Chapter 1

Chapter 5: Effect of G-Quadruplexes on expression and splicing of the Extensin gene family

Introduction

Methods

Plant growth conditions and drug treatments

- Growth conditions & drug treatments for qPCR
- Growth conditions & drug treatments for RNAseq
- Growth conditions for sanger sequencing

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).

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`.

Quantitative PCR experiments

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, 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

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

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 1.1). 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 (Fig1.1, 50% of expressed genes in `GO:0005199` were downregulated by NMM, FDR $=$ 9.6e-7).

			Coding Strand PG4 Log2 Enrichment	Coding Strand FDR	Strand PG4 Log2 Enrichment	eT 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 1.1: Gene Ontology groups enriched in template stranded putative G-Quadruplexes 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 1.2a). 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 1.2b).

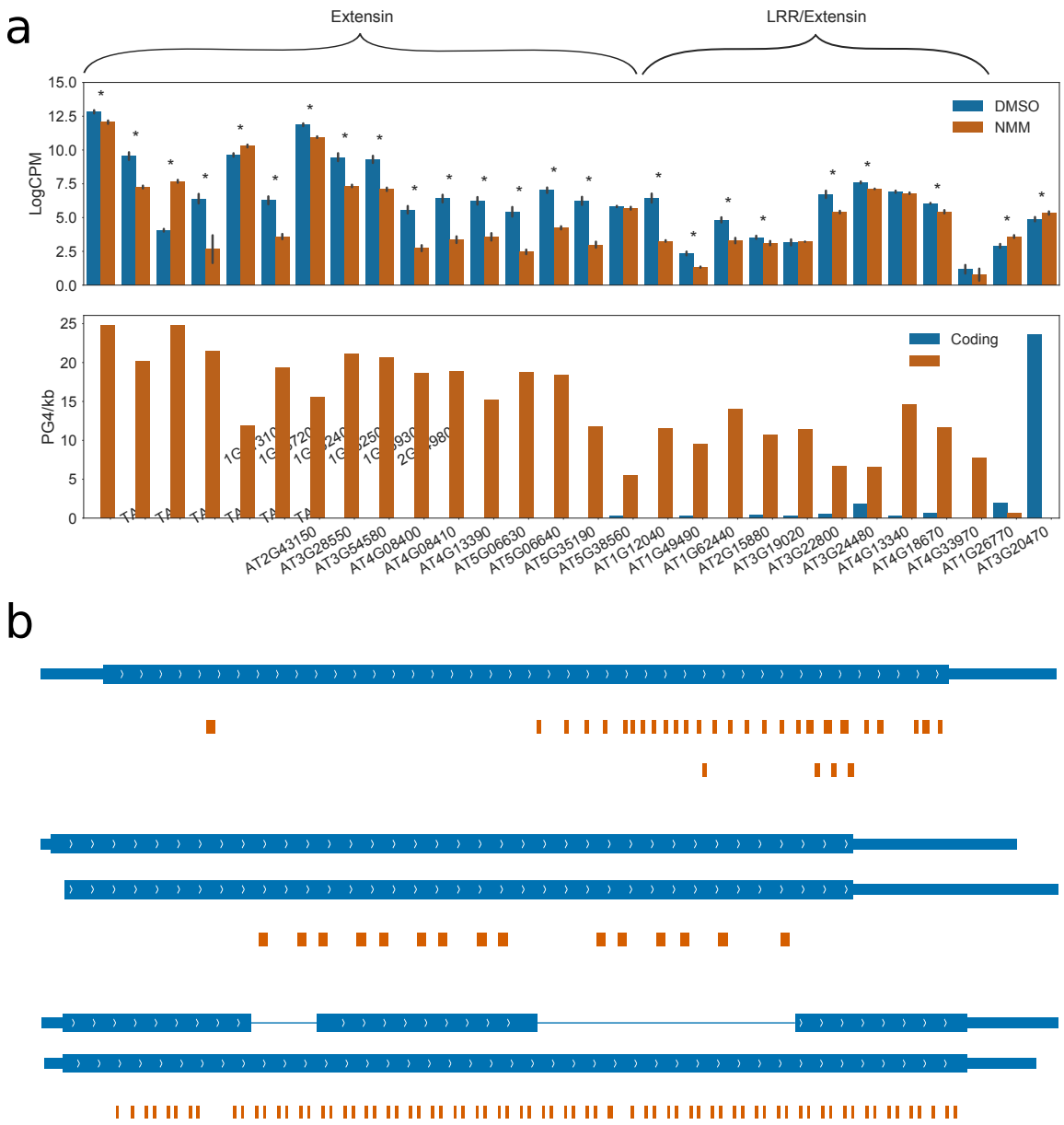


Figure 1.2: Expression and PG4 density of genes in the Cell Wall Structural Ontology group `GO:0005199` **a)** Panels showing gene expression (top panel) and PG4 density (bottom panel) for genes in the `GO:0005199` group. Expression in DMSO (blue) and NMM (orange) conditions is shown at log2 counts per million (from root RNAseq dataset). Errorbars show standard deviation. 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.

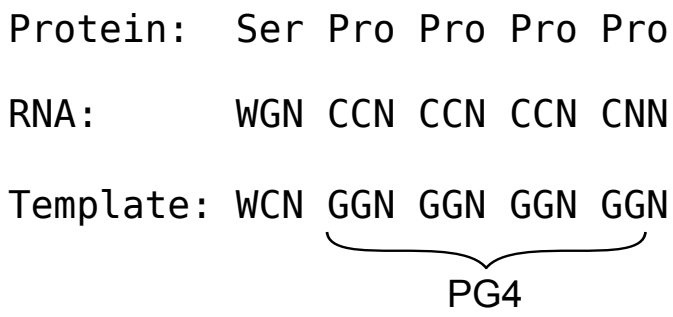
From a search of the literature, we discovered that Extensin genes are highly repetitive, proline rich proteins. These proteins polymerise to function as a structural matrix in the protein component of the plant cell wall. In particular, we noted that these proteins are characterised by large numbers 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 1.4a). 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 1.3). 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
	TAEXT20	SP5/SP4	Roots	2	1	40	186	84	1773	3.72	6.80	1.1e-168
	TAEXT21	SP5/SP4	Roots	7	0	28	139	69	1627	-3.40	5.57	5.1e-10
	TAEXT22	SP5/SP4		3	1	13	269	42	1353			
	TAEXT18	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	PERK	Chimeric	11 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 1.3: 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 1.4b). Removing the PG4 by mutating the sequence (mutated sequence: AGAGGTGATGGTGGTATG) or removing potassium ions from the buffer abolished this absorbance profile.

a



b

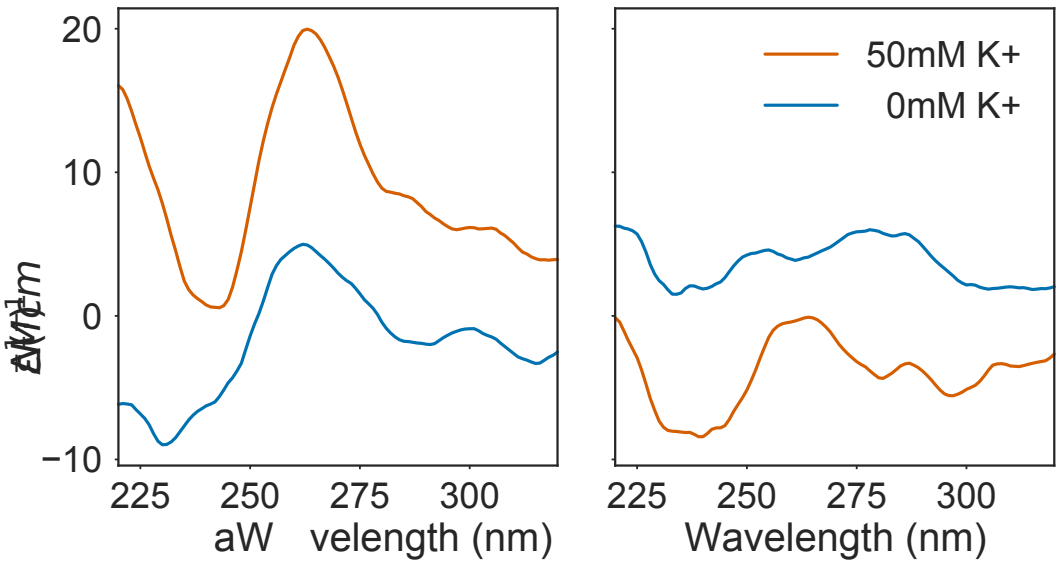


Figure 1.4: The Extensin SP4 motif forms a G-Quadruplex *in vitro*. a) Schematic showing how the Extensin SP4 protein motif hardcodes a two tetrad PG4 into the template strand of the gene body. **b)** CD spectroscopy of an Extensin repeat sequence (left) and a mutated control which does not conform the 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 1.5a). Treatment with the G4 intercalating drug, Berberine, also caused strong downregulation of EXT13 and LRX1 (Fig 1.5b). 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.

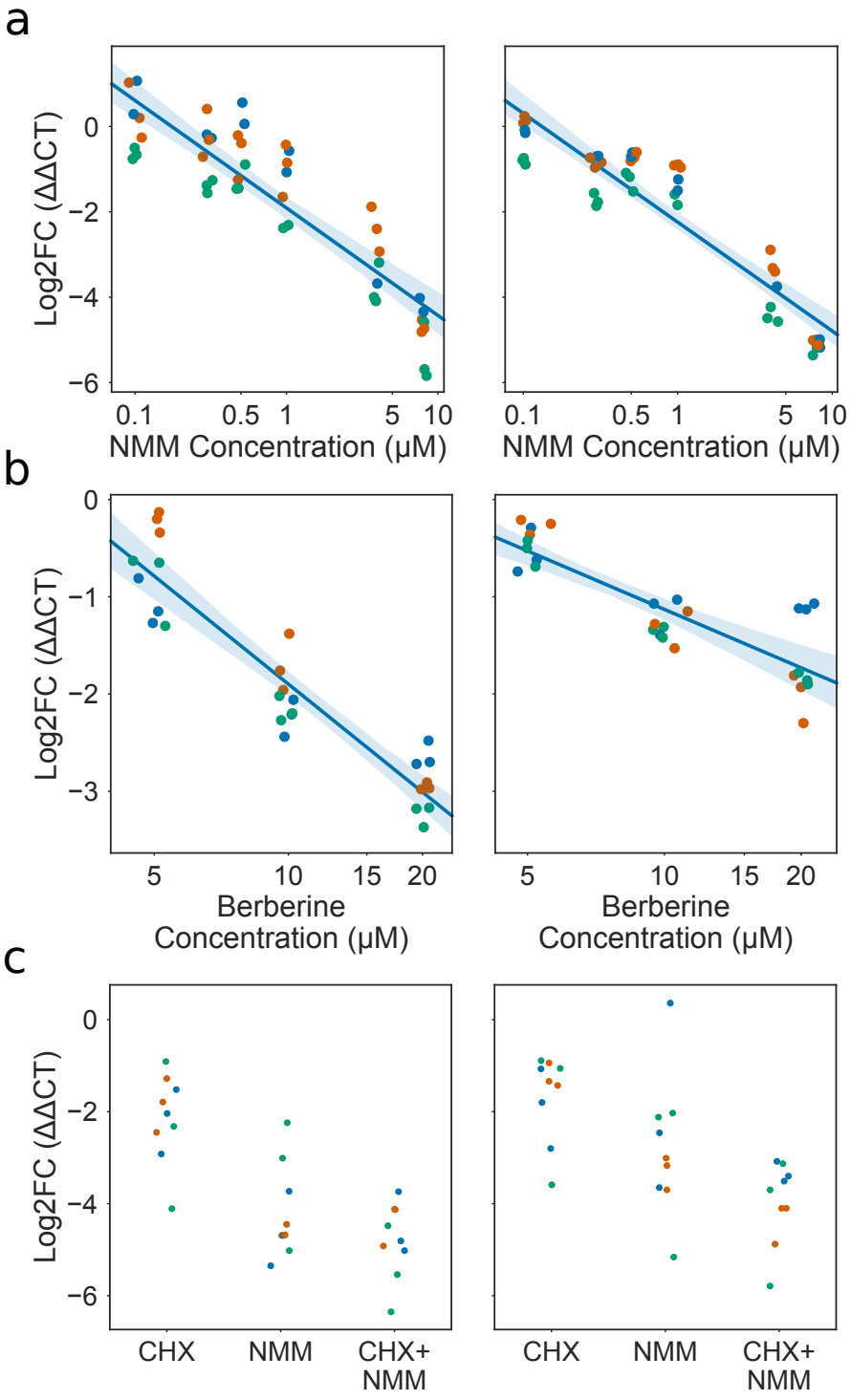


Figure 1.5: Expression of EXT13 and LRX1 during treatment with G4-binding ligands Scatter/strip plots showing qPCR results for EXT13 (left panels) and LRX1 (right panels). Log2 fold change in expression (ΔΔCT) of Extensin genes decreases with increasing concentrations of **a)** NMM and **b)** Berberine. **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 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 1.6a). 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 1.6b). 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 1.6c). These logos showed only a weak enrichment for the GT/AG motif in EXT9, and CT/AC in LRX1.

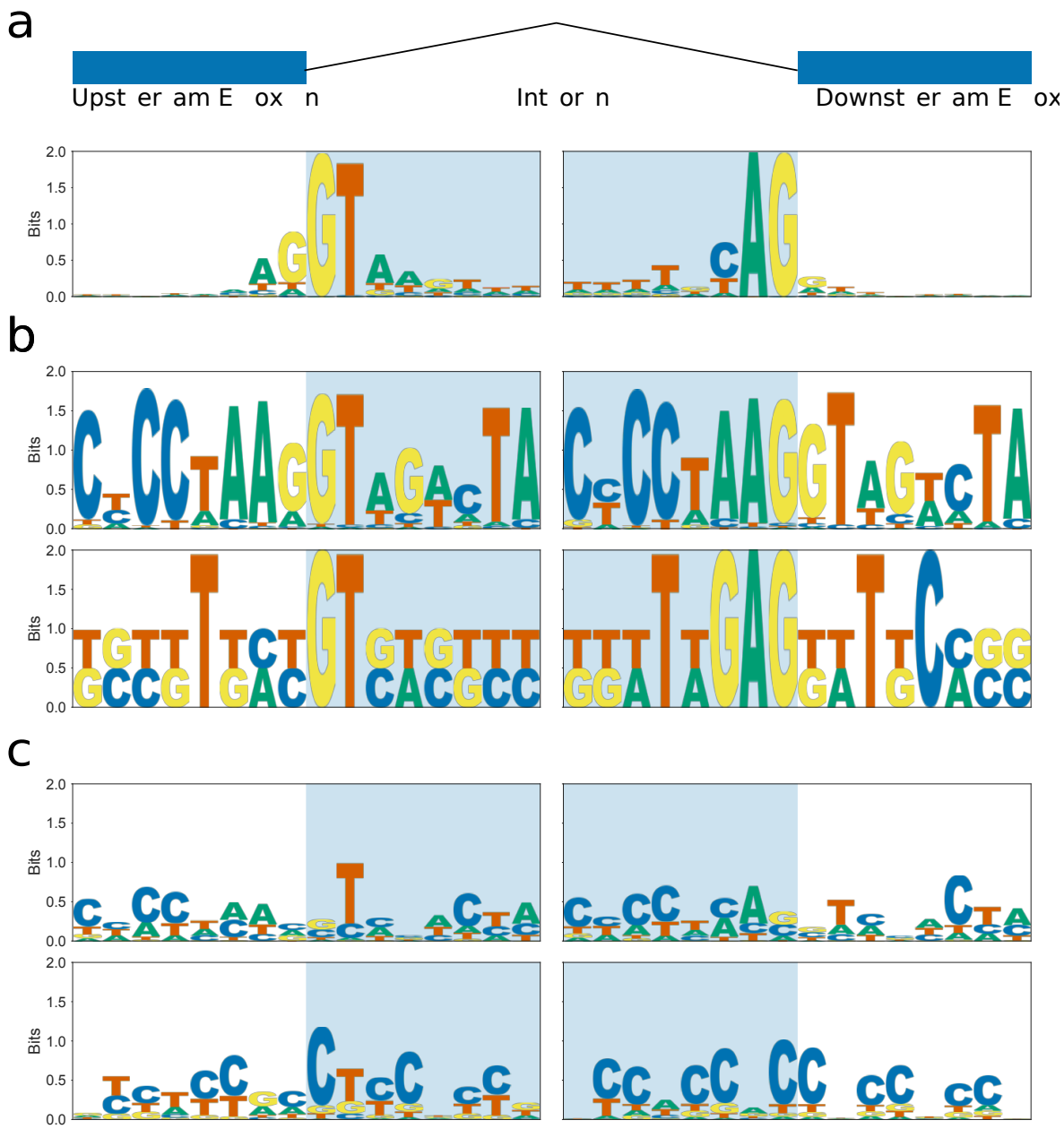


Figure 1.6: Splice junction motifs for EXT9 and LRX3 Sequence logo plots showing consensus splice site sequences around donor (left panels) and acceptor (right panels) splice sites. Putative intronic sequences are shown on shaded blue background. **a)** Splice junction consensus sequence logo for Arabidopsis, 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 1.7a-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 1.8a). 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 1.7c).

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 1.7d).

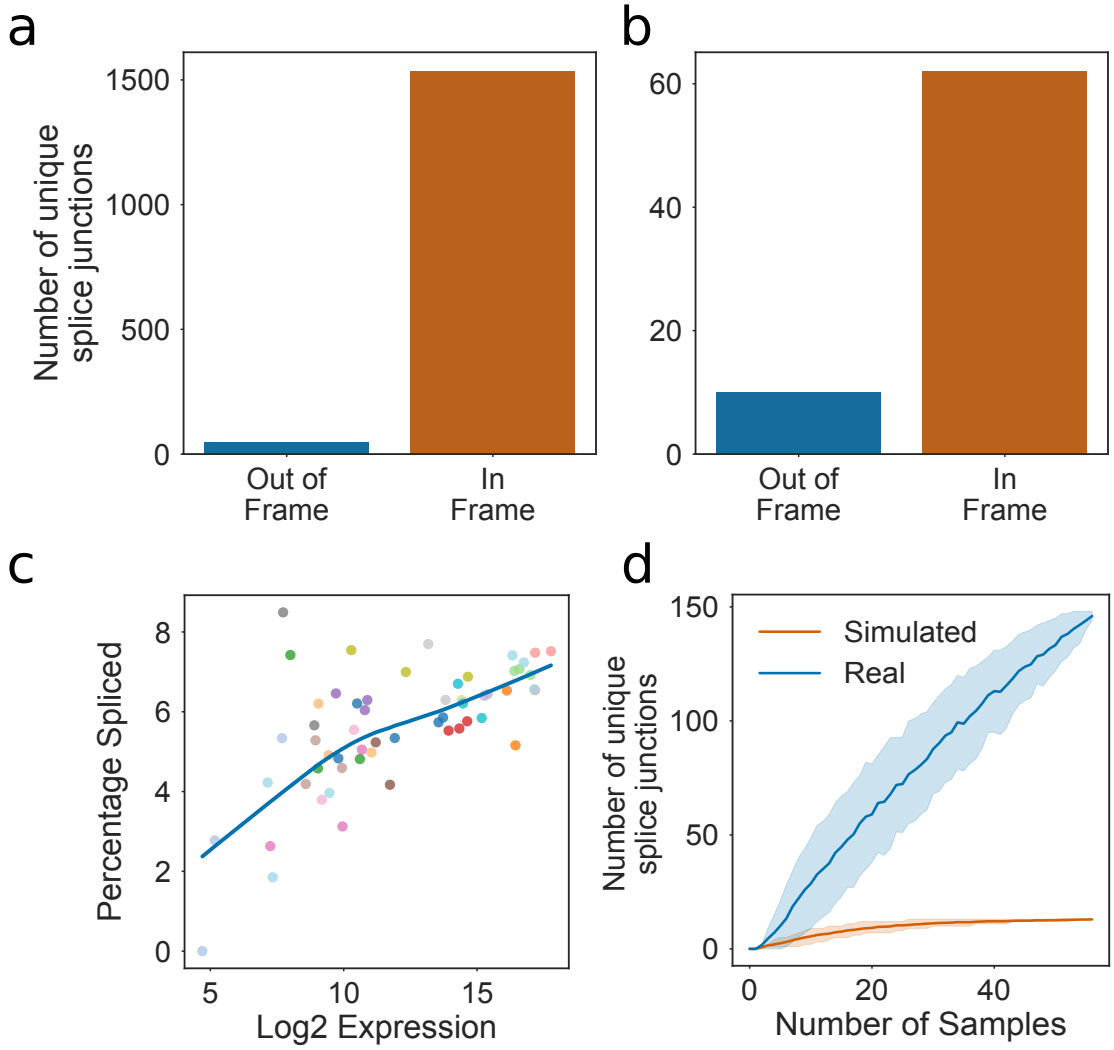


Figure 1.7: Splice junction motifs for EXT9 and LRX3 a & **b)** Frequency barplots showing number of In/Out of frame splice junctions for EXT9 and LRX3 respectively. **c)** Scatterplot showing percentage of reads with splicing versus log2 counts per million for EXT9. **d)** Bootstrapped splicing simulation 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 1.8b). For EXT9, the regions with lowest mappability are clearly those with the most annotated splice sites, including in the Araport11 reference (Fig 1.8c).

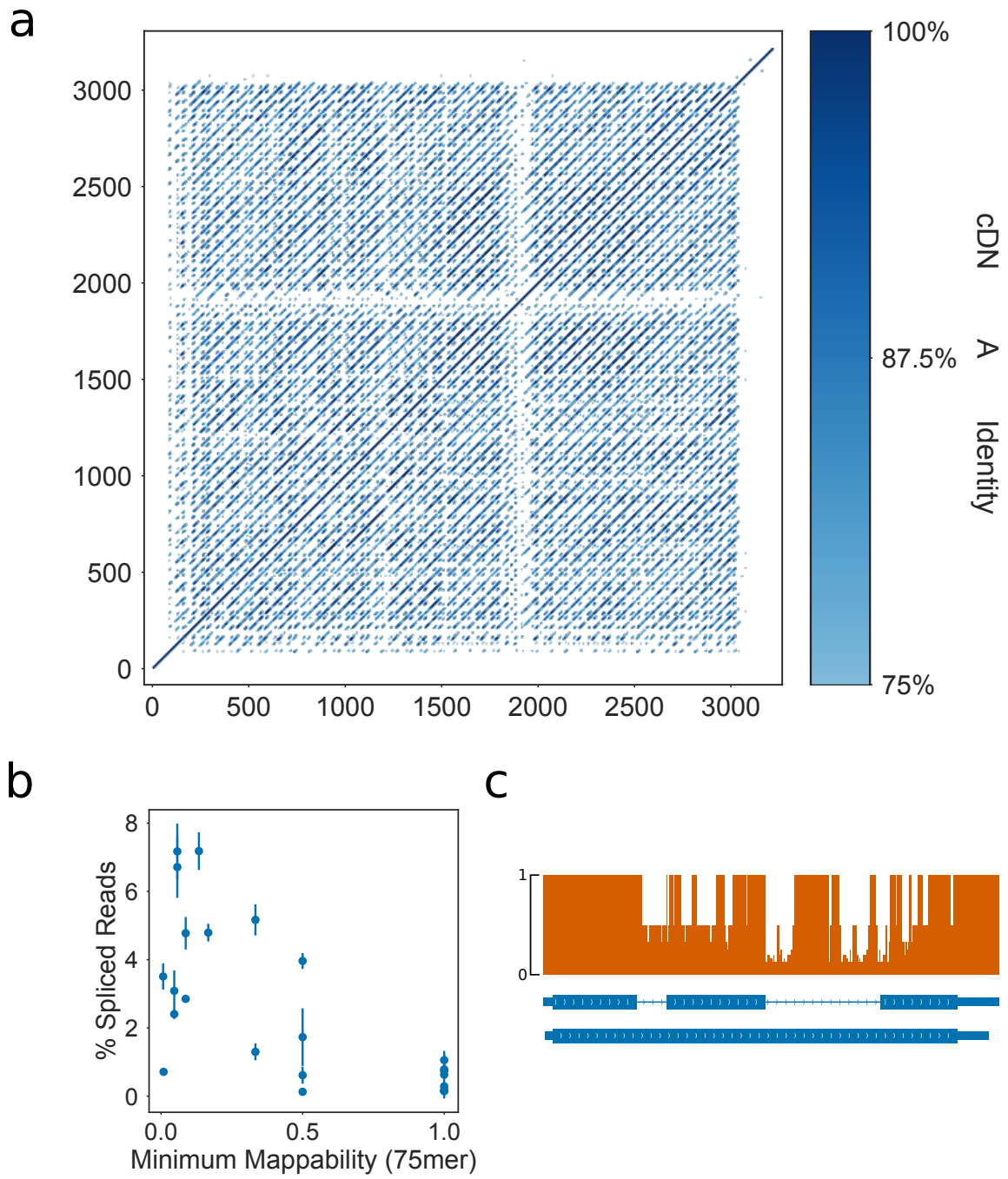


Figure 1.8: Extensin genes with greater spliced mapped reads have low mappability **a)** Dotplot showing self homology of the EXT9 cDNA (unspliced isoform). Positional identity was calculated using 15bp windows across the gene. Positions with identity less than 75% were filtered to remove noise from 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. **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.

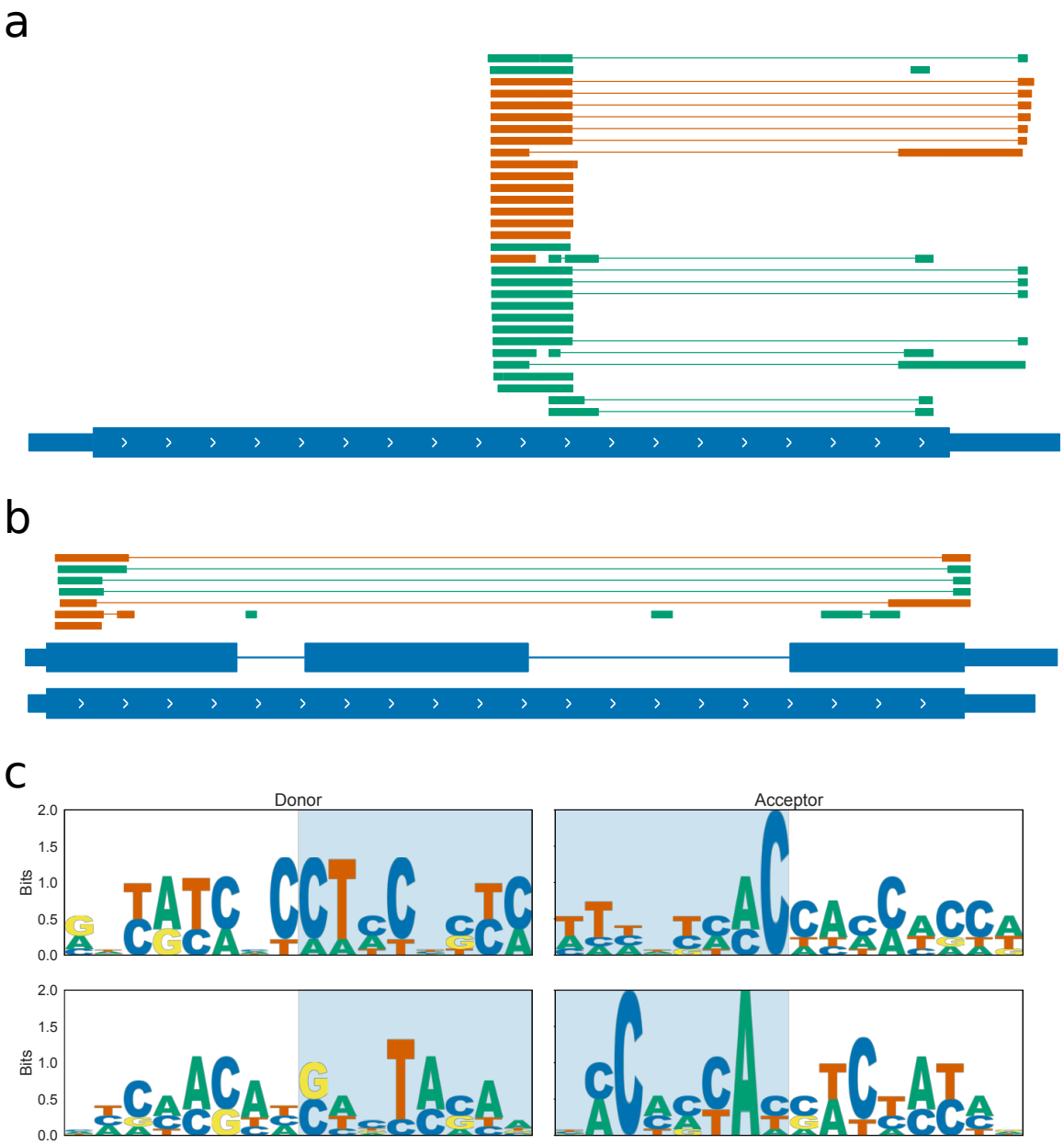


Figure 1.9: Sanger sequencing of LRX1 and EXT9 cDNA identifies spliced forms a) Gene track showing aligned sanger sequencing products for **a)** LRX1 and **b)** EXT9. Products aligned to the forward strand are shown in green, and products aligned to the negative strand are shown in orange. Gene models 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.

Discussion