	2	0	26	8	1399	0.01	5.21	0.96
AT1G12040	LRX1	Chimeric	Roots	1	17	7	96	47	2690	-3.14	5.63	7.5e-46
AT1G49490	PEX2	Chimeric	Pollen	1	13	1	22	22	2952	-0.99	1.95	6.5e-10
AT1G62440	LRX2	Chimeric	Roots	4	12	6	96	45	2361	-1.43	4.27	4.5e-16
AT2G15880	PEX3	Chimeric	Pollen	2	16	9	143	43	2806	-0.35	3.35	0.011
AT3G19020	PEX1	Chimeric	Pollen	1	19	5	51	39	3414	0.14	3.21	0.34
AT3G22800	LRX6	Chimeric	Root	1	0	2	36	36	1938	-1.20	6.23	2.1e-10
AT3G24480	LRX4	Chimeric		2	1	3	30	12	1682	-0.35	7.38	0.0029
AT4G13340	LRX3	Chimeric		4	13	15	306	91	3088	-0.02	6.86	0.92
AT4G18670	LRX5	Chimeric		3	1	5	159	52	3333	-0.49	5.78	7.2e-09
AT4G33970	PEX4	Chimeric	Pollen	4	10	4	245	36	3108	-0.28	1.07	0.35
AT5G25550	LRX7	Chimeric	Stamen	1	0	1	3	3	1926			
AT1G10620	PERK11	Chimeric	Pollen	2	0	0	45	10	3321			
AT1G23540	PERK12	Chimeric	Pollen	1	2	0	17	9	3242	-1.51	0.74	1.3e-10
AT1G26150	PERK10	Chimeric		4	2	1	90	26	4228	-0.22	5.00	0.17
AT1G49270	PERK7	Chimeric	Pollen	1	4	1	17	13	3209			
AT1G52290	PERK15	Chimeric		0	0	0	2	2	2763	-0.01	0.85	0.98
AT1G70460	PERK13	Chimeric	Roots	3	2	2	97	16	3618	-2.35	4.36	7e-78
AT2G18470	PERK4	Chimeric	Pollen	1	0	1	9	5	3679	-0.62	1.50	0.028
AT3G18810	PERK6	Chimeric	Pollen	1	1	2	22	14	3291			
AT3G24540	PERK3	Chimeric		0	1	1	26	9	2849	-0.60	0.09	0.091
AT3G24550	PERK1	Chimeric		3	0	0	23	16	3474	0.18	7.34	0.078
AT4G32710	PERK14	Chimeric		0	0	0	42	20	3593	-0.95	3.91	2.8e-19
AT4G34440	PERK5	Chimeric	Pollen	2	0	0	3	3	3310			
AT5G38560	PERK8	Chimeric		5	2	2	31	13	4048	-0.03	5.76	0.77
AT3G11030	EXT50	Chimeric		0	5	0	11	11	3124	-0.36	5.51	2.2e-06
AT3G19430	EXT51	Chimeric	Root	0	7	0	60	21	2458	-5.92	3.60	3.9e-14
AT3G53330	EXT52	Chimeric		0	3	0	11	11	1009			
AT1G62760	HAE1	AGP/EXT hybrid	Pollen	2	0	2	6	6	1273	2.78	0.66	7.4e-34
AT3G50580	HAE2	AGP/EXT hybrid	Stamen	1	2	1	19	5	889			
AT4G11430	HAE3	AGP/EXT hybrid		2	0	2	47	47	1444			
AT4G22470	HAE4	AGP/EXT hybrid	Leaves	2	1	0	31	23	1317	-1.51	0.92	2.6e-06

Figure 5.4: The Extensin gene family contains large numbers of hardcoded PG4s Table showing extended Extensin gene family, their expression patterns, SP4 motif counts, PG4 counts and expression during NMM treatment. Adapted from Showalter et al. 2010

To demonstrate that the PG4 from Extensin genes could form a G4 structure in vitro we used circular dichroism spectroscopy (CD). We performed these experiments at physiologically relevant temperatures for Arabidopsis. An oligo representative of the SP4 repeat was designed (AGAGGTGGTGGTGGTATG) using 3bp flanks upstream and downstream of the PG4. CD showed the G4 oligo had peak absorbance at 260nm and trough at 240nm, indicative of a parallel G4 structure (Fig 5.5b). Removing the PG4 by mutating the sequence (mutated sequence: AGAGGTGATGGTGGTATG) or removing potassium ions from the buffer abolished this absorbance profile.

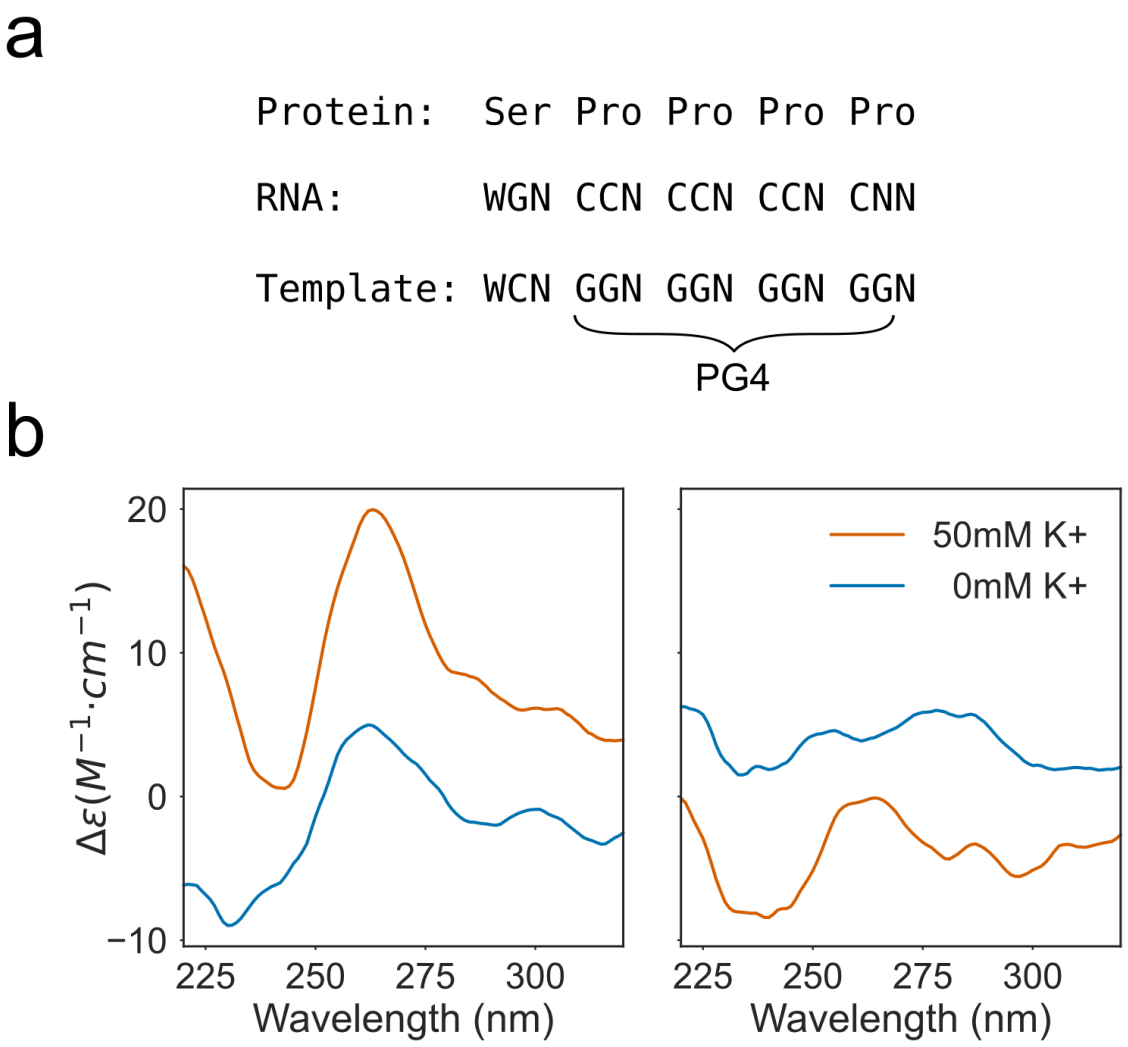

Figure 5.5: The Extensin SP4 motif forms a G-Quadruplex *in vitro*. a) Schematic showing how the Extensin SP4 protein motif hardcodes a two tetrad PG4 into the template strand of the gene body. **b)** CD spectroscopy of an Extensin repeat sequence (left) and a mutated control which does not conform the the Quadparser motif (right) show that the Extensin repeat forms a G4 *in vitro*. This is indicated by the peak in ellipticity at 260nm and the trough at 240nm, which are characteristic of a parallel G4.

Extensins are strongly downregulated by NMM and Berberine

To confirm that the Extensin genes are downregulated by NMM, we performed RNA extraction and quantitative RT-PCR (qPCR) on root tissue from 7 day old Arabidopsis seedlings treated for 6 hours with NMM at varying concentrations. EXT13 and LRX1 were chosen as representative classical and chimeric Extensins, respectively. The change in expression of both genes upon treatment was negatively correlated with the concentration of NMM applied (Fig 5.6a). Treatment with the G4 intercalating drug, Berberine, also caused strong downregulation of EXT13 and LRX1 (Fig 5.6b). Since NMM and berberine are very different drugs which stabilise G4s through different methods, taken together our results suggest downregulation of Extensins is caused by G4 stabilisation.

Downregulation of Extensins by NMM is translation independent

To confirm whether downregulation of EXT13 and LRX1 by NMM was direct, or the result of a perturbation the levels of a transcription factor, we conducted qPCR experiments with combinatorial treatment of NMM and Cyclohexamide (CHX). CHX is an inhibitor of translation which is commonly used to determine whether interactions by transcription factors on a gene are direct. If effects are indirect (i.e. if the transcription factor of interest regulates transcription of some intermediate transcriptional factor, which regulates the gene of interest), then treatment with CHX will prevent regulation, since any intermediate factors will not be able to be translated. In the case of NMM treatment, this was used to see whether NMM acts directly on EXT13 and LRX1 through G4 stabilisation, or through other changes in the transcriptome. Seedlings were pretreated with CHX for two hours, before NMM was added and treatment was continued for another 6 hours. This experiment showed that that Extensin downregulation by NMM still occurs even when translation is blocked, suggesting that NMM acts directly upon the Extensin genes.

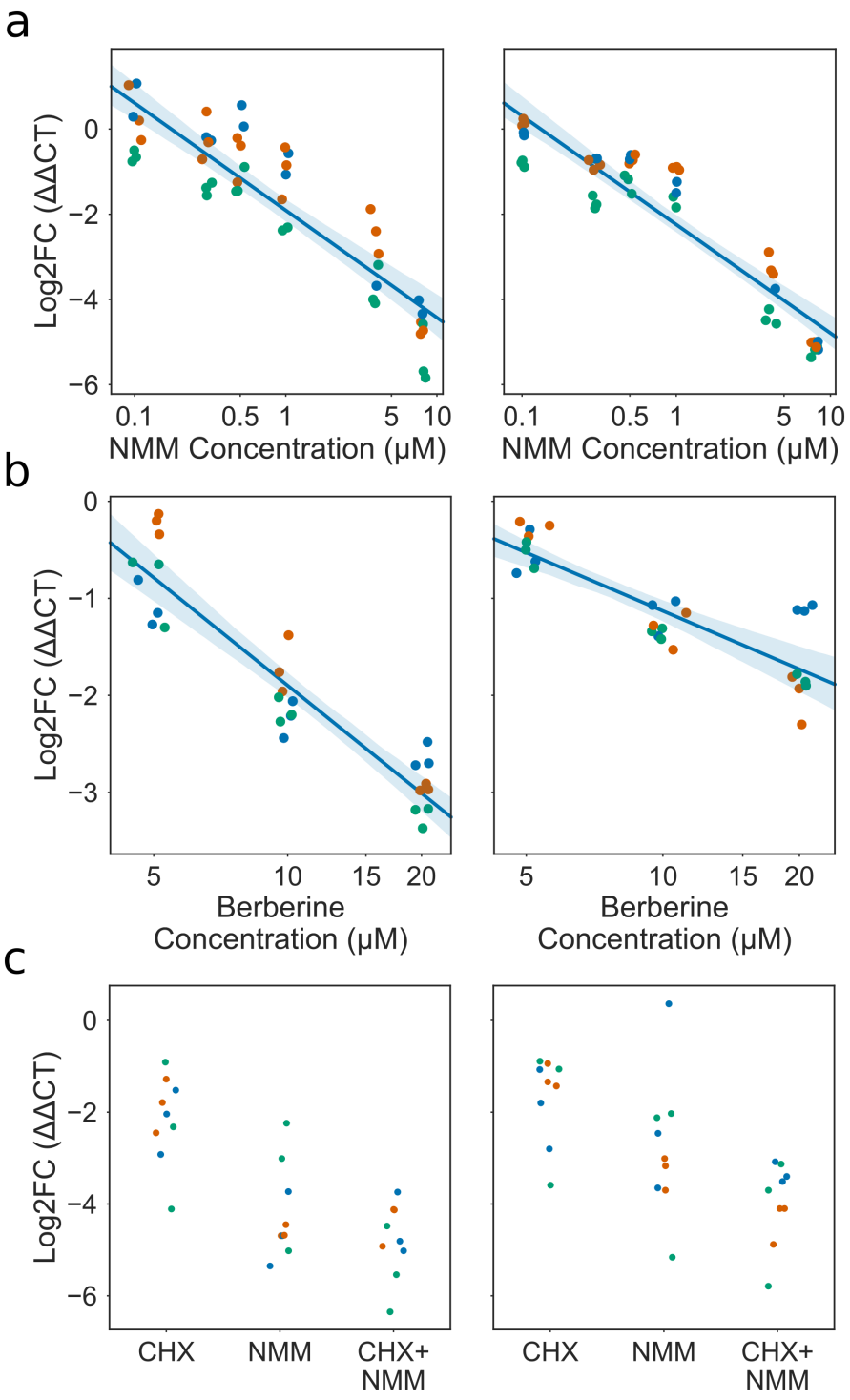

Figure 5.6: Expression of EXT13 and LRX1 during treatment with G4-binding ligands Scatter/strip plots showing qPCR results for EXT13 (left panels) and LRX1 (right panels). Log2 fold change in expression (ΔΔCT) of Extensin genes decreases with increasing concentrations of **a)** NMM and **b)** Berberine. **c)** NMM downreguation of EXT13 and LRX1 is not affected by concurrent Cyclohexamide treatment, suggesting a mechanism independent of translation. For all panels, each point is a single technical replicate, and colours represent different biological replicates. A small amount of jitter has been added to the X axis for better visualisation of results.

RNAseq suggests Extensin genes contain exitronic splice sites

We noted from studying *de novo* assembled splice isoforms from a root specific RNAseq dataset (Li et al. 2016) that many of the Extensin genes had large numbers of novel spliced isoforms. EXT9 was found to have the most novel spliced forms of any gene in the dataset. These were not present in the annotation. In fact, the majority of Extensin domains are annotated in both TAIR10 and Araport11 as intronless. These splice isoforms are presumably therefore a product of "exitronic" splicing (Marquez et al. 2015), where sections of constitutive exons, flanked on both sides by exonic sequence, are spliced out of a gene. We hypothesised that these unusual exitrons could be a result of slow Pol II elongation through PG4 dense regions, allowing splicing to occur at weak splice sites.

A hallmark of most true splice junctions is the GT/AG intron motif, which is the conserved canonical sequence in all higher eukaryotes, including Arabidopsis (Fig 5.7a). To determine whether Extensin exitrons had canonical splice motifs, we produced splice junction sequence logos for predicted introns from the dataset produced by Li et al., for EXT9 and LRX3, both of which were highly spliced. We found that splice junctions in these genes had near universal GT/AG motifs (Fig 5.7b). Upon inspection of the methods for Li et al., however we discovered that `CuffLinks` was used for *de novo* transcript assembly. Since the RNAseq dataset is unstranded, `Cufflinks` requires the upstream mapping tool (here, `STAR`) to annotate the orientation of spliced reads using the intron motif (i.e. positive strand for GT/AG and negative strand for CT/AC). This setting means reads which do not conform to the intron motif are discarded, leading to serious bias.

To remove this bias, we remapped reads from the Li et al. dataset using STAR without filtering by intron motif. Since assemblers like `CuffLinks` and `StringTie` require strandedness information derived from the intron motif, transcript assembly was not possible. Instead, we simply extracted spliced reads aligning to EXT3 and LRX3 and identified all unique splice site starts and ends. The corresponding sequences were then used to produce sequence logos (Fig 5.7c). These logos showed only a weak enrichment for the GT/AG motif in EXT9, and CT/AC in LRX1.

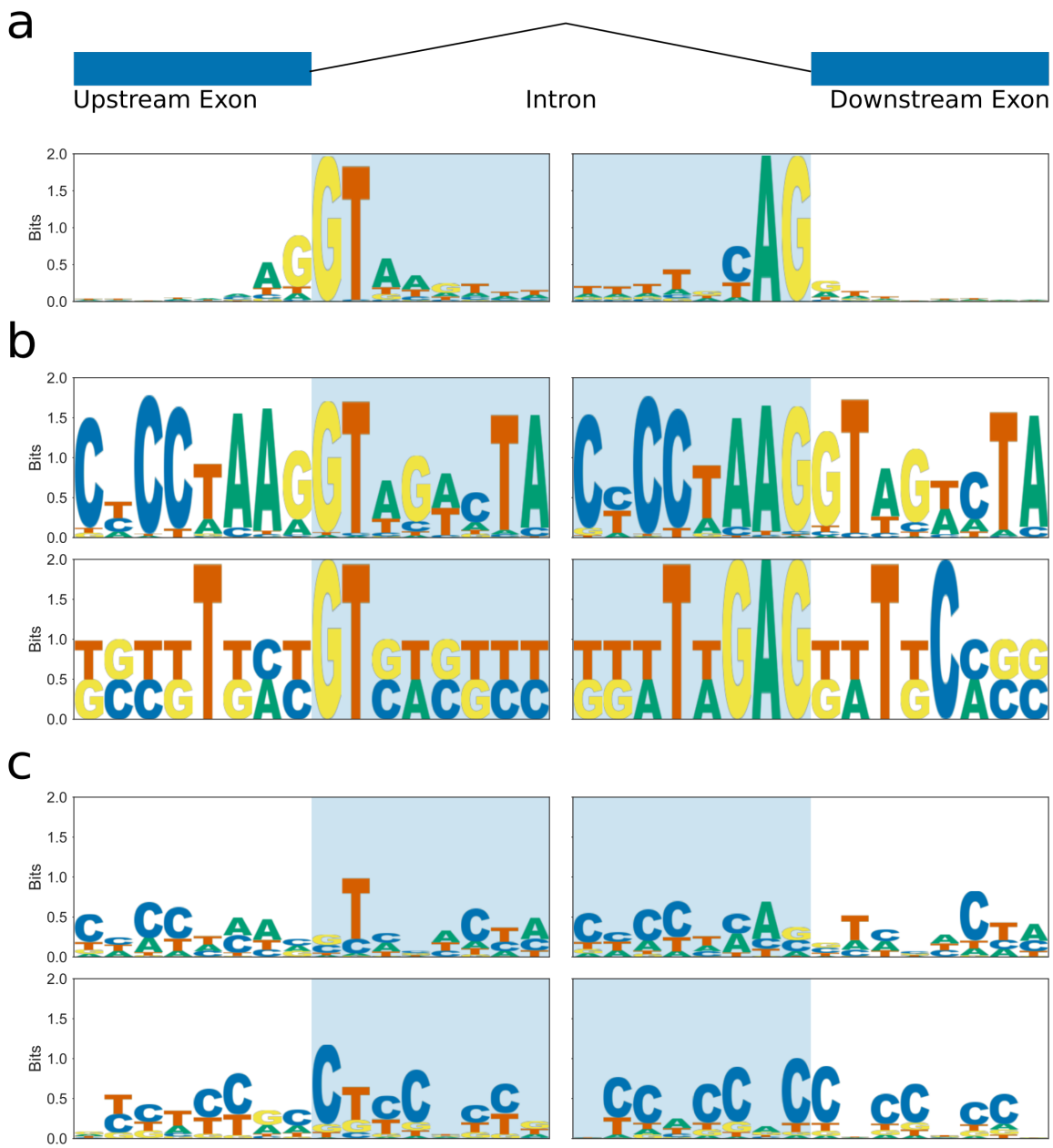

Figure 5.7: Splice junction motifs for EXT9 and LRX3 Sequence logo plots showing consensus splice site sequences around donor (left panels) and acceptor (right panels) splice sites. Putative intronic sequences are shown on shaded blue background. **a)** Splice junction consensus sequence logo for Arabidopsis, calculated from junctions in the Araport11 annotation. **b)** Splice junction consensus sequence logo produced from **de novo** assembled transcripts (Li et al. 2016) for EXT9 and LRX3. **c)** Splice junction consensus sequence logo produced from unique donor/acceptor pairs identified from spliced reads on EXT9 and LRX3.

Since the Extensin exitrons appear in coding regions of DNA, if the spliced out region is not a multiple of three, then the resulting mRNA would be frameshifted, producing truncated and potentially deleterious proteins. We therefore tested whether the spliced reads in EXT9 and LRX3 contained gaps which were multiples of three or not. For both genes, almost all of the unique splice junction pairs were multiples of three (Fig 5.8a-b). This could be evidence that these exitrons are genuine and produce function gene products. On the other hand, we noted that splicing tended to occur between regions with high protein and DNA sequence level homology. EXT9, for example, is an incredibly repetitive gene with high self-homology (Fig 5.9a). These in frame splice junctions could therefore simply be the result of mapping errors from the spliced aligner `STAR`, which utilises heuristics which may result in some reads from contiguous parts of the genome being mapped as spliced. If the homologous regions which could cause mapping errors within a gene also have homology at the protein level, as is the case in EXT9, then it is probable that erroneously spliced reads would be a multiple of three in intron length.

If spliced reads mapping to EXT9 were the result of some systematic error in mapping, one might not expect to see much variation in the percent of reads mapping to a gene being spliced. We therefore correlated the expression of EXT9 in each sample from the root RNAseq dataset (measured in log2 counts per million or needs to be!) with the percent of reads which mapped with a splice site. We found a slight positive correlation between expression and splicing (Fig 5.8c).

As a further precaution against these erroneous spliced mappings, we performed read simulation for each sample in the root RNAseq dataset. Put simply, the expression of each gene was quantified for each sample by counting the number of mapped reads, then `Polyester` (an Illumina sequencing read simulator) was run to generate reads from the reference transcriptome with the same read counts. These simulated reads were then remapped with `STAR` using the same parameters as the original mapping. We then performed a bootstrap analysis for EXT9 where we sampled one or more real/simulated sample pairs, and counted the number of unique splice donor/acceptor pairs that occurred in each. Junctions with the same exonic flanking sequence (using 20bp overhangs) or with edit distance of only one base were

collapsed. Any junctions with flanking sequence that appeared as a contiguous kmer in the reference sequence of EXT9 were also removed. Despite this, we saw a consistently larger number of unique donor/acceptor splice pairs in the real data than in the simulated data (Fig 5.8d).

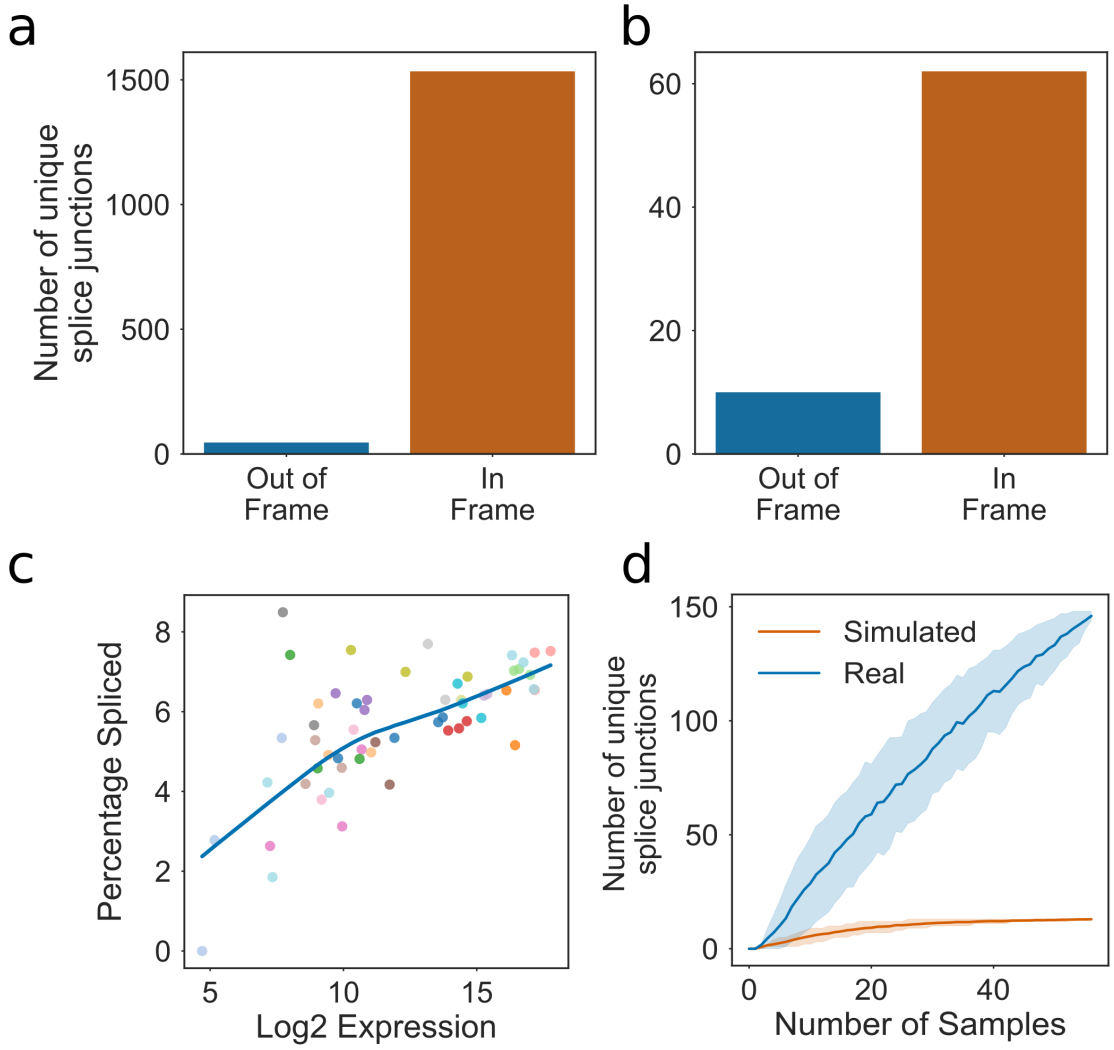

Figure 5.8: Splice junction motifs for EXT9 and LRX3 a & **b)** Frequency barplots showing number of In/Out of frame splice junctions for EXT9 and LRX3 respectively. **c)** Scatterplot showing percentage of reads with splicing versus log2 counts per million for EXT9. **d)** Bootstrapped splicing simulation showing number of unique EXT9 splice junctions discovered with increasing numbers of samples for real root RNAseq data versus paired simulated RNAseq data. Errorbars are 67% confidence intervals.

Since the Extensin genes are highly repetitive, this reduces the ability of read aligners to map to them. This "mappability" can be quantified using tools such as `GEM` which measure, for each genomic position, how often the sequence kmer that is found there occurs in the rest of the genome. We utilised `GEM` to score the mappability of the Arabidopsis genome, and compared the median mappability score for extensin genes to the percent of spliced reads for each gene. Only genes which were annotated as intronless (i.e. no constitutive introns, genes with exitrons were allowed) were included. We found a clear negative correlation between mappability and the number of mapped spliced reads (Fig 5.9b). For EXT9, the regions with lowest mappability are clearly those with the most annotated splice sites, including in the Araport11 reference (Fig 5.9c).

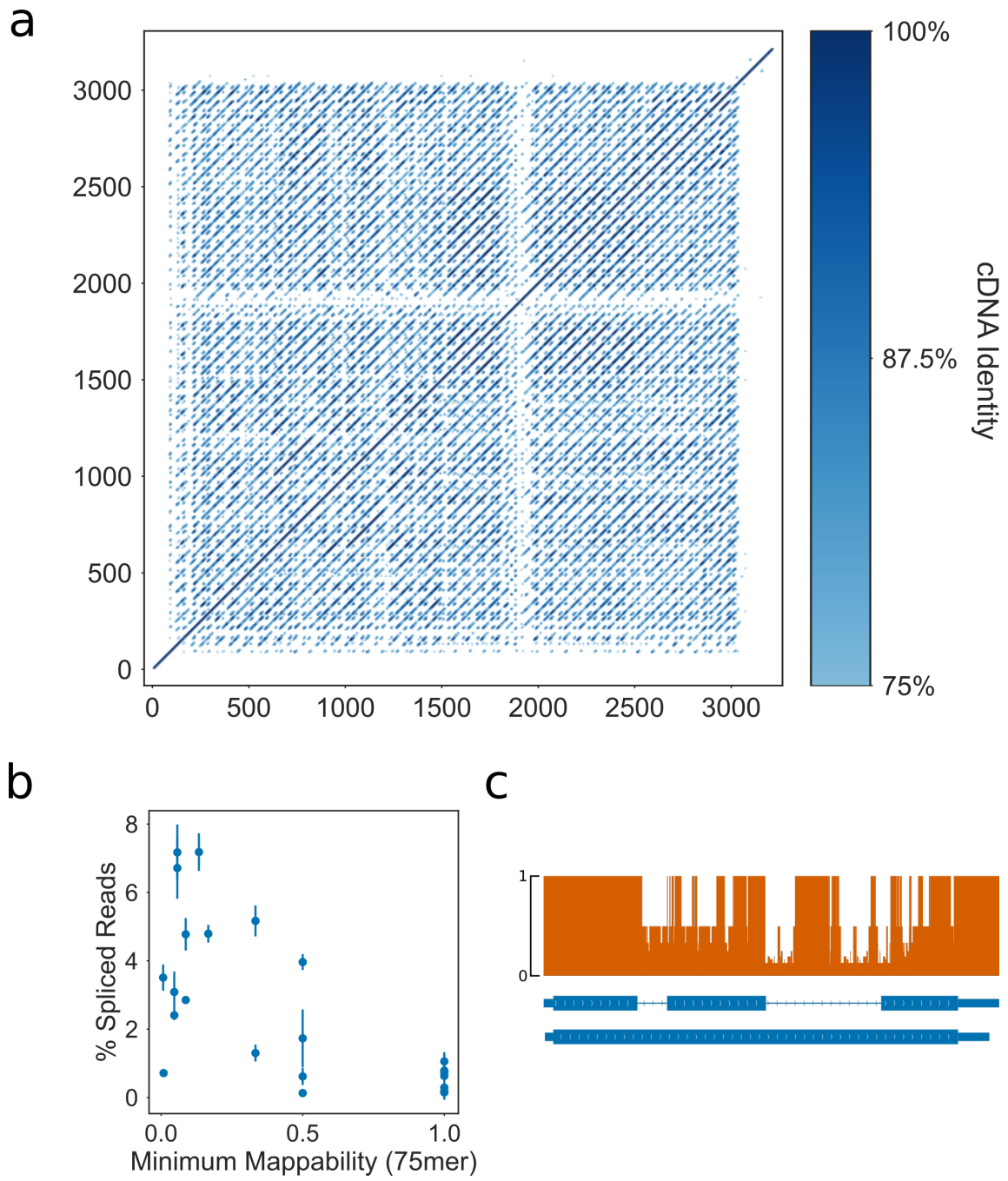

Figure 5.9: Extensin genes with greater spliced mapped reads have low mappability **a)** Dotplot showing self homology of the EXT9 cDNA (unspliced isoform). Positional identity was calculated using 15bp windows across the gene. Positions with identity less than 75% were filtered to remove noise from the plot. **b)** Scatter plot showing the minimum mappability score of unspliced Extensin genes against the percentage of spliced reads for that gene in mature root RNAseq samples (Li et al. 2016). Errorbars are standard deviation of three biological replicates. **c)** Gene track showing the mappability score across EXT9 (orange). Most splice forms cross these low mappability regions, including in the reference annotation Araport11 (shown in blue).

Sanger sequences identifies LRX1 and EXT9 splice variants

To experimentally confirm whether Extensin gene exitron splicing exists, we performed RT-PCR of LRX1 and EXT9 mRNAs. PCR products of both genes showed multiple products, characteristic of several spliced forms (data not shown). PCR products which did not correspond to the full length of the unspliced mRNA were gel extracted, cloned and sanger sequenced to identify their origin. We identified a number of mRNA fragments originating from the LRX1 and EXT9 genes. Alignment of these products using `BLAT` identified a number of spliced isoforms in both genes. To identify whether these isoforms contained canonical splice sites, we produced sequence logos. Neither gene showed a clear pattern conforming to the canonical intron motif GT/AG, though the products from LRX1 showed the reverse complement of this pattern, CT/AC.

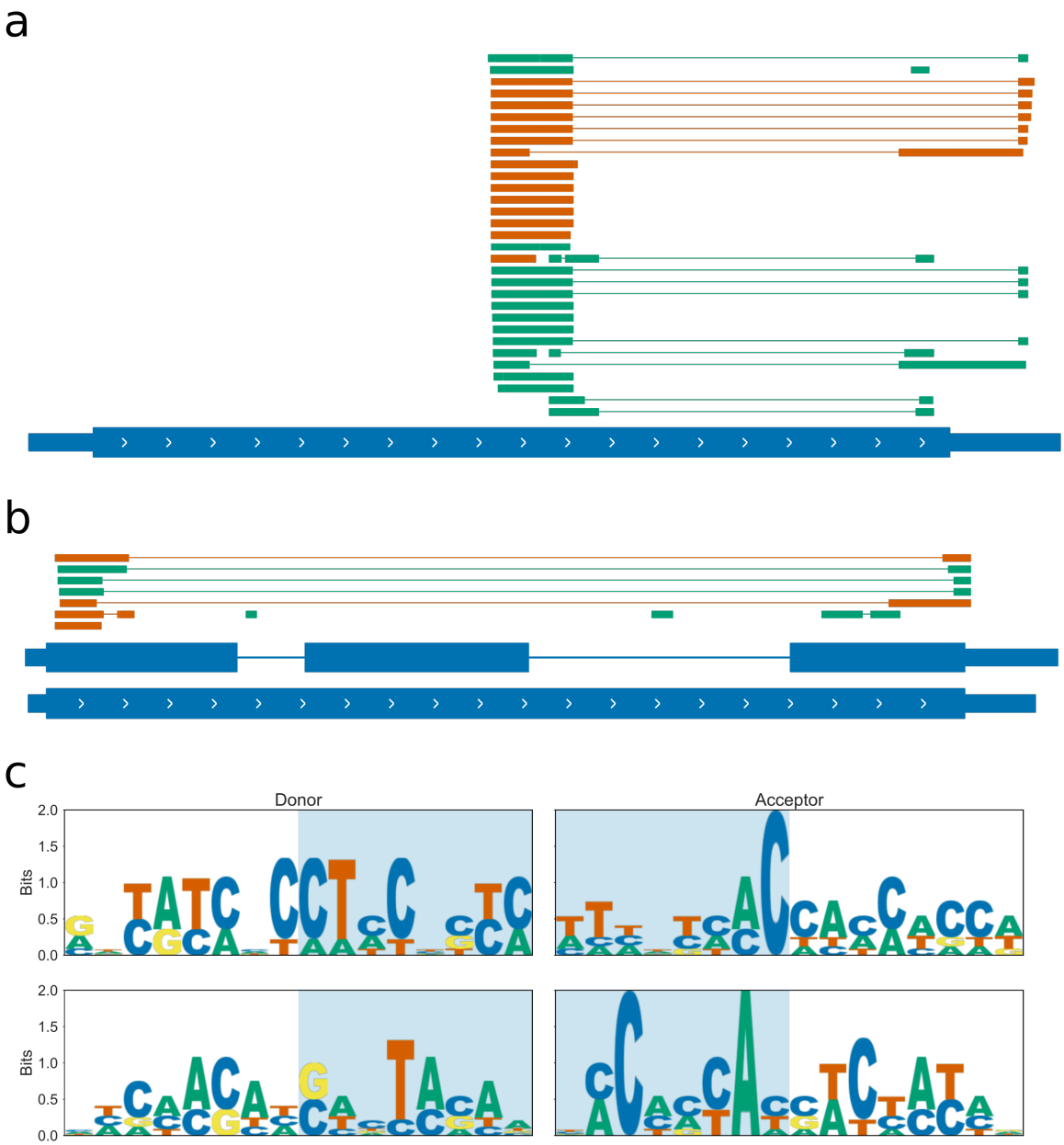

Figure 5.10: Sanger sequencing of LRX1 and EXT9 cDNA identifies spliced forms a) Gene track showing aligned sanger sequencing products for **a)** LRX1 and **b)** EXT9. Products aligned to the forward strand are shown in green, and products aligned to the negative strand are shown in orange. Gene models are from the Araport11 annotation. **c)** Sequence logos for sanger product splice junctions for LRX1 (top panel) and EXT9 (lower panel).

NMM treated plants do not have increased splicing of Extensin genes.

We hypothesise that G4s cause the exitronic splicing of Extensin genes, by slowing down Pol II elongation. To test whether increased stabilisation of G4s changes this splicing pattern, we performed RNAseq of root tissue from plants treated with NMM using 220bp paired reads, to identify novel splicing isoforms. Mapping parameters for `STAR` were made more stringent than defaults in an attempt to increase the precision of mapping over Extensin genes without attenuating recall of splice junctions too strongly. A common method for conducting differential splicing analysis is to use differential exon usage methods such as `DEXseq`. This is sometimes conducted on exon "chunks" which are the contiguous genomic ranges which each appear in a distinct set of transcripts of a gene. Since there are many overlapping exitrons in the Extensin genes, these chunks would be extremely short and be complex to interpret. Furthermore, if spliced transcripts are in lower abundance than full length transcripts, a large change in the use of a particular exitron may only lead to a small change in the expression of the exon chunk which is being spliced out. We therefore opted for counting the number of reads which support each unique junction in a gene (junction counts), and performing differential junction usage on this. The downside to this analysis is that the number of reads supporting each junction may be lower than the number per exon, leading to reduced power to detect differential usage.

In order to perform junction level differential splicing analysis, we identified all spliced reads with the correct first-in-pair strand orientation overlapping each gene. Only splice junctions with at least 20 supporting reads total across the 6 samples were kept for analysis. Splice junctions were categorised into five types based on how they related to the flattened reference annotation in Araport11: constitutive, alternative, retained intron/exitronic, exon skipping, or other. See Fig 5.11a legend for an explanation of these different categories.

We performed differential junction usage using `limma-voom` and `limma-diffSplice`. Using an FDR threshold of 0.2 and an absolute fold change threshold of 0.5, we identified 338 junctions in 302 genes with increased use during NMM treatment, and 189 junctions in 162 genes with decreased usage. 27 genes contained junctions which showed both increased and

decreased usages, i.e. some type of switching. None of the junction isoforms identified in the Extensin genes were significantly differentially utilised, however.

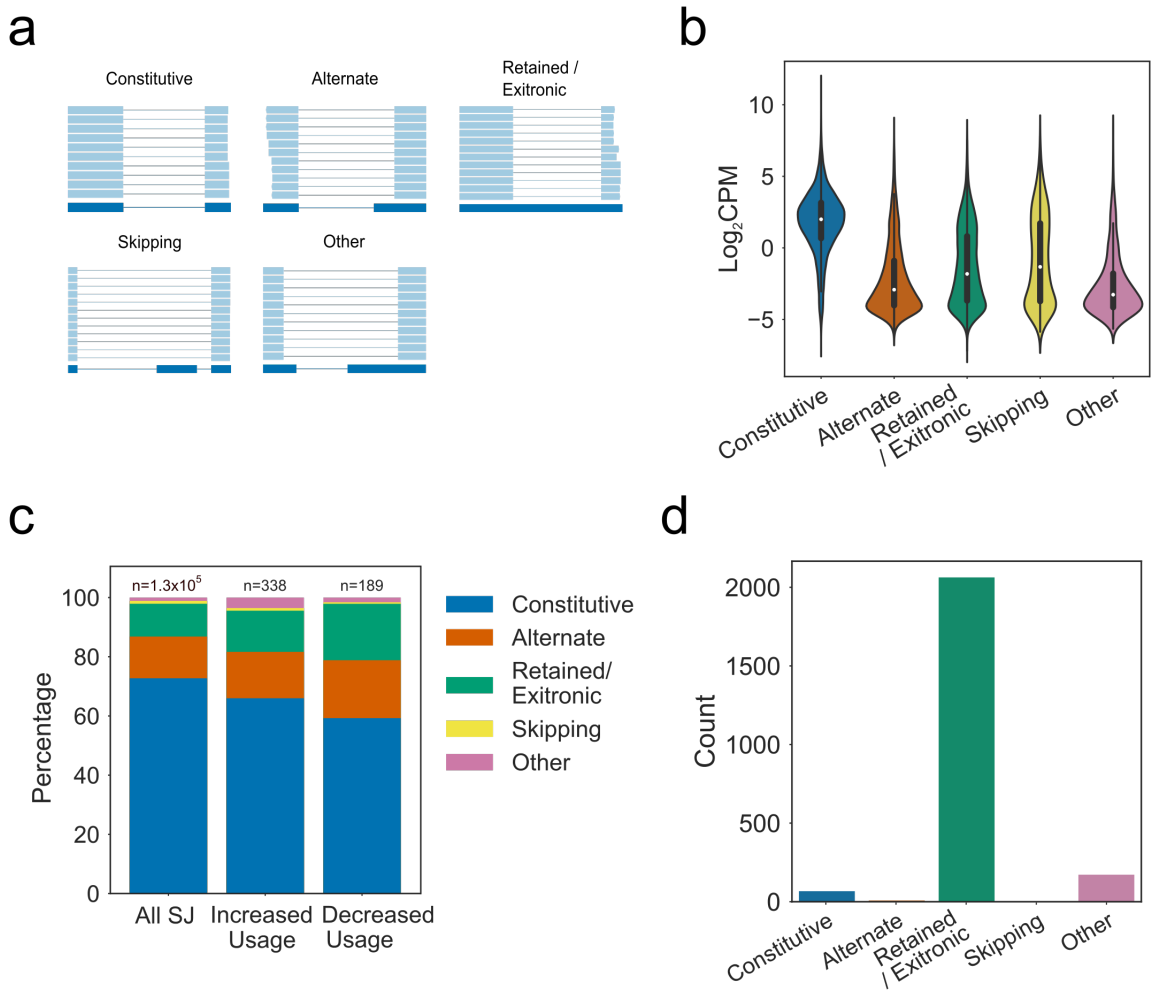

Figure 5.11: Differential Junction Usage during NMM treatment a) Schematic showing the five classes used to categorise splice junctions based on flattened reference annotation. Junctions which are present as introns in the reference are labelled constitutive junctions. Constitutive junctions which cause skipping of an exon are labelled skipping junctions. Junctions which share a donor or acceptor site with one in the reference are labelled alternate junctions. Junctions which are wholly contained within a single exon of the reference are labelled retained/exitronic junctions. Finally, junctions which share no donor or acceptor with the reference and span a mixture of exonic and intronic sequence are labelled "other". **b)** Violin plot showing distribution of average expression in log2 counts per million for each junction class. **c)** Proportion of junction in each class for all splice junctions vs. those with significantly increased or decreased usage during NMM treatment (Absolute logFC $>$ 0.5, FDR $<$ 0.2). **d)** Categorisation of detected junctions in Extensin genes.

The linear model fit by `diffSplice` is looking for differences in the usage of a junction relative to the usage of the rest of the junctions in that gene. It is possible therefore, if utilisation of all splice junctions in the Extensin genes were changed by a relatively similar amount, that `diffSplice` would not detect any differentially utilised junctions. We therefore examined the total percentage of reads mapping to each Extensin gene which were exitronic, and looked to see if this percentage changed during NMM treatment. Despite large gene-level effects on many Extensin genes with Extronic splicing (Fig 5.12a, there was no strong effect on the overall level of spliced reads mapping to these genes (Fig 5.12b), suggesting that either NMM does not affect the splicing of Extensins, only the expression, or that spliced mapping to these genes is a systematic mapping error that occurs at approximately the same rate per gene regardless of the read count.

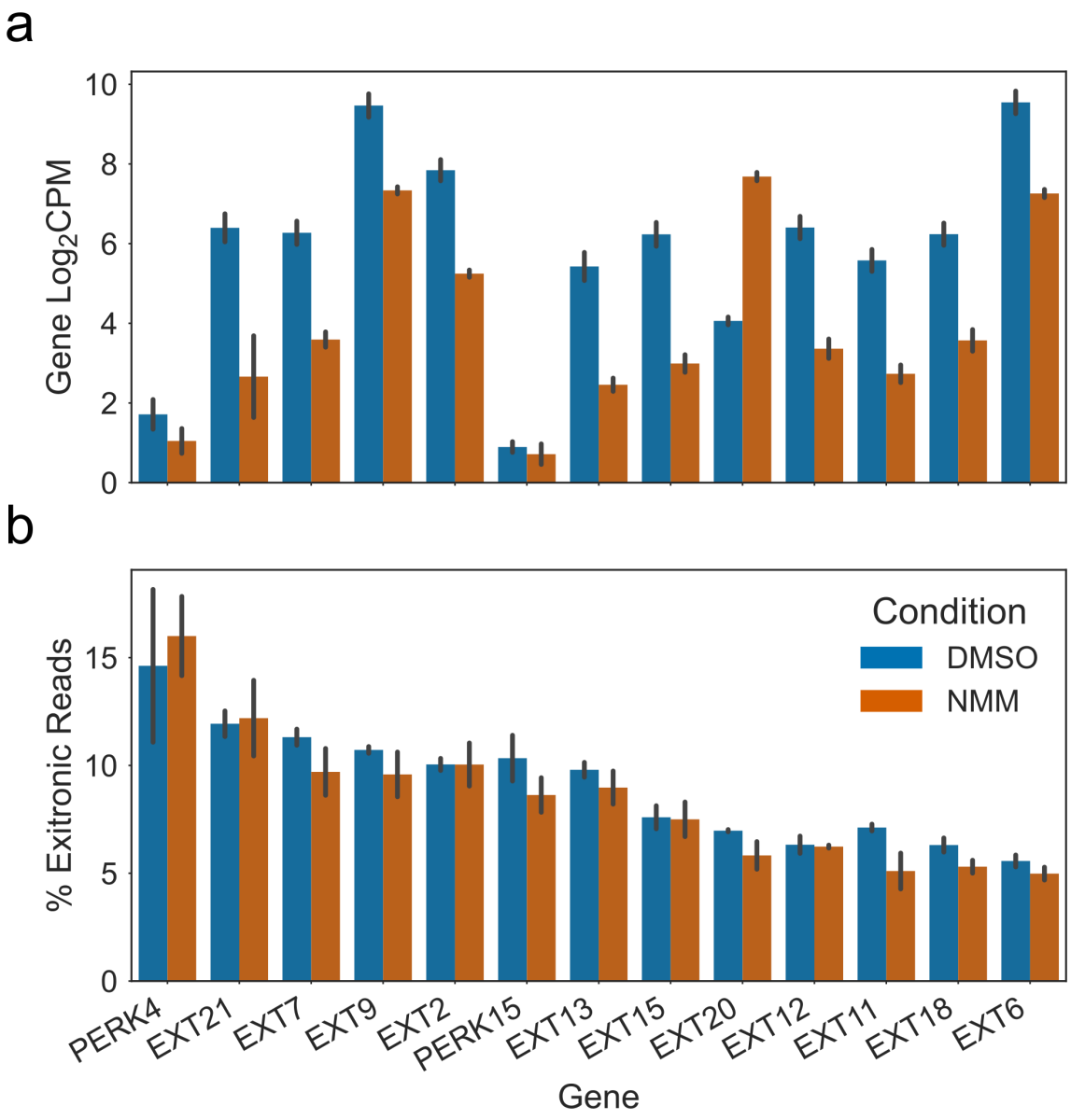

Figure 5.12: NMM does not cause significant changes in the amount of Exitronic Splicing of Extensins Barplots showing **a)** gene expression in log2 counts per million, and **b)** percentage of reads exhibiting Exitronic splicing, for expressed Extensin genes with an average of at least 5% spliced reads. Levels during mock (DMSO) treatment are shown in blue, and levels during NMM treatment are shown in orange. Ordering of genes (by average percent spliced) in upper and lower panels is matched. Errorbars are standard deviation of three biological replicates.

Discussion

We have identified a new family of G4 regulated genes: the Extensins. This is an extremely unusual and difficult to study family of genes, due to high levels of self homology and paralogy. The repetition which makes Extensins less tractable is also what makes them interesting to us: it is caused by large numbers of polyproline rich sequences, called SP4 motifs, which give rise to template stranded two tetrad PG4s. Our *in vitro* CD spectroscopy work suggests that these PG4 sequences indeed form G4s which may be stable in the physiological temperature range that Arabidopis lives at.

SP4 motifs are found throughout the coding regions of Extensin domains, usually in regularly spaced patterns. Since they contain large numbers of the motif, Extensins are some of the most PG4 dense genes in Arabidopsis, potentially in any organism. The density of PG4 sequences may also lead to large combinatorial increases in the G4 topologies that can form: an analysis of all the possible overlapping PG4 structures in Extensins suggests that some have more than 500 potential conformations per kilobase. Greater numbers of potential G4 conformations of these sequences may increase the entropy and therefore the thermodynamic stability of the folded population. This seems likely to have biological implications in transcription or regulation of the genes.

The majority of Extensin family genes exhibit a sharp reduction in gene expression when plants are treated with the G4-stabilising ligand NMM. We have demonstrated this robustly through independent analyses of microarray, RNAseq and qPCR data. Furthermore, treatment with another G4 binding ligand, Berberine, also downregulated selected Extensins, shown by qPCR. Berberine has a very different structure and binds G4s through a totally different mechanism to NMM, suggesting that this is likely to be a direct G4 effect. As we showed in **??**, NMM appears to have a strong global effect on the expression of genes with template stranded PG4s. Whilst this effect may be in part attributable to the Extensin genes, it is not solely due to them, since 50% more NMM downregulated genes than might be expected by chance contain template stranded PG4s. Our hypothesis is that template stranded PG4s cause blockages which slow the progression of Pol II along the gene during transcription. When NMM is added, this may

cause increased slowing or even premature termination of transcription, resulting in decreased expression. We demonstrated that the effect of NMM is translation independent (i.e. not caused by transcription factor activity) by treating plants with both Cyclohexamide and NMM at the same time, and showing that NMM still caused downregulation.

Examination of *de novo* assembled splice isoforms collated from many RNAseq samples in Li et al 2016 led us to speculate about splicing of Extensin genes. EXT9 was found to have the most different splice isoforms of any gene in this dataset. These splice sites tended to occur within exonic sequence and remove regions which on the template strand were PG4 rich. Analysis of how these splicing patterns might affect the protein encoded by the mRNA showed that they were multiples of three and would not cause frameshifts. Instead they would simply cause greater variation in the length of protein products. This is an intriguing idea since the Extensins are structural components of the cell wall protein matrix, and it is possible that different lengths of building block might result in differences in flexibility of the wall. Furthermore, pervious work has shown that truncated versions of LRX1 with fewer SP4 repeats still function and are able to rescue an *lrx1* null mutation (Baumberger et al 2001). It became clear, however, that the number and type of splice junctions which could be discovered in Extensin genes was highly dependent on parameters used in mapping, suggesting that at least some of these splice junctions are spurious and caused by technical error.

To try to identify true splice isoforms of Extensin genes, we performed RNAseq, using longer 220bp paired end reads to attempt to capture more splice junctions and reduce the number of multimapping reads. Since splicing occurs co-transcriptionally, and Pol II elongation speed is linked to utilisation of weak splice sites, we hypothesised that splicing of Extensins might be controlled by formation of G4s which slow down transcription. We therefore also performed RNAseq with NMM treated plants to see how G4 stabilisation affects splicing. Our RNAseq dataset identified large numbers of potential splice variants, however again many of these did not have the hallmarks of canonical splice junctions, and their abundances were highly sensitive to changes in mapping parameters, suggesting the possibility of some technical defect. Furthermore we did not see any strong or consistent change in the percentage of spliced reads identified when plants were treated with NMM, despite some very strong changes in the

expression of Extensin genes as a whole. Despite these negative results, we were able to identify PCR products from cDNA which appeared to be truncated forms of both EXT9 and LRX1. Again, these products did not have any consistent traits of splice variants. One explanation for these results, and the results of the RNAseq, could be artefacts introduced during PCR amplification. These can occur in repetitive regions due to incorrect annealing of single stranded DNA and can result in deletions or expansions. To overcome these issues in the future, the ideal techniques for detecting true splice forms would be Northern blots on a gene by gene basis, or using direct RNA Nanopore sequencing for global identification.