Chapter 1

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.

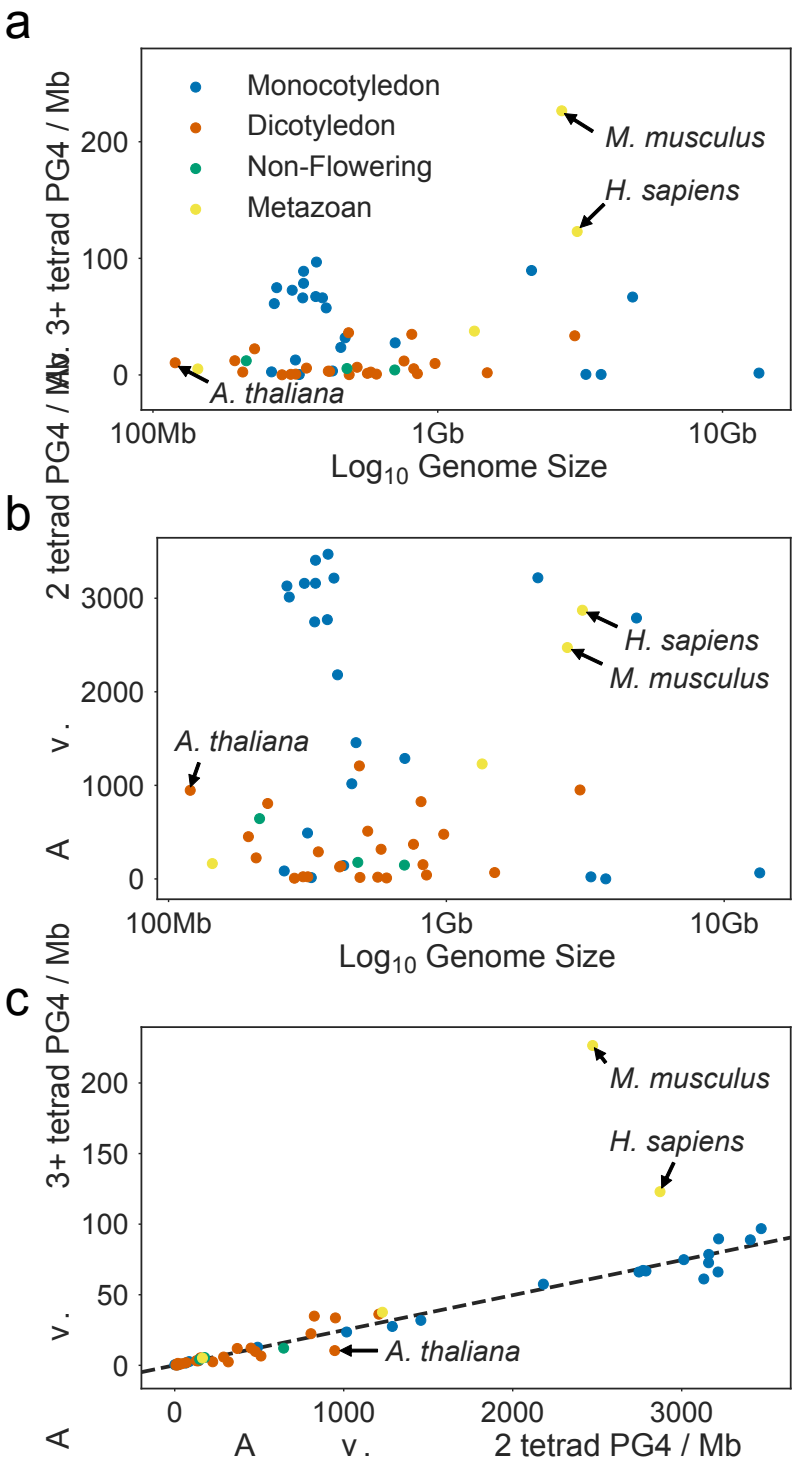

Figure 1.1: PG4 density of Plant Genomes a) and **b)** Scatter plots showing log10 genome size vs. the average **a)** three tetrad or more and **b)** two tetrad PG4 density per Megabase for 22 Monocotyledons, 21 Dicotyledons, 3 Non-flowering plants and 4 metazoans. Plant genomes are much poorer in three 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.

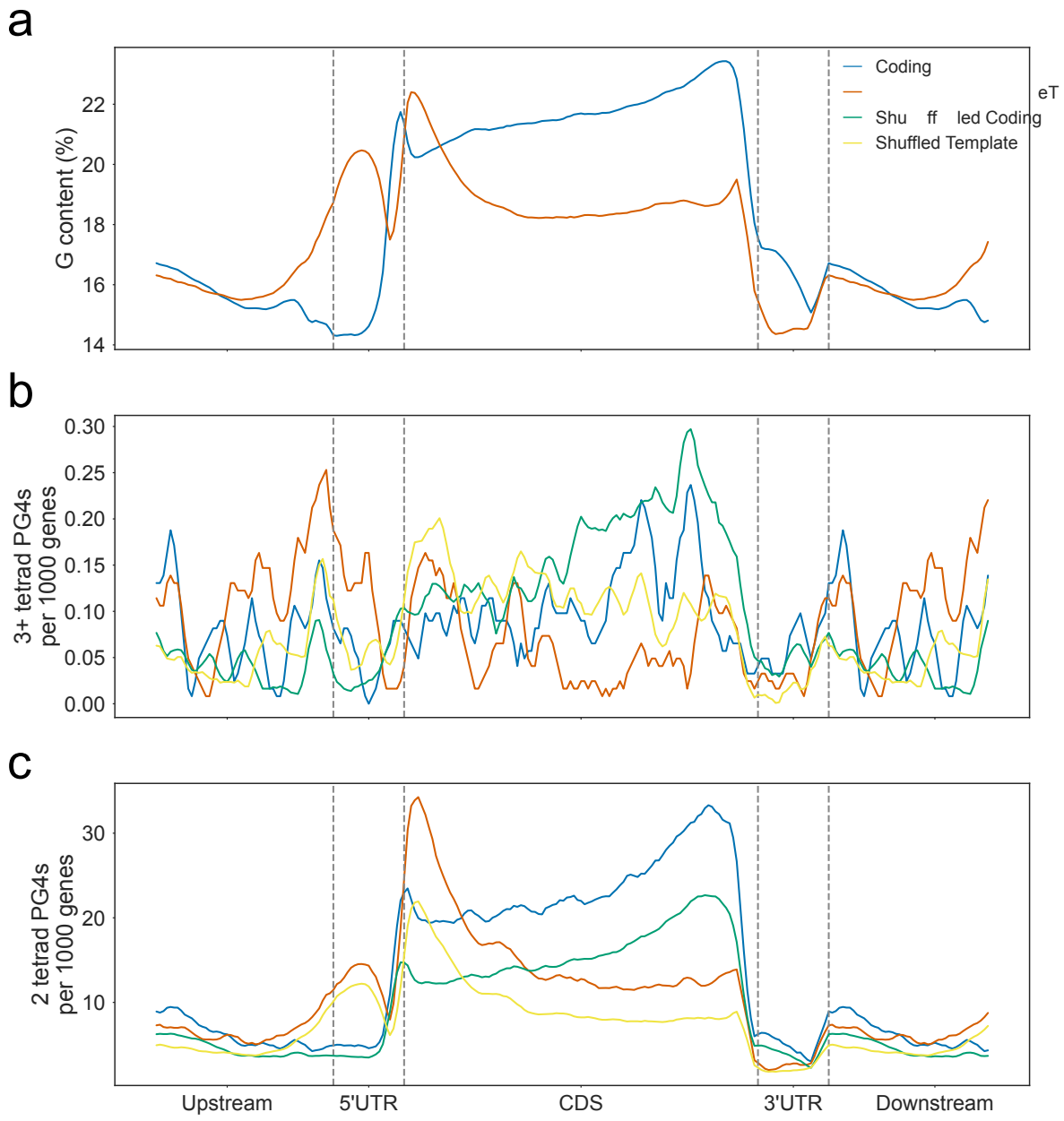

Figure 1.2: Metagene Profile of GC content and PG4 density Metagene profiles showing the **a)** G content, **b)** three or more tetrad PG4 content and **c)** two tetrad PG4 content of protein coding genes on the coding (blue) and template (orange) strands. Green and yellow lines show the average coding 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 1.3a). 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 1.3b & 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 1.3b & 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 1.3b & 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.

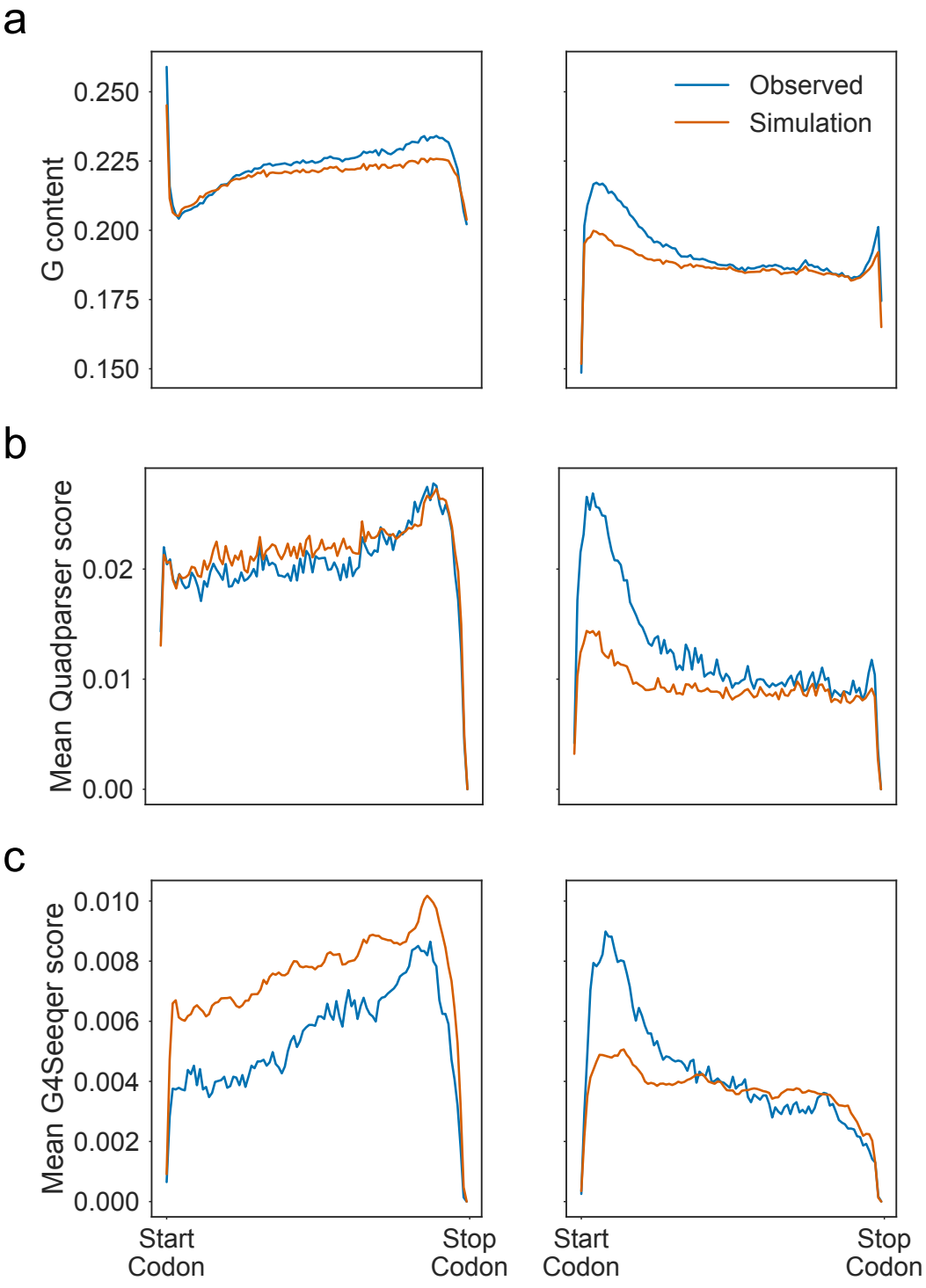

Figure 1.3: Reverse Translation Simulation shows that PG4s are enriched at the Start Codon proximal region of the template strand. Metagene profiles showing **a)** G content **b)** Quadparser PG4s and **c)** G4Seeqer PG4s for real CDS regions (blue) vs. reverse translation simulated potential coding 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 1.4a). 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.

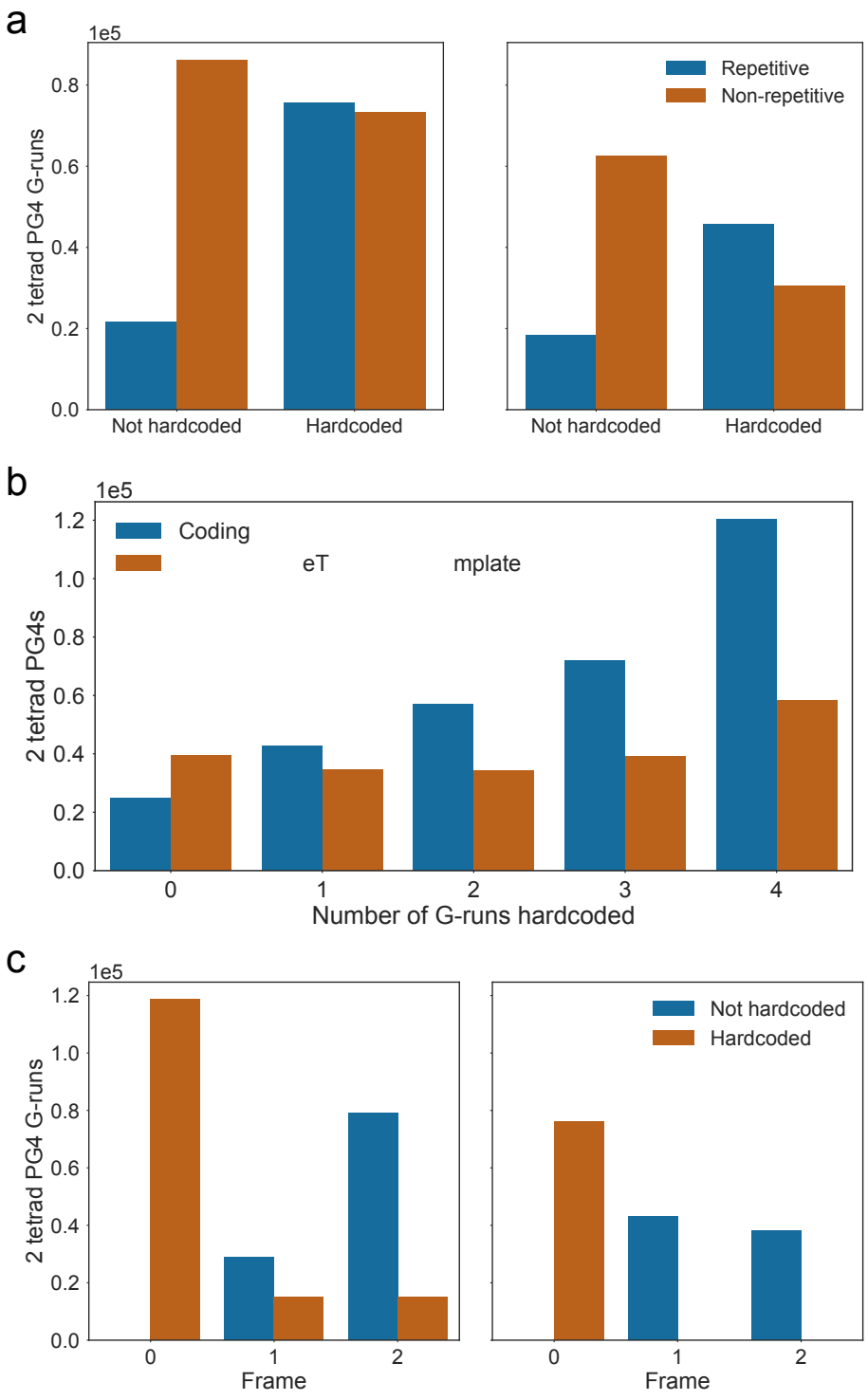

Figure 1.4: 54% of CDS PG4 G-runs are hardcoded by protein sequence. a) Frequency plot showing the total number of G-runs contributing to PG4s which are hardcoded and repetitive. Left and right panels show frequencies on coding and template strands, respectively. Hardcoded G-runs are defined as those 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 1.5). 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 <ref{hc_metagene}a). 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 1.5b). This shows that PG4s are selected by codon usage in the start codon proximal region of the template strand.

Figure 1.5: 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.

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 1.6). The majority of these G-runs are hardcoded, because the codon for glycine is GGN (Fig 1.6a, 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 1.6b). 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 1.6a, 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.

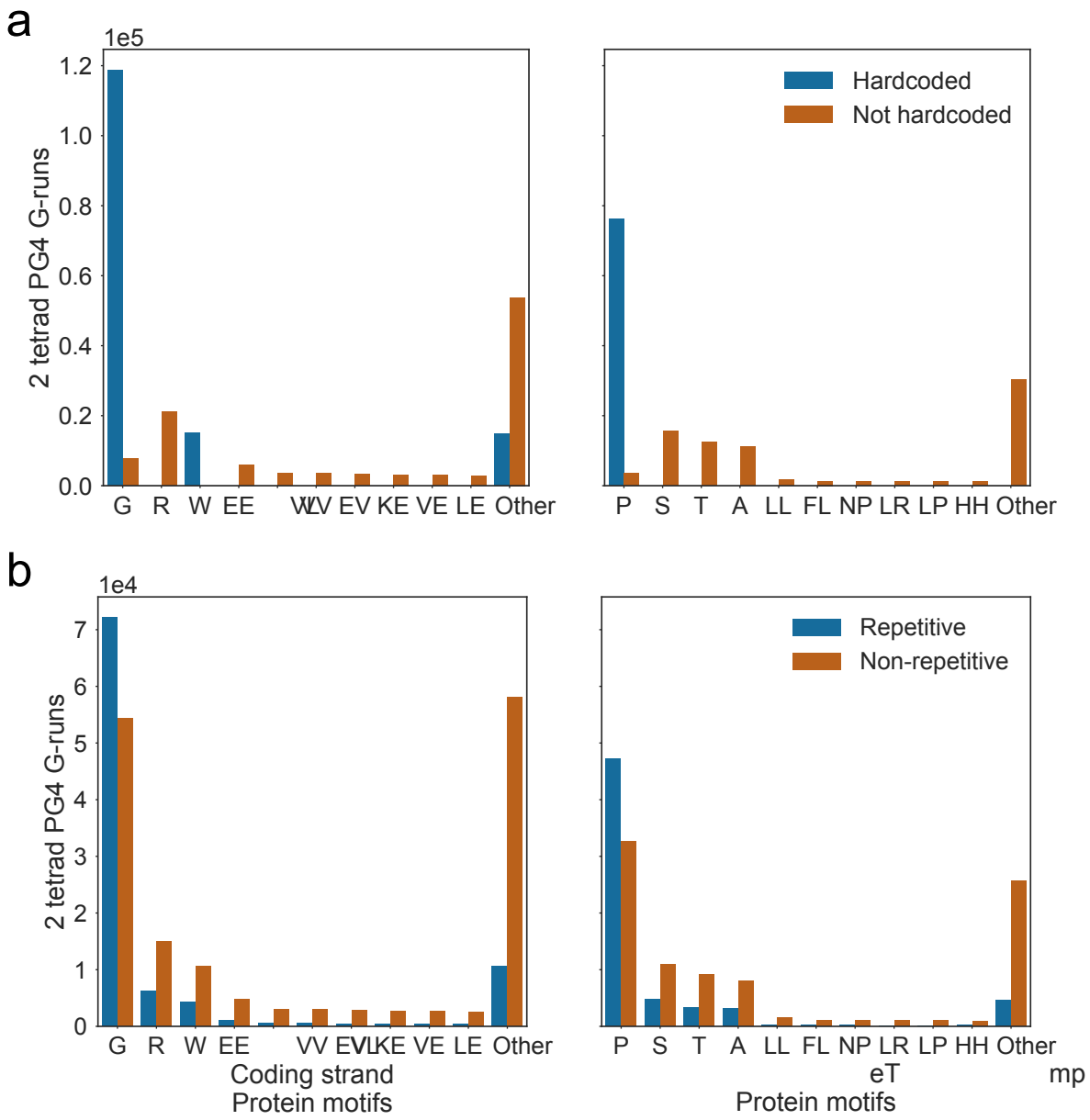

Figure 1.6: Protein motifs that are coded by PG4 G-runs. Frequency plots showing the 10 most common amino acids motifs which PG4 G-runs contribute to the coding of. Left and right panels are for coding and template strands, respectively. Bars are coloured by the frequency of **a)** hardcoded and **b)** 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.