
Understanding and engineering *cis*-regulatory functions in *Arabidopsis thaliana*

Samuel Witham

A thesis presented for the degree of Doctor of Philosophy



October, 2023

© This copy of the thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with the author and that use of any information derived therefrom must be in accordance with current UK Copyright Law. In addition, any quotation or extract must include full attribution.

Abstract

Nitrate (N) is essential for plant growth and metabolic processes. The use of N-containing fertilisers improved crop yields over several decades but had negative environmental impacts. In response to N-availability, plants alter the expression of thousands of genes, analysis of which enabled the identification of gene regulatory networks (GRNs) containing putative interactions between transcription factors (TFs) and their targets. GRNs contain network motifs that likely affect network dynamics and are important for plant N-responses. Rational engineering of GRNs has the potential to improve plant traits but requires a detailed understanding of network components and their interactions. In this thesis, I characterised interactions within an N-response subnetwork using *in vitro* protein-DNA relative affinity and *in planta* gene expression assays (Chapter 3). I investigated the perturbation of this subnetwork using synthetic biology approaches using a CRISPR library to disrupt edges in the promoters/5' UTRs of four TFs (Chapter 4). Five mutant lines were identified which had altered root growth and gene expression in response to N. A second method of perturbation is the introduction of synthetic genetic feedback controllers built from synthetic elements which are predicted to change network dynamics. To inform the design of synthetic promoters that respond to specific TFs in the N-subnetwork I investigated promoter architectural differences between four groups of genes with different expression patterns (Chapter 5). I successfully built and tested synthetic N-responsive promoters that responded to nitrate and specific TFs in the subnetwork, and synthetic TFs which activated or repressed *ARF18* expression. These synthetic elements were co-assembled into synthetic feedback controllers and were transformed into plants (Chapter 6). Together, the work in this thesis demonstrates that the engineering of root growth in response to nitrate can be facilitated by an improved understanding of an N-response subnetwork, providing a path towards the engineering of improved nitrogen-use efficiency.

Acknowledgements

I am deeply grateful to the numerous individuals who have supported me throughout my PhD journey and contributed to the completion of this thesis.

First, I would like to express my gratitude to all the members of the Patron lab and Earlham Foundry, present and past, for their invaluable assistance in the lab and constructive feedback during lab meetings. I would like to thank Tufan, Yaomin, Kalyani, Quentin, Connor, Melissa, Oleg, Dasha, Hong, Rachel, Lesley, Jose, and Adam for their invaluable contributions.

I am also grateful to those who generously provided proofreading and feedback on my thesis, including Nicola Patron, Wilfried Haerty, Will Nash, Anita Scoones and my parents. Their insights and suggestions were invaluable in improving the final thesis.

I am particularly grateful to my primary supervisor, Nicola Patron, for her tireless scientific support and guidance throughout my PhD, and to my secondary supervisor, Wilfried Haerty, for his invaluable support with my bioinformatics. I am also grateful to Will Nash for his help with my bioinformatics.

On a personal level, I am thankful to my friends at Earlham for their moral support and encouragement during this journey, including Matthew Madgwick, Peter Osborne, and Anita Scoones. I am also grateful to the members of the ESB for organizing student socials and events. I would like to express my deepest gratitude to Rebecca for her unwavering support, patience, and encouragement. Finally, I would like to thank my parents for their support throughout my entire academic career.

I am also grateful to the Norwich Research Park Doctoral Training Partnership (NRPDTP) and the Biological Sciences Research Council (BBSRC) for providing me with the opportunity to complete my PhD. I would like to extend my thanks to horticultural services for keeping my plants alive, and to Genomics Pipelines for their support. I am also grateful to everyone at the Earlham Institute for making me feel welcome and for the enjoyable conversations at lunch and in the kitchen.

Access Condition and Agreement

Each deposit in UEA Digital Repository is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the Data Collections is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form. You must obtain permission from the copyright holder, usually the author, for any other use. Exceptions only apply where a deposit may be explicitly provided under a stated licence, such as a Creative Commons licence or Open Government licence.

Electronic or print copies may not be offered, whether for sale or otherwise to anyone, unless explicitly stated under a Creative Commons or Open Government license. Unauthorised reproduction, editing or reformatting for resale purposes is explicitly prohibited (except where approved by the copyright holder themselves) and UEA reserves the right to take immediate 'take down' action on behalf of the copyright and/or rights holder if this Access condition of the UEA Digital Repository is breached. Any material in this database has been supplied on the understanding that it is copyright material and that no quotation from the material may be published without proper acknowledgement.

Contents

Abstract	2
Acknowledgements	3
List of abbreviations	11
List of Figures	23
List of Tables	30
List of accompanying materials	30
1 Introduction	31
1.1 Gene regulatory networks	31
1.2 Plant responses to nitrate	33
1.2.1 The low-nitrate response	34
1.2.2 The replete-nitrate response	34
1.2.3 The nitrate-starvation response	37
1.2.4 The role of hormones in the nitrate response	37
1.3 Regulation of gene expression	42
1.3.1 <i>Cis</i> -regulatory DNA sequences	43
1.3.2 <i>Trans</i> -acting DNA regulation	45
1.3.3 Identification of CREs	48
1.4 Identification of chromatin accessible regions	48
1.5 Identification of specific DNA-protein interactions	50
1.5.1 Yeast one-hybrid	50
1.5.2 <i>In planta</i> expression methods	51

1.5.3	DNase I	52
1.5.4	Electrophoretic mobility shift assay	53
1.5.5	Systematic evolution of ligands by exponential enrichment	53
1.5.6	Chromatin immunoprecipitation	53
1.5.7	DNA adenine methyltransferase identification (DamID)	54
1.5.8	Protein binding microarray	54
1.5.9	DNA affinity purification sequencing	55
1.5.10	<i>In silico</i> prediction of TFBSs	56
1.6	Engineering regulation	56
1.6.1	CRISPR-Cas9 gene editing	58
1.6.2	Development of synthetic promoters for microbes	59
1.6.3	Synthetic regulatory elements for plants	60
1.6.4	Synthetic TFs	61
1.6.5	Towards engineering regulation in plants	63
1.7	Gaps in knowledge and experimental approaches	63
2	Methods	67
2.1	Assembly and validation of plasmid constructs	67
2.1.1	Design and construction of promoter parts	69
2.1.2	Luciferase reporter constructs	70
2.1.3	Transcription factor parts	70
2.1.4	Transcription factor bacterial expression constructs	72
2.1.5	Transcription factor plant expression constructs	73
2.1.6	Digestion-ligation reaction	73
2.1.7	Transformation of competent <i>E. coli</i>	74
2.1.8	Colony PCR	74
2.1.9	Preparation and verification of plasmid DNA	74
2.1.10	Agarose gel	75
2.2	Plant growth and transformation	75
2.2.1	Seed sterilisation	75
2.2.2	Arabidopsis growth	76
2.2.3	Transformation of electrocompetent <i>A. tumefaciens</i>	76

2.2.4	Arabidopsis transformation	77
2.2.5	Selection of transgenic seeds	77
2.2.6	Protoplast preparation	78
2.2.7	Protoplast transfection	78
2.2.8	Assessment of root system architecture	79
2.3	Protein expression, purification and analysis	80
2.3.1	Recombinant protein expression in <i>E. coli</i>	80
2.3.2	Protein purification	80
2.3.3	Protein quantification	81
2.3.4	SDS-PAGE	82
2.3.5	Western blot	82
2.4	Transcription factor relative binding affinity	83
2.5	Expression analysis	84
2.5.1	Ratiometric quantification of luciferase expression	84
2.5.2	Quantification of transgene luciferase expression	86
2.5.3	Root RNA extraction	87
2.5.4	cDNA preparation	87
2.5.5	Quantitative real-time PCR	88
2.6	CRISPR library	89
2.6.1	Design and assembly of constructs for a targeted CRISPR library	89
2.6.2	Growth and selection of a population of edited plants	90
2.6.3	Polymerase chain reaction for amplicon genotyping	90
2.6.4	Generation of amplicon sequencing library	90
2.6.5	Demultiplexing of samples	91
2.6.6	Variant calling	93
2.6.7	Mutation analysis	93
2.7	<i>Cis</i> -regulatory module analysis	95
2.7.1	Extraction of <i>cis</i> -regulatory modules	96
2.7.2	Transcription factor binding site identification	96
2.7.3	Gene selection	96
2.7.4	Sliding window creation	97
2.7.5	GC content	97

2.7.6	Transcription factor binding site coverage	98
2.7.7	Open chromatin coverage	98
2.7.8	TF diversity	98
2.7.9	TATA box enrichment	99
2.7.10	GO-term analysis	99
2.7.11	Visualisation of TFBSs in promoters	100
3	Validation of a regulatory subnetwork for nitrate-associated metabolism and growth	101
3.1	Preface	101
3.2	Introduction	101
3.3	Aims	105
3.4	Experimental approach	105
3.5	Results	107
3.5.1	Identification of candidate binding sites	107
3.5.2	Expression and validation of <i>AtARF18</i> in TRAM assays	112
3.5.3	Validation of TF:DNA interactions within the subnetwork	113
3.5.4	Assembly of plasmids for luciferase assay	114
3.5.5	Optimisation of reporter co-expression assays	115
3.5.6	Testing and selection of calibrators	116
3.5.7	Regulatory effects of TF expression on $p_{ANAC032}$ -LucN	118
3.5.8	Regulatory effects of TF expression on p_{ANR1} -LucN	121
3.5.9	Regulatory effects of TF expression on p_{ARF18} -LucN	123
3.5.10	Regulatory effects of TF expression on p_{DREB26} -LucN	125
3.5.11	Regulatory effects of TF expression on p_{NIR1} -LucN	127
3.5.12	Regulatory effects of TF expression on p_{NLP6} -LucN	129
3.5.13	Regulatory effects of TF expression on p_{NLP7} -LucN	131
3.5.14	Regulatory interactions between TGA1 and target genes	133
3.5.15	Investigating the role of ARF binding sites in the <i>ANAC032</i> up- stream promoter and untranslated region.	136
3.5.16	Investigating the role of TGA1 binding sites in the <i>ANAC032</i> promoter	138

3.5.17	Investigating the role of NLP binding sites in the <i>NIR1</i> upstream promoter and untranslated region.	139
3.5.18	Investigating the role of ANAC032 binding sites in the <i>NLP7</i> promoter	141
3.5.19	Reporter co-expression assays provide insights into indirect regulation within the subnetwork	142
3.6	Discussion	146
3.6.1	Reporter co-expression assays support direct edges within the subnetwork	146
3.6.2	TF-DNA interaction data supports the architecture of the subnetwork	149
3.6.3	Reporter co-expression assays are a useful tool for understanding TF function	150
3.6.4	Conclusions	154
4	Development of a CRISPR library for engineering variation into <i>cis</i>-regulatory sequences	156
4.1	Preface	156
4.2	Introduction	156
4.3	Aims	159
4.4	Results	160
4.4.1	Selection of genetic regions	160
4.4.2	Identification of targets	160
4.4.3	Construction of a CRISPR library	161
4.4.4	A wide range of promoter mutations were generated	162
4.4.5	Selection of plant lines for further study	167
4.4.6	Characterisation of plant lines	169
4.5	Discussion	184
4.5.1	Conclusions	190
5	Architectural differences in the promoters of constitutive and variably expressed genes	192
5.1	Preface	192
5.2	Introduction	192
5.3	Aims	195

5.4	Results	195
5.4.1	Gene selection	195
5.4.2	Gene ontology analysis	199
5.4.3	Analysis of chromatin availability	199
5.4.4	Analysis of GC content	202
5.4.5	Transcription factor binding site coverage	204
5.4.6	TF diversity	209
5.4.7	TATA box enrichment	213
5.5	Discussion	215
5.5.1	The CRMs of constitutive and non-specific genes contain more open chromatin than variable and tissue-specific genes	216
5.5.2	Constitutive and non-specific genes contain increased GC content in 400 bp upstream region than variable and tissue-specific genes	216
5.5.3	Variable and tissue-specific genes contain more TFBSs in the 400 bp region upstream of the ATG start codon than constitutive and non-specific genes	217
5.5.4	TF diversity did not differ between gene categories	217
5.5.5	TATA boxes were enriched in variable and tissue-specific promoters	218
5.5.6	Analysis of promoter architecture is useful for learning design features for designing synthetic promoters	219
5.5.7	Conclusions	220
6	Engineering of a genetic feedback loop	222
6.1	Preface	222
6.2	Introduction	222
6.3	Aims	225
6.4	Results	225
6.4.1	Design and construction of programmable synthetic TFs	225
6.4.2	Assessing the performance of synthetic TFs	228
6.4.3	Development of minimal synthetic N-responsive promoters	230
6.4.4	Regulatory effects of TGA1 expression on a synthetic promoter-LucN construct containing TGA1 TFBSs	235

6.4.5	Regulatory effects of TF expression on the luminescence of a synthetic promoter-LucN construct containing TGA1 and NLP7 binding sites	240
6.4.6	Construction of synthetic genetic feedback controllers	246
6.5	Discussion	247
6.5.1	Synthetic transcription factors regulate the expression of p_{ARF18} -LucN	247
6.5.2	Synthetic minimal promoters that respond to specific TFs	248
6.5.3	Synthetic promoters respond to nitrate	252
6.5.4	Towards the development of synthetic genetic feedback controllers	255
6.5.5	Conclusions	256
7	Discussion	257
	Appendices	263
	Appendix Appendix A: Supplementary data for Chapter 2	263
	Appendix Appendix B: Supplementary data for Chapter 4	264
	Appendix Appendix C: Supplementary data for Chapter 5	266
	References	333

List of abbreviations

A. thaliana *Arabidopsis thaliana* 33, 67, 69, 94, 188, 223

A. tumefaciens *Agrobacterium tumefaciens* 74

ABA abscisic acid 39, 40

ABCG ATP-binding cassette G 36

AGI Arabidopsis Genome Initiative 92

AGL16 agamous-like 16 105, 107

ANAC032 NAC domain containing protein 32 62, 67, 68, 99, 102–108, 110, 111, 113–115, 117, 118, 120, 122, 124, 126, 128, 130, 132–140, 142–146, 148–153, 157, 166, 170, 181, 184, 185, 217, 222, 247, 248, 254, 258–261

ANOVA analysis of variance 77, 167

ANR1 Arabidopsis nitrate regulated 1 33, 67, 68, 102, 105–108, 110, 111, 113–115, 118–120, 122, 124, 126, 128, 130, 141, 143, 144, 146, 148, 150, 166, 170, 181, 184, 185, 248, 255, 261

ARF auxin response factor 62, 106, 112, 117–127, 129, 133–135, 144–147, 149, 150, 152, 157, 175, 185, 253

ARF18 auxin response factor 18 62, 63, 67, 68, 85, 87, 89, 90, 93, 99, 103–105, 108–114, 116, 118, 120–122, 124, 126, 128, 130, 132, 134, 139, 143–148, 150, 151, 153, 156–158, 161–163, 165, 166, 173, 175, 183, 185, 216–218, 220–222, 237–239, 244, 245, 247–250, 254–261

ARF2 auxin response factor 2 105, 107

ARF9 auxin response factor 9 [62](#), [68](#), [85](#), [87](#), [89](#), [90](#), [102](#), [105–107](#), [110–116](#), [118](#), [120](#), [122](#), [124](#), [126](#), [128](#), [143–147](#), [152](#), [156–158](#), [161–163](#), [165](#), [166](#), [171](#), [175](#), [176](#), [178](#), [181](#), [183](#), [185](#), [248](#), [249](#)

ASA1 anthranilate synthase alpha subunit 1 [41](#)

AsCpf1 *Acidaminococcus* sp. BV3L6 Cpf1 [69](#), [219](#)

ATAC-seq assay for transposase-accessible chromatin using sequencing [48](#), [87](#), [96](#), [98](#), [107](#), [108](#), [117](#), [119](#), [121](#), [123](#), [125](#), [127](#), [129](#), [135–138](#), [158](#), [193](#), [195](#), [211](#), [218](#), [254–260](#), [268](#)

ATAF1/ANAC002 Arabidopsis NAC domain containing protein 2 [105](#), [107](#), [144](#)

ATP adenosine triphosphate [71](#)

AtUBQ10 *A. thaliana* ubiquitin 10 [68–70](#), [219](#), [221](#)

AtuNOS *Agrobacterium tumefaciens* nopaline synthase [68](#), [69](#), [82](#), [83](#), [88](#), [111](#), [112](#), [114](#), [115](#), [145](#), [216](#)

BAK1 BRI1-associated kinase 1 [37–39](#)

BED browser extensible data [91](#)

BES1 BRI1-ethyl methanesulphonate-suppressor 1 [38](#), [39](#)

bHLH basic helix-loop-helix [244](#)

BIN2 brassinosteroid-insensitive 2 [38](#), [39](#)

BKI1 BRI1-kinase inhibitor 1 [38](#), [39](#)

bp base pair [42](#), [45](#), [57–59](#), [67](#), [81](#), [87](#), [89–92](#), [94](#), [95](#), [97](#), [98](#), [106–108](#), [111](#), [136](#), [144](#), [146](#), [147](#), [156](#), [158](#), [160](#), [162](#), [163](#), [165–167](#), [170](#), [171](#), [173](#), [175](#), [178](#), [185](#), [189](#), [191](#), [193–195](#), [197](#), [199–203](#), [205–211](#), [218](#), [219](#), [223–233](#), [239–241](#), [249](#), [268](#), [269](#), [271](#), [272](#), [275](#), [278](#)

BPC1 basic pentacysteine 1 [167](#), [173](#), [178](#), [181–183](#)

BPC5 basic pentacysteine 5 [167](#), [173](#), [178](#), [181–183](#)

BPC6 basic pentacysteine 6 [178](#), [183](#)

BRI1 brassinosteroid-insensitive 1 [37–39](#)

BRL1 BRI1-like 1 [37](#), [39](#)

BSA bovine serum albumin [71](#), [76](#), [81](#)

BSK3 brassinosteroid signalling kinase 3 [38](#), [39](#)

BSU1 BRI1-suppressor 1 [38](#), [39](#)

bZIP basic leucine-zipper [38](#), [45](#), [144](#), [239](#), [244](#)

bZIP3 basic leucine-zipper 3 [68](#), [223–225](#), [230](#), [231](#), [237](#), [239](#), [241](#), [243–246](#), [250](#)

BZR1 brassinazole-resistant 1 [38](#), [39](#)

CaMV35s cauliflower mosaic virus *35s* [58](#), [59](#), [67](#), [68](#), [71](#), [82](#), [83](#), [111–115](#), [117](#), [119](#), [121](#),
[123](#), [125](#), [127](#), [129–133](#), [135–142](#), [145](#), [221](#), [222](#), [224–234](#), [237](#), [241](#), [242](#)

Cas CRISPR associated [56](#), [57](#), [155](#), [249](#)

Cas12a CRISPR associated protein 12a [69](#), [180](#), [218–222](#), [237](#), [238](#), [245](#)

Cas9 CRISPR associated protein 9 [54](#), [56](#), [57](#), [61](#), [62](#), [69](#), [70](#), [88](#), [155](#), [156](#), [159](#), [161](#), [180](#),
[181](#), [216](#), [218](#), [238](#), [248](#)

CCS circular consensus sequencing [90](#), [154](#), [160](#)

cDNA complementary DNA [85](#), [86](#)

CDS coding sequence [66](#), [68–71](#), [94](#), [111](#), [151](#), [155](#), [188](#)

CEU cohesive end unit [71](#), [72](#)

ChIP chromatin immunoprecipitation [50–53](#)

ChIP-seq chromatin immunoprecipitation sequencing [52](#), [100](#), [144](#)

CHX cycloheximide [50](#)

CIPK23 calcineurin B-like protein-interacting protein kinase 23 [33](#)

CLV CLAVATA [156](#)

CML38 calmodulin-like 38 [38](#)

CNGC15 cyclic nucleotide-gated channel 15 [33](#), [35](#)

CNN convolutional neural network [54](#), [240](#)

Col-0 Columbia-0 74, 75, 77, 86, 88, 94, 98, 159, 167, 168, 170, 171, 173, 175, 176, 178, 179, 181–184, 237, 242

CPK calcium-dependent protein kinase 33, 37

Cq quantification cycle 86

CRE *cis*-regulatory element 41–44, 54, 58, 189

CRISPR clustered regularly interspaced short palindromic repeats 25, 30, 54, 56, 57, 60–63, 75, 88, 91–93, 98, 152–157, 159, 160, 165, 173, 181, 185, 214, 217, 218, 244, 245, 248–250

CRM *cis*-regulatory module 26–29, 42, 43, 45, 47, 58, 61, 96, 105, 145, 151, 152, 190, 193–208, 249, 253, 266–278

crRNA CRISPR RNA 70, 219–221, 237, 238, 245

CTR1 constitutive triple response 1 40

CV coefficient of variation 26, 94, 95, 190–192, 209, 212

CYCB1.1 cyclin-dependent protein kinase B 1.1 36

CYP707A2 cytochrome P450 707A2 40

DamID DNA adenine methyltransferase identification 52

DAP-seq DNA affinity purification sequencing 46, 53, 81, 94, 98, 100, 103, 105, 106, 144, 190, 198–202, 263–265

dCas9 endonuclease deficient Cas9 57, 60, 61, 69, 70, 218, 219, 221, 222, 238

DEX dexamethasone 49, 50, 101, 104, 132, 133, 135, 136, 139–142, 149, 151, 153

DHS DNase I hypersensitive site 46, 47

DMSO dimethyl sulphoxide 67, 87, 88

DNA deoxyribonucleic acid 32, 38, 41–57, 59–62, 67, 71–74, 76, 78, 81, 82, 85, 87, 89, 99–101, 104, 105, 108, 110, 111, 115, 118, 120, 122, 124, 144, 146, 147, 151–153, 155, 181, 189, 190, 215, 216, 244

DNase deoxyribonuclease 46, 47, 50, 51, 85, 188, 209

DNase I-seq DNase-chip with sequencing [47](#), [48](#), [50](#), [53](#)

DNase-seq DNase I hypersensitive sites sequencing [102](#)

dNTP deoxynucleotide triphosphate [67](#), [72](#), [85](#)

DREB26 dehydration response element-binding protein 26 [40](#), [62](#), [67](#), [68](#), [85](#), [87](#), [89](#), [90](#), [98](#), [99](#), [102–108](#), [110](#), [111](#), [113–115](#), [122](#), [123](#), [130](#), [143–149](#), [152](#), [153](#), [157](#), [158](#), [161–163](#), [165](#), [166](#), [178](#), [180](#), [184](#), [185](#), [217](#), [247–249](#), [254](#), [257–262](#)

dsDNA double stranded DNA [57](#), [81](#), [82](#), [92](#), [104](#), [107](#), [109](#), [110](#)

DTT dithiothreitol [71](#), [81](#), [85](#)

DWF1 dwarf 1 [38](#)

E. coli *Escherichia coli* [52](#), [57](#), [62](#), [70](#), [72](#), [78](#), [109](#)

EAR ERF-associated amphiphilic repression [60](#)

EDTA ethylenediaminetetraacetic acid [74](#), [81](#), [85](#)

EF1 α elongation factor 1 alpha [85](#), [86](#)

EIN2 ethylene insensitive 2 [40](#)

EIN3 ethylene insensitive 3 [40](#)

EMSA electrophoretic mobility shift assay [101](#)

EPD Eukaryotic Promoter Database [97](#), [208](#), [278](#)

ERF ethylene response factor [40](#)

ERF2 ethylene response factor 2 [60](#)

ETR1 ethylene response 1 [40](#)

FACS fluorescence-activated cell sorting [49](#)

FAIRE formaldehyde-assisted isolation of regulatory elements [47](#)

FAIRE-seq FAIRE assay with sequencing [47](#)

FDR false discovery rate [98](#), [192](#), [193](#)

FRS9 far1-related sequence 9 [167](#), [178](#), [181](#), [183](#)

GAT Genomic Association Tester [97](#)

GB1 B1 domain of streptococcal protein G [69](#)

GO gene ontology [97](#), [192](#), [193](#), [211](#)

GR glucocorticoid receptor [49](#), [50](#), [71](#), [101](#), [104](#), [105](#), [130](#), [132](#), [135](#), [136](#), [138–142](#), [149](#), [151](#), [215](#)

GRN gene regulatory network [31](#), [36](#), [46](#), [54–56](#), [58](#), [61–63](#), [99–102](#), [152](#), [154](#), [155](#), [189](#), [214–216](#), [247](#), [248](#), [250](#)

HCA2 high cambial activity 2 [173](#), [182](#), [183](#)

HDR homology directed repair [56](#), [57](#), [155](#)

HHO2 HRS1 homolog 2 [68](#), [223–225](#), [232](#), [233](#), [242](#)

HiFi high fidelity [90](#)

IAA indole-3-acetic acid [36](#), [112](#), [116](#), [118](#), [120](#), [122](#), [124](#), [126](#), [128](#), [134](#), [146](#), [147](#), [149](#), [150](#), [152](#), [182](#), [253](#)

IDT Integrated DNA Technologies [67](#), [81](#)

IPA1 ideal plant architecture 1 [156](#)

IPTG Isopropyl β -D-1-thiogalactopyranoside [72](#), [78](#)

kb kilobase [105](#), [108](#), [151](#), [156](#)

KEGG Kyoto Encyclopedia of Genes and Genomes [98](#), [193](#)

KRAB *Staphylococcus aureus* dCas9-Kruppel-associated box domain [60](#)

LB lysogeny broth [72](#), [75](#), [78](#), [88](#)

LBD37 lateral organ boundaries-domain 37 [34](#)

LucF firefly luciferase [68](#), [82](#), [112](#), [114](#), [115](#)

LucN NanoLuc® luciferase 24, 27, 68, 82–84, 111–115, 117–130, 132–142, 145, 146, 148–152, 214, 220–222, 225–229, 232, 234, 236–243

MES 2-(*N*-morpholino)ethanesulphonic acid 74, 76

MMEJ microhomology-mediated end joining 181

MMG 4 mM MES pH 5.6 containing 0.4 M mannitol and 15 mM MgCl₂ 76

MNase-seq micrococcal nuclease digestion followed by sequencing 48

MPRA massively parallel reporter assay 57–59

mRNA messenger ribonucleic acid 37, 41, 44, 59, 182

MS Murashige and Skoog 88, 234

N. benthamiana *Nicotiana benthamiana* 216, 245, 250

NAA 1-Naphthaleneacetic acid 112, 116, 118, 120, 124, 126, 128, 134, 152

ncRNA non-coding RNA 41, 43, 44, 46

NEB New England Biolabs 67, 71–73, 78

NHEJ non-homologous end joining 56, 57, 155, 181

NIA1 nitrate reductase 1 59

NID1 nitrogen response deficiency 1 39

NIR1 nitrite reductase 1 59, 67, 68, 85, 105, 108, 111, 124, 125, 130, 132, 136, 137, 142–145, 148–151, 156, 167, 171, 173, 176, 178, 181–183, 223, 224, 227, 234, 236, 239, 243, 247, 258, 262, 263

NLP NIN-like protein 36, 40, 118, 120, 122, 124, 126, 136, 137, 145, 146, 148, 150, 239, 243

NLP6 NIN-like protein 6 34, 36, 67, 68, 98, 99, 102, 104–108, 110, 111, 113–115, 118, 120, 122, 124, 126–128, 130, 137, 143–146, 148–150, 170, 255–260, 262

NLP7 NIN-like protein 7 33–37, 62, 63, 67, 68, 85, 87, 89, 90, 99, 103–105, 108, 110, 111, 113, 114, 118, 120, 122, 124, 126, 128–130, 132, 137, 138, 140–146, 148–153, 156–158,

161–163, 165–167, 170, 171, 178, 181, 182, 184, 185, 216, 217, 222–224, 226–232, 237–250, 255–260, 262, 263

NLS nuclear localisation signal 69, 221, 245

NOME-seq nucleosome occupancy and methylome sequencing 48

NPR1 nonexpressor of pathogenesis-related gene 1 38

NPTII neomycin phosphotransferase II 88

NRE nitrate-responsive *cis*-element 59, 105, 136, 137, 150, 223, 224, 226–230, 239, 240, 243

NRP nitrate-regulated promoter¹ 59, 68–70, 229, 234, 237, 241, 243, 245

NRT nitrate transporter 40, 41, 223, 242

NRT1 nitrate transporter 1 33

NRT1.1 nitrate transporter 1.1 33–35, 37, 39, 40

NRT1.13 nitrate transporter 1.13 33

NRT2 nitrate transporter 2 33

NRT2.4 nitrate transporter 2.4 156

NUE nitrogen-use efficiency 61, 157, 186, 247–250

O. sativa *Oryza sativa* 188

OCS *Agrobacterium tumefaciens* octopine synthase 88, 114

PAM protospacer adjacent motif 56, 57, 87, 158, 218, 219

PBM protein binding microarray 52, 53, 100

PBS phosphate-buffered saline 81

PCR polymerase chain reaction 47, 48, 51, 53, 57, 67, 70, 72, 86–90, 100, 156, 159, 160, 218

PEAT paired-end analysis of TSSs 188

PEG polyethylene glycol 57, 76

PEPR2 perception of the Arabidopsis danger signal peptide 2 [38](#)

PIN PIN-formed [41](#)

PP2C protein phosphatases of type 2C [39](#), [40](#)

PWMs position weight matrices [54](#), [101](#), [103](#), [105](#), [108](#), [181](#)

PYL pyrabactin resistance-like [39](#)

PYR pyrabactin resistance [39](#)

qRT-PCR quantitative real-time PCR [47](#), [86](#), [87](#), [104](#), [139–142](#), [154](#), [157](#), [167](#), [170](#), [173](#), [175](#), [239](#), [248](#)

Q–Q quantile-quantile [77](#)

RCAR regulatory component of abscisic acid receptor [39](#)

RCF relative centrifugal force [75](#), [76](#), [78](#), [79](#), [83](#), [84](#), [88](#)

Rex human T-cell leukaemia virus type 1 Rex protein [69](#)

RFP red fluorescent protein [49](#), [69](#), [75](#), [237](#), [242](#)

RHD6 root hair defective 6 [40](#)

ΔRn normalized reporter value [86](#)

RNA ribonucleic acid [42](#), [43](#), [45](#), [46](#), [55](#), [56](#), [74](#), [85–87](#), [102](#), [155](#), [167](#), [170](#), [175](#), [215](#), [240](#)

RNase ribonuclease [85](#)

RNA-seq RNA-sequencing [46](#), [50](#), [100](#), [102–104](#), [183](#), [211](#), [212](#)

RPM revolutions per minute [72](#), [75](#), [78](#)

RSA root system architecture [33](#), [36](#), [61](#), [74](#), [157](#), [167](#), [170](#), [173](#), [175](#), [178](#), [181–184](#), [248](#), [249](#)

RSL4 root hair defective 6-like 4 [40](#)

RTV1 related to vernalisation 1 [173](#), [182](#), [183](#)

S. hygrosopicus *Streptomyces hygrosopicus* [69](#)

SAM sequence alignment map [91](#)

SBOL Synthetic Biology Open Language [65](#)

scDNase-seq single cell DNase I-seq [48](#)

scFv single chain variable fragment [69](#), [70](#), [218](#), [219](#), [221](#), [238](#)

SD standard deviation [86](#)

SDS sodium dodecyl sulphate [80](#), [109](#)

SDS-PAGE sodium dodecyl sulphate-polyacrylamide gel electrophoresis [80](#), [109](#)

SELEX systematic evolution of ligands by exponential enrichment [51](#), [53](#)

SELEX-seq systematic evolution of ligands by exponential enrichment followed by sequencing [100](#)

sfGFP superfolder GFP [69](#), [70](#), [218](#), [219](#), [221](#), [238](#)

sgRNA single guide RNA [56](#), [57](#), [60](#), [69](#), [70](#), [87](#), [88](#), [156](#), [159](#), [180](#), [181](#), [218](#), [219](#), [237](#), [238](#), [245](#)

SID mSin Interaction Domain [60](#)

SMRT single molecule real time [89](#), [90](#), [160](#)

SNP single nucleotide polymorphism [162](#)

SnRK1 sucrose non-fermenting 1-related protein kinase 1 [34](#)

SnRK2 sucrose non-fermenting 1-related protein kinase 2 [39](#), [40](#)

SpCas9 *Streptococcus pyogenes* Cas9 [25](#), [87](#), [158](#), [159](#), [218](#)

SRDX SUPERMAN repression domain X [60](#), [69](#), [70](#), [218–221](#), [237–239](#), [245](#), [246](#)

ssDNA single stranded DNA [81](#)

STAP synthetic TALE-activated promoter [216](#)

SV40 simian virus 40 [69](#)

TAD transcription activator-like effector activation domain motif [60](#), [69](#), [70](#), [218–221](#), [237](#), [238](#), [245](#)

TAIR the Arabidopsis information resource [87](#), [91](#), [94](#)

TALE transcription activator-like effector [59–61](#), [216](#)

TAR2 tryptophan aminotransferase related 2 [37](#), [156](#)

TARGET transient assay reporting genome-wide effects of transcription factors [49](#), [50](#),
[101](#), [102](#), [104](#), [130](#), [132](#), [149–153](#)

TARSET transient assay reporting specific effects of transcription factors [99](#), [104](#),
[138–142](#), [145](#), [149–151](#), [153](#), [248](#)

TBE Tris/Borate/EDTA [73](#)

TCP20 TCP domain family protein 20 [36](#), [68](#), [152](#)

T-DNA transfer-DNA [61](#), [156](#), [180](#), [219](#), [220](#), [238](#)

TE Tris-EDTA [85](#), [104](#)

TF transcription factor [24](#), [26](#), [27](#), [29](#), [30](#), [33](#), [35](#), [38](#), [40–46](#), [48–54](#), [58–63](#), [70](#), [71](#), [81](#), [83](#),
[87](#), [94–97](#), [100–105](#), [107](#), [108](#), [110–129](#), [131](#), [133–147](#), [149–157](#), [160](#), [165](#), [170](#), [172](#),
[175](#), [178](#), [180–182](#), [184](#), [185](#), [188–190](#), [202–206](#), [210](#), [211](#), [214–218](#), [220–234](#), [237–241](#),
[243–250](#), [253](#), [266](#), [267](#), [273–277](#)

TFBS transcription factor binding site [26](#), [28–30](#), [41](#), [44–46](#), [50–54](#), [57](#), [59](#), [62](#), [63](#), [67](#),
[81](#), [92–103](#), [105–108](#), [110](#), [111](#), [117](#), [119](#), [121](#), [123](#), [125–127](#), [129](#), [130](#), [133](#), [135–138](#),
[143–146](#), [149–153](#), [155](#), [157](#), [158](#), [165–167](#), [170](#), [172](#), [173](#), [175](#), [178](#), [180](#), [181](#), [183–185](#),
[189](#), [190](#), [198–204](#), [210](#), [212](#), [215](#), [218](#), [223–233](#), [239–242](#), [244](#), [246–249](#), [266](#), [270–272](#)

TFIIIA transcription factor IIIA [167](#), [178](#), [181](#), [183](#)

TGA TGACG sequence-specific binding protein [38](#), [39](#)

TGA1 TGACG sequence-specific binding protein 1 [38](#), [39](#), [68](#), [98](#), [99](#), [102](#), [104](#), [106](#), [107](#),
[110](#), [112–115](#), [130](#), [132](#), [135](#), [136](#), [138](#), [143](#), [144](#), [149–151](#), [153](#), [223](#), [224](#), [226](#), [231](#),
[232](#), [239](#), [241](#), [242](#), [255](#), [256](#), [259](#)

ΩTMV tobacco mosaic virus [68](#), [69](#), [71](#), [83](#), [88](#), [112](#), [114](#), [224](#)

TPL TOPLESS [112](#)

TRAM transcription factor relative binding affinity measurement [82](#), [99](#), [104](#), [109](#), [110](#), [142–147](#), [153](#), [247](#), [248](#)

Tris tris(hydroxymethyl)aminomethane [85](#)

TSR transcription start region [44](#), [188](#), [189](#)

TSS transcription start site [42](#), [44](#), [57](#), [58](#), [67](#), [91](#), [92](#), [94](#), [97](#), [105](#), [111](#), [147](#), [151](#), [158](#), [159](#), [163](#), [164](#), [166](#), [173](#), [188](#), [189](#), [191](#), [194](#), [203](#), [208–210](#), [212](#), [218](#), [219](#), [225](#), [232](#), [233](#), [242](#), [278](#)

TSV tab-separated values [86](#)

UTR untranslated region [42](#), [44](#), [62](#), [63](#), [67–69](#), [71](#), [90](#), [92](#), [94](#), [99](#), [103](#), [105](#), [111](#), [117](#), [119](#), [121](#), [123](#), [125](#), [127](#), [129](#), [133](#), [135–137](#), [143](#), [144](#), [150](#), [151](#), [153](#), [156–158](#), [160](#), [167](#), [171](#), [173](#), [175](#), [178](#), [180](#), [182–185](#), [190](#), [191](#), [193](#), [194](#), [213](#), [224](#), [239](#), [242](#), [248](#), [249](#)

VP64 tetrameric repeat of the minimal activation domain of herpes simplex virus VP16² [60](#), [69](#), [70](#), [218–221](#), [237](#), [238](#), [245](#)

VRN1 vernalisation 1 [173](#), [182](#), [183](#)

W5 2 mM MES (pH 5.7) containing 154 mM MgCl₂ and 5 mM KCl [76](#)

WT wild type [134–138](#), [150–152](#)

Y1H yeast one-hybrid [48–50](#), [61](#), [62](#), [101–103](#), [110](#), [143–145](#), [157](#), [248](#)

YFP yellow fluorescent protein [68](#), [83](#), [112](#), [114](#), [115](#), [117](#), [119](#), [121](#), [123](#), [125](#), [127](#), [129](#), [131](#), [133](#), [135–142](#), [145](#), [214](#), [225–234](#), [236](#), [237](#), [242](#), [243](#)

List of Figures

1.1	Examples of feedforward and feedback loops.	32
1.2	Mechanisms underlying plant nitrate responses.	36
1.3	Nitrate and brassinosteroid response mechanisms.	40
1.4	<i>Cis</i> -regulatory module (CRM) classification.	43
1.5	Yeast one-hybrid assay schematic.	51
1.6	TARGET assay schematic.	52
1.7	DNase I-seq schematic.	53
1.8	DAP-seq schematic.	55
1.9	AND, OR and NOR logic gates with two inputs.	57
1.10	CRISPR-Cas9 gene editing.	58
2.1	Key of Synthetic Biology Open Language symbols used in diagrams throughout this thesis.	67
2.2	Schematic of loop assembly.	68
2.3	The SunTag system.	72
2.4	Protoplast co-expression ratiometric luciferase assay.	85
3.1	A subnetwork of TFs predicted to cross-regulate each other in <i>Arabidopsis</i> in response to N.	105
3.2	<i>In vitro</i> TRAM protein-DNA binding affinity assay.	106
3.3	TARSET assay schematic.	107
3.4	Schematic of the upstream regulatory regions of genes in the nitrogen-response subnetwork showing gene structure and locations of open chromatin and candidate TFBSs.	111
3.5	ARF18 lysate, flow-through and eluate on stained gel after SDS-PAGE.	112

3.6	Relative binding affinity of ARF18::HiBiT fusion to auxin response element positive control compared to random DNA probe.	113
3.7	Edges supported by <i>in vitro</i> TRAM protein-DNA binding affinity assay. . .	114
3.8	Relative luminescence from $p_{CaMV35s}$ -LucN is reduced when co-expressed with $p_{CaMV35s}$ -YFP.	116
3.9	Relative luminescence of $p_{CaMV35s}$ -LucN following co-expression with Arabidopsis transcription factors (TFs).	117
3.10	Relative luminescence of p_{AtuNOS} -NanoLuc® luciferase (LucN) following co-expression with Arabidopsis TFs.	118
3.11	Normalised luminescence of $p_{ANAC032}$ -LucN following co-expression with Arabidopsis TFs.	120
3.12	Normalised luminescence of p_{ANR1} -LucN following co-expression with Arabidopsis TFs.	122
3.13	Normalised luminescence of p_{ARF18} -LucN following co-expression with Arabidopsis TFs.	124
3.14	Normalised luminescence of p_{DREB26} -LucN following co-expression with Arabidopsis TFs.	126
3.15	Normalised luminescence of p_{NIR1} -LucN following co-expression with Arabidopsis TFs.	128
3.16	Normalised luminescence of p_{NLP6} -LucN following co-expression with Arabidopsis TFs.	130
3.17	Normalised luminescence of p_{NLP7} -LucN following co-expression with Arabidopsis TFs.	132
3.18	Normalised luminescence of promoter-LucN reporters following co-expression with $p_{CaMV35s}$ -TGA1.	134
3.19	Normalised luminescence of promoter-LucN reporters following co-expression with $p_{CaMV35s}$ -TGA1::GR with and without dexamethasone treatment.	136
3.20	Normalised luminescence of $p_{ANAC032}$ -LucN with mutated ARF TFBSs following co-expression with ARF TFs.	138
3.21	Normalised luminescence of $p_{ANAC032}$ -LucN with a mutated TGA1 TFBS following co-expression with TGA1 TFs.	139

3.22	Normalised luminescence of <i>NIR1</i> -LucN with a mutated NLP TFBS following co-expression with NLP TFs.	140
3.23	Normalised luminescence of p_{NLP7} -LucN with mutated ANAC032 TFBSs following co-expression with ANAC032.	141
3.24	Summary of significant regulatory interactions determined between TFs and their targets in the N-response regulatory subnetwork.	144
3.25	Summary of significant edges determined between TFs and their targets in the N-response regulatory subnetwork using co-expression luciferase assays .	145
4.1	Selected genetic regions for clustered regularly interspaced short palindromic repeats (CRISPR) targeting.	161
4.2	Schematic of <i>Streptococcus pyogenes</i> Cas9 (SpCas9) expression cassettes. . .	161
4.3	Generation of a <i>cis</i> -regulatory mutation library in Arabidopsis.	162
4.4	Distribution of \log_{10} read counts per amplicon.	163
4.5	Distribution of \log_{10} read counts at each mutation site for each gene.	164
4.6	Deletion size counts in the promoter/5' UTRs of four genes targeted using CRISPR.	165
4.7	Distribution of mutations relative to the TSS.	166
4.8	Characterisation of line plntEPSWT30069-9 containing <i>nlp7-m1-11bp</i>	172
4.9	Characterisation of line plntEPSWT30125-4 containing <i>arf9-m1-12bp</i> , <i>nlp7-m2-13bp</i> / <i>nlp7-m3+1bp</i> and <i>nlp7-m4-1bp</i> / <i>nlp7-m5+1bp</i>	175
4.10	Characterisation of line plntEPSWT30130-4 containing <i>arf18-m1-14bp</i> / <i>arf18-m2+1bp</i> and <i>arf18-m3+1bp</i>	178
4.11	Characterisation of line plntEPSWT30139-9 containing <i>arf9-m2-31bp</i>	181
4.12	Characterisation of line plntEPSWT30134-3 containing <i>dre26-m1-5bp</i>	184
5.1	Mean \log_2 expression of constitutive, variable, non-specific and tissue-specific gene categories in <i>Arabidopsis thaliana</i> across 79 different tissues and developmental stages.	197
5.2	Expression coefficient of variation (CV) distribution of the top 100 constitutive and variable genes compared to all CV ranked genes.	197
5.3	Tau tissue-specificity distribution of the top 100 non-specific and tissue-specific genes compared to all Tau ranked genes.	198

5.4	Mean percentage open chromatin in Arabidopsis <i>cis</i> -regulatory module (CRM) expression categories.	200
5.5	Sliding window analysis of percentage open chromatin in Arabidopsis CRMs.	201
5.6	Percentage open chromatin in Arabidopsis CRMs centred around the transcription start site (TSS).	202
5.7	Mean percentage GC content in Arabidopsis CRMs.	203
5.8	Sliding window analysis of percentage GC content in Arabidopsis CRMs. . .	203
5.9	Mean percentage GC content of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.	204
5.10	Mean percentage transcription factor binding site (TFBS) coverage of Arabidopsis CRMs.	205
5.11	Mean percentage coverage of TFBS within open chromatin regions of Arabidopsis CRMs.	206
5.12	Sliding window analysis of percentage TFBS coverage in Arabidopsis CRMs.	207
5.13	Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.	208
5.14	Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions. . . .	209
5.15	Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs.	210
5.16	Sliding window analysis of TF Shannon diversity of 100 constitutive (blue) and 100 variable (orange) Arabidopsis CRMs.	211
5.17	Two PCA components accounting for the highest percent of variation for Shannon diversity of TF families binding promoters of constitutive and variable genes.	212
5.18	Two PCA components accounting for the highest percent of variation for Shannon diversity of TF families binding promoters of non-specific and tissue-specific genes.	212
5.19	\log_2 -fold enrichment of 15 bp TATA boxes in Arabidopsis CRMs.	214
6.1	Schematic of new edges added by synthetic feedback controller.	225
6.2	Locations of protospacer targets in <i>ARF18</i> of synthetic TFs.	226

6.3	Schematic of SunTag synthetic transcription factors.	227
6.4	Schematic of synthetic dCas12a transcriptional repressors.	228
6.5	Normalised luminescence from p_{ARF18} -LucN following co-expression with synthetic transcriptional activators.	229
6.6	Normalised luminescence from p_{ARF18} -LucN following co-expression with synthetic transcriptional repressors.	230
6.7	Synthetic promoter design.	232
6.8	Synthetic promoter structure.	233
6.9	Normalised luminescence of $p_4 \times [bZIP3\text{-}random]$ -LucN following co-expression with $p_{CaMV35s\text{-}AtbZIP3}$	234
6.10	Normalised luminescence of $p_4 \times [TGA1\text{-}random]$ -LucN following co-expression with $p_{CaMV35s\text{-}AtTGA1}$	235
6.11	Normalised luminescence of $p_4 \times [NRE]$ -LucN following co-expression with $p_{CaMV35s\text{-}AtNLP7}$	236
6.12	Normalised luminescence of $p_4 \times [NLP7\text{-}random]$ -LucN following co-expression with $p_{CaMV35s\text{-}AtNLP7}$	237
6.13	Normalised luminescence of $p_4 \times [NLP7\text{-}random] + spacing$ -LucN following co-expression with $p_{CaMV35s\text{-}AtNLP7}$	237
6.14	Normalised luminescence of $p_4 \times [NLP7\text{-}TATA]$ -LucN following co-expression with $p_{CaMV35s\text{-}AtNLP7}$	238
6.15	Normalised luminescence of $p_4 \times [NLP7\text{-}TATA] + spacing$ -LucN following co-expression with $p_{CaMV35s\text{-}AtNLP7}$	239
6.16	Normalised luminescence of p_{NRP} -LucN following co-expression with $p_{CaMV35s\text{-}AtNLP7}$	239
6.17	Normalised luminescence of $p_2 \times [bZIP3\text{-}random\text{-}NLP7\text{-}random]$ -LucN following co-expression with bZIP3 and/or NLP7.	240
6.18	Normalised luminescence of $p_2 \times [TGA1\text{-}random\text{-}NLP7\text{-}random]$ -LucN following co-expression with TGA1 and/or NLP7.	241
6.19	Normalised luminescence of $p_3 \times HHO2\text{-}minsyn\text{-}105$ -LucN following co-expression with $p_{CaMV35s\text{-}AtHHO2}$	242
6.20	Normalised luminescence of $p_{minsyn\text{-}105\text{-}3 \times HHO2}$ -LucN following co-expression with $p_{CaMV35s\text{-}AtHHO2}$	242

6.21	Normalised luminescence from the roots of transgenic lines containing synthetic promoter-LucN::YFP constructs.	245
6.22	Schematic of synthetic feedback controllers.	247
S1	Location of mutation relative to transcription start site in line 69-9.	264
S2	Location of mutation relative to transcription start site in line 130-4.	265
S3	Location of mutation relative to transcription start site in line 134-3.	265
S4	Mean percentage open chromatin in Arabidopsis CRM expression categories.	267
S5	Sliding window analysis of percentage open chromatin in Arabidopsis CRMs.	268
S6	Percentage open chromatin in Arabidopsis CRMs centred around the transcription start site (TSS).	268
S7	Mean percentage GC content in Arabidopsis CRMs.	269
S8	Sliding window analysis of percentage GC content in Arabidopsis CRMs.	269
S9	Mean percentage GC content of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.	270
S10	Mean percentage TFBS coverage of Arabidopsis CRMs.	270
S11	Mean percentage coverage of TFBS within open chromatin regions of Arabidopsis CRMs.	271
S12	Sliding window analysis of percentage TFBS coverage in Arabidopsis CRMs.	271
S13	Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.	272
S14	Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions.	272
S15	Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs.	273
S16	Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs within open chromatin regions.	274
S17	Sliding window analysis of Shannon diversity of individual TFs predicted to bind Arabidopsis CRMs.	275
S18	Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.	276

S19	Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions.	277
S20	\log_2 -fold enrichment of 15 bp TATA boxes in Arabidopsis CRMs.	278

List of Tables

3.1	DAP-seq sequence logos used to identify candidate TFBSs using FIMO. . .	109
3.2	Calibrators used in the TF co-expression assay.	118
4.1	Details of the genotypes, mutation types and overlapping TFBSs of the four CRISPR library gene targets.	167
4.2	Types of mutations in plant lines selected for further analysis.	168
6.1	DAP-seq consensus sequence logos or sequences used for adding TFBSs in synthetic promoters.	231

List of accompanying materials

File S1 - Table of plasmids used in this thesis (electronic supplementary materials).

File S2 - Table of primers used in this thesis (electronic supplementary materials).

Chapter 1

Introduction

1.1 Gene regulatory networks

To cope with variable environmental conditions without the ability to relocate, plants modulate their metabolism and growth in response to environmental signals. These changes are mediated by complex gene regulatory networks (GRNs) comprised of suites of genes which interact with each other³. GRNs are composed of nodes representing genes, and edges which represent regulatory relationships between nodes. The topology of GRNs can be defined by metrics such as the number of edges each node has to other nodes (node connectivity); the ratio of the number of edges to the number of nodes (network density); the number of edges between the most distal parts of the network (network diameter); and regularly occurring subgraphs or patterns (network motifs)⁴. Information about the topology of GRNs can be used to infer function and elucidate mechanisms that underlie phenotypic robustness, allowing for identification of strategies for engineering the network response⁵. Network robustness often correlates with the cooperativity of nodes, where multiple nodes can interact and form dimers, co-regulating the same target genes⁶.

Network motifs, such as feedforward and feedback loops, are common features of GRNs⁷. Feedforward loops consist of an input gene, X, that regulates an intermediate gene, Y, where both X and Y regulate an output gene, Z⁸ (fig. 1.1). There are eight different types of feedforward loops depending on whether X and Y are positive or negative regulators⁸. Different types of network motifs cause different network dynamics. For example, incoherent feedforward loops, where direct regulation of Z by X is in opposition to the indirect

regulation of Z by X through Y, can cause pulses of gene expression⁷. Negative feedback, where a downstream gene Z represses an upstream gene X which regulates Z directly or indirectly, can improve robustness of the network⁷, and can generate oscillations⁹. Positive feedback often leads to bistability, with genes involved either in the on or off state¹⁰.

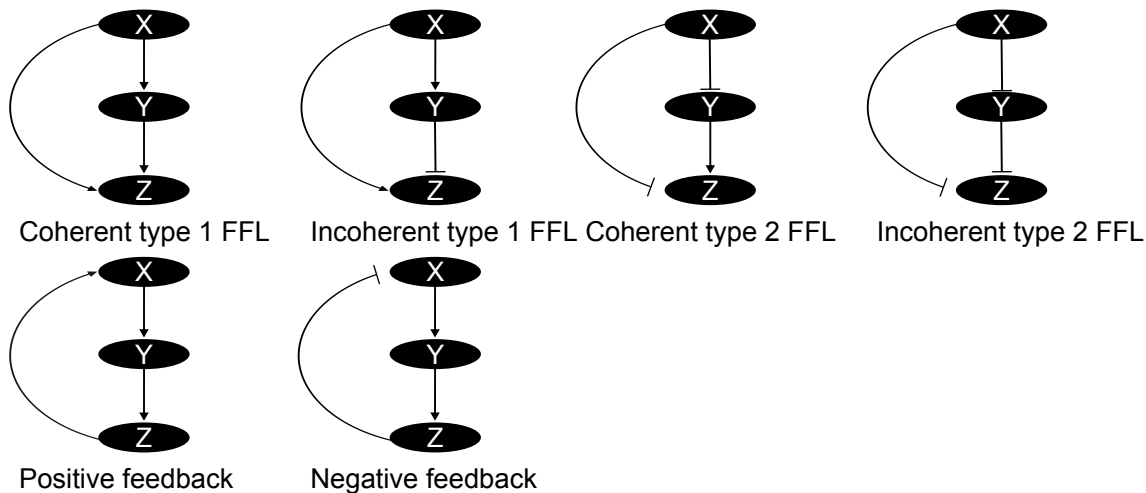


Figure 1.1: **Examples of feedforward and feedback loops.** In coherent feed forward loops (FFLs), the sign of the regulation of the output gene Z by the input gene X is the same as the sign of the regulation of Z by the intermediate gene Y. In incoherent FFLs, the signs of the two paths are the opposite to each other. In positive feedback loops, the downstream gene Z activates the upstream gene X, which regulates Z directly or indirectly. In negative feedback loops, the downstream gene Z represses the upstream gene X, which regulates Z directly or indirectly.

There are multiple levels of biological regulation, including regulation during transcription, splicing, translation, post-translation and degradation of messenger ribonucleic acid (mRNA)¹¹. The timing and rate of transcription can be regulated by various factors including TFs, chromatin structure, deoxyribonucleic acid (DNA) methylation, enhancers and silencers. These factors can activate or repress the transcription of genes in response to environmental signals¹²⁻¹⁷. During splicing, introns are removed from precursor messenger RNA (pre-mRNA) to form mature mRNA. This process can be regulated by alternative splicing, where different combinations of exons are spliced together to form different mRNA variants. Different protein isoforms with unique functions can be generated by alternative splicing, increasing the complexity and diversity of the proteome¹⁸. mRNA degradation is regulated by non-sense mediated decay and microRNA (miRNA)-mediated ribonucleic acid (RNA) decay, and affects how much protein is translated¹⁹. The rate

and efficiency of translation can be regulated by factors such as mRNA structure, RNA-binding proteins and small RNAs²⁰. After protein synthesis, proteins can be modified by post-translational modifications such as phosphorylation, ubiquitination, glycosylation, sumoylation and acetylation, which can alter their activity, localisation, interacting partners and stability²¹. Protein degradation can be regulated by various mechanisms such as the ubiquitin-proteasome system and lysosomal proteolysis²².

This thesis will focus on transcriptional regulation. The availability of genome-wide technologies for investigating expression analysis and protein-DNA interactions enabled plant scientists to collect data from plant cells exposed to stresses such as salt²³, heat²⁴ and nutrient availability^{25,26} and infer the regulatory networks that underpin plant responses to stresses. This thesis will concentrate on plant responses to nitrate.

1.2 Plant responses to nitrate

Nitrogen (N) is an essential nutrient required for plant growth and development²⁷. During the 1960s green revolution, crop yields were improved by adoption of semi-dwarf varieties, and the increased application of nitrogen fertilisers²⁸. Overuse of nitrogen fertilisers can cause negative environmental impacts such as eutrophication of waterways²⁹ and release of nitrous oxide, a potent greenhouse gas³⁰. The development of crops that forage and grow robustly in lower nitrate concentrations is therefore highly desirable. The available forms of nitrogen for plants are nitrate and ammonium³¹. Plants have to cope with variable N concentrations in soil and adjust their root system architecture (RSA) to forage for pockets of N within soils and to adapt to different N conditions. Plants growing in replete nitrate conditions develop fewer lateral roots than plants growing in low nitrate conditions³². N-dependent development of lateral roots allows foraging for nitrate when it is scarce, and conserving energy when it is abundant.

Coordination of these adaptive mechanisms in response to N availability is mediated by sensing systems to balance nitrate uptake with internal supply and demand³³. Nitrate acts as a molecular signal which controls short-term (minutes to hours) to long-term (days) plant responses³⁴. This nitrate induced response causes large-scale transcriptional changes in genes involved with nitrate acquisition, assimilation, signalling, metabolism, carbon metabolism, auxin and organ growth, among other functions²⁵. In *Arabidopsis thaliana*,

nitrate uptake in the roots is mediated by low-affinity nitrate transporter 1 (NRT1) or high-affinity nitrate transporter 2 (NRT2) transporters in the plasma membrane³⁴ (fig. 1.2). An exception to this is the nitrate transporter 1.1 (NRT1.1) (CHL1) transporter, which has both low and high affinity for nitrate³⁵.

1.2.1 The low-nitrate response

In response to limiting nitrate conditions, plant roots undergo several molecular events^{36–38} such as post-transcriptional, calcium and phosphorylation-dependent signalling to induce lateral root development through the regulation of N-transport, signalling, assimilation, hormone and carbon metabolism pathways^{32,39–44}. This enables them to forage for nitrate and adjust their nitrate metabolism. Additionally, these metabolic changes lead to a reduction in shoot and leaf growth. An important post-transcriptional mechanism in the low-nitrate response involves the NRT1.1 transporter. NRT1.1 can switch from a low-affinity to a high-affinity state in response to the nitrate concentration in the environment³⁵. When nitrate levels are low, threonine 101 in NRT1.1 is phosphorylated by calcineurin B-like protein-interacting protein kinase 23 (CIPK23) kinase⁴⁵, switching NRT1.1 from a low to a high-affinity state. NRT1.1 forms a heteroduplex with another membrane protein, cyclic nucleotide-gated channel 15 (CNGC15), which inhibits the Ca²⁺ channel activity of CNGC15, preventing the Ca²⁺ mediated activation of the nitrate response, leading to the inhibition of lateral root growth⁴⁶.

1.2.2 The replete-nitrate response

At replete nitrate concentrations, CNGC15 dissociates from NRT1.1, activating the CNGC15 Ca²⁺ channel, causing the influx of Ca²⁺ into cells⁴⁶. This activates the calcium-dependent protein kinase (CPK)10/30/32 kinases which phosphorylate serine 205 in NIN-like protein 7 (NLP7), a TF which is known to mediate the expression of several genes involved in nitrate metabolism and transport, leading to its accumulation in the nucleus⁴¹.

NRT1.1 also indirectly activates Arabidopsis nitrate regulated 1 (ANR1) which promotes root growth in the presence of nitrate⁴⁷. The membrane protein nitrate transporter 1.13 (NRT1.13) was recently identified as another potential nitrate sensor, and although it cannot transport nitrate, it controls flowering time and branching at low-nitrate condi-

tions⁴⁸. This suggests that plants have multiple ways of sensing and responding to nitrate availability.

Recently, NLP7 was shown to directly bind to nitrate intracellularly, causing post-translational derepression of NLP7⁴⁹. Nuclear retention of NLP7 along with its post-translational derepression activates its transcriptional responses. NLP7 regulates the expression of ~60 % of nitrate-responsive genes⁵⁰, including the upregulation of negative regulators lateral organ boundaries-domain 37 (LBD37) and 38^{26,40}. LBD37 and LBD38 act as repressors of *NRT1.1*^{44,50,51}, reducing nitrate influx and sensing and lowering the Ca²⁺ mediated accumulation of NLP7 in the nucleus; this is an example of negative feedback of *NLP7* expression.

Presence of nitrate causes the degradation of sucrose non-fermenting 1-related protein kinase 1 (SnRK1).1, allowing NLP7 to translocate to the nucleus and induce downstream target genes, leading to enhanced root growth⁵². Presence of ammonium has no effect on the stability of SnRK1.1. SnRK1.1 was also found to interact with NIN-like protein 6 (NLP6)⁵³. The NLP7 nitrate response relies on post-translational regulation and its transcription is not significantly affected by nitrate²⁶.

Thus at replete nitrate concentrations, plants improve nitrate uptake by increasing root growth, and at low nitrate concentrations, plants conserve energy by reducing root growth while switching to higher affinity nitrate uptake. This control of nitrate uptake can be exploited to improve nitrogen-use efficiency (NUE) in crops. For example, overexpression of *NRT1.1* in rice and *A. thaliana* improved NUE in both high and low nitrate concentrations⁵⁴. Changing the expression of other nitrate transporters (NRTs) can also improve NUE^{55–57}. In rice, overexpression of NIN-like protein (NLP)4, a key regulator of N-uptake, assimilation and signalling genes, improved NUE by 38.0 % under low nitrate and 47.2 % under replete nitrate conditions in field trials⁵⁸.

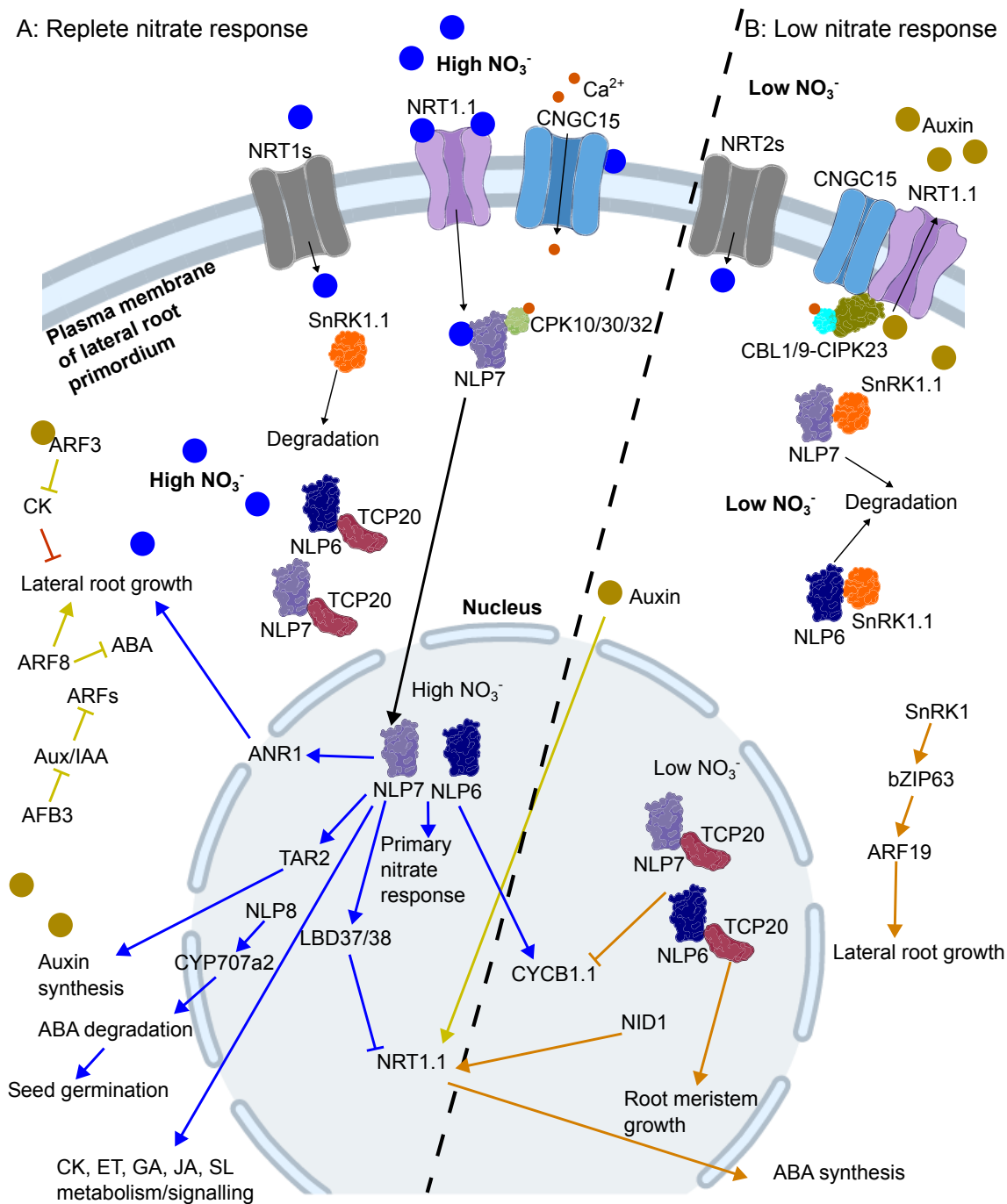


Figure 1.2: **Mechanisms underlying plant nitrate responses.** A: The replete nitrate response. At higher nitrate concentrations, NRT1.1 auxin transport is inhibited, leading to accumulation of auxin in the lateral root primordia, stimulating growth⁵⁹. In replete-nitrate conditions, TCP20-NLP6/7 complexes reside primarily outside of the nucleus. NLP7 binds to nitrate, causing post-translational derepression of NLP7⁴⁹. Increased nitrate represses the NRT1.1 mediated auxin transport and activates CNGC15 Ca_2^+ transport, causing Ca_2^+ influx into the cytoplasm. Presence of nitrate causes the degradation of SnRK1.1, allowing NLP7 to translocate to the nucleus and induce downstream target genes⁵³. The calcium-dependent protein kinases phosphorylate NLP7 causing its nuclear retention and activation of primary nitrate responses⁴⁶. These nitrate responses include nitrate transport, hormone synthesis, organ growth, nitrate assimilation, nitrogen signalling and carbon metabolism²⁵. NLP7 upregulates LBD37 and LBD38 which repress NRT1.1^{44,50,51}, reducing nitrate influx and sensing.

B: The low nitrate response. At limiting nitrate concentrations, NRT1.1 and CNGC15 form a complex which inhibits the CNGC15 Ca_2^+ transport. NRT1.1 transports auxin away from the root tips into basal roots, lowering the auxin concentration in the lateral root primordia, inhibiting growth^{59,60}.

Nitrate starvation: During the carbon/nitrogen starvation response, SnRK1 phosphorylates and activates the TF bZIP63, causing it to regulate downstream genes⁶¹, including ARF19⁶², which enhances lateral root growth in the presence of auxin⁶³. TCP20-NLP6/7 complexes accumulate in the nucleus and up-regulate nitrate assimilation and signalling, support root meristem growth, and down-regulate the G2/M cell-cycle gene CYCB1.1⁶⁴. Figure adapted from Wang et al.⁴⁶, Guan⁶⁵ and Shakeel et al.⁶⁶. CK, cytokinin; ET, ethylene; GA, gibberellic acid; JA, jasmonic acid; SL, strigolactone; ABA, abscisic acid. Arrows represent positive regulation, blunted lines represent negative regulation. The black dashed line separates the two panels.

1.2.3 The nitrate-starvation response

NLP7, NLP6 and TCP20 can form homodimers and heterodimers in all combinations. Under nitrate-starvation (0 mM KNO_3), TCP20-NLP6/7 complexes accumulate in the nucleus and up-regulate nitrate assimilation and signalling, support root meristem growth, and down-regulate the G2/M cell-cycle gene CYCB1.1⁶⁴ (fig. 1.2). TCP20 also regulates root foraging when exogenous nitrate supply is heterogeneous⁴⁴. Unlike NLP homodimers, in replete-nitrate conditions, TCP20-NLP6/7 complexes reside primarily outside of the nucleus.

SnRK1 is an evolutionarily conserved energy kinase sensor which regulates nitrogen and carbon homeostasis in response to energy levels in the plant. Upon nitrogen or carbon starvation, the SnRK1 subunit SnRK1.1 (KIN10) was recently found to phosphorylate NLP7 at serine 125 and serine 306, restricting NLP7 to the cytoplasm and promoting its degradation⁵³.

1.2.4 The role of hormones in the nitrate response

Phytohormones including cytokinins, auxins, ethylene, abscisic acid, gibberellins, jasmonic acid, strigolactones and brassinosteroids control nitrate GRNs and metabolism through signalling feedback⁶⁷, leading to changes in root growth and development⁶⁶. There is crosstalk between many of these phytohormones, with many genes responding to multiple phytohormones⁶⁶.

Cytokinins

Cytokinins adjust shoot development in response to the levels of nitrate sensed in the roots⁶⁷. Cytokinins are satiation hormones, preventing nutrients from accumulating to toxic levels, acting as negative regulators of mineral nutrient uptake⁶⁸. The cytokinin *trans*-zeatin is required for the response of roots to heterogeneous nitrate. Nitrate triggers *trans*-zeatin accumulation in roots, which is transported to the shoot via ATP-binding cassette G (ABCG)14-ABCG11 transporter heterodimers⁶⁹ where it modifies gene expression and metabolism⁷⁰. Primary root elongation⁶⁶ and lateral root initiation^{71,72} is inhibited by cytokinin, but it is currently unknown whether this is due to long distance (root to shoot to root) signalling, or whether it is due to local responses in the roots⁷³.

Auxins

Auxins play a major role in regulation of RSA⁷⁴. The main form of auxin, indole-3-acetic acid (IAA), is primarily produced in young shoot tissues, but is also produced in roots⁷⁵. An auxin gradient is maintained in the lateral root primordia to allow for correct growth, with the highest concentration at the apex⁷⁶. Influx transporters (AUX/LAX family proteins) facilitate the entry of auxin into cells, while efflux transporters PIN and ABCB family proteins move auxin out of cells. As well as being a nitrate transceptor (transporter and receptor), NRT1.1 also acts as an auxin influx carrier in lateral roots, transporting auxin into cells⁶⁰. *NRT1.1* expression is induced by auxin in both roots and shoots⁷⁷. Although *NRT1.1* transcription is strongly induced by nitrate⁷⁸, and its mRNA is more stable in the presence of nitrate, NRT1.1 protein accumulation is repressed by nitrate in lateral root primordia through tissue-specific post-transcriptional regulation⁵⁹. The auxin influx activity of NRT1.1 is inhibited by increasing levels of nitrate⁶⁰ (fig. 1.2). Therefore, NRT1.1 is only active in lateral root primordia at low nitrate concentrations⁵⁹, where it moves auxin into the outer cell layer of the lateral root primordia, which is then exported away from root tips into basal roots (basipetal transport). This lowers the auxin concentration in the lateral root primordia, inhibiting growth⁷⁹. At higher nitrate concentrations, NRT1.1 auxin transport is inhibited, leading to accumulation of auxin in the lateral root primordia, stimulating growth. Auxin levels are also boosted at high nitrate levels through stimulation of auxin biosynthesis. As previously mentioned, at

high nitrate concentrations, NRT1.1 causes an influx of Ca^{2+} into cells⁴⁶, which causes CPK10/30/32 kinases to phosphorylate serine 205 in NLP7, leading to its accumulation in the nucleus⁴¹. NLP7 then directly activates *tryptophan aminotransferase related 2 (TAR2)*, which increases auxin biosynthesis, promoting lateral root development⁵².

There is interplay between auxin, cytokinin and nitrate signalling. Under optimal nutrient and water conditions, plants have a high shoot:root growth ratio. Increasing the auxin:cytokinin ratio promotes root development, while decreasing the auxin:cytokinin ratio promotes shoot development⁶⁸.

Brassinosteroids

As well as the auxin:cytokinin antagonistic control of root growth through nitrate sensing⁶⁸, there is interplay between nitrate, auxin and brassinosteroid signalling. Brassinosteroids are sensed by brassinosteroid-insensitive 1 (BRI1) and its homologues, BRI1-like 1 (BRL1) and 3, in complex with BRI1-associated kinase 1 (BAK1) located on the cell membrane. Binding of brassinosteroids to the receptor complex causes a phosphorylation cascade that phosphorylates a negative regulator of the brassinosteroid signalling pathway, BRI1-kinase inhibitor 1 (BKI1). BKI1 phosphorylation leads to it dissociating from the BRI1 receptor complex, which activates the brassinosteroid signalling effectors brassinazole-resistant 1 (BZR1) and BRI1-ethyl methanesulphonate-suppressor 1 (BES1) (BZR2)⁸⁰. High auxin increases BKI1 expression, which represses brassinosteroid signalling⁸⁰. Under low nitrate conditions, auxin levels are reduced in the lateral root primordia, leading to repression of BKI1 expression, which enhances brassinosteroid signalling which promotes primary root elongation⁸⁰ (fig. 1.3).

Under mild nitrate-deficiency (0.55 mM), the expression of brassinosteroid synthesis gene *dwarf 1 (DWF1)* is upregulated⁸¹. Brassinosteroids enhance root foraging for nutrients under mild nitrate-deficiency by activating the BAK1-brassinosteroid signalling kinase 3 (BSK3) protein complex on the cell surface⁸², which activates BRI1-suppressor 1 (BSU1), which inhibits brassinosteroid-insensitive 2 (BIN2)⁸³. Under severe nitrate-deficiency (0.3 mM), brassinosteroids activate calmodulin-like 38 (CML38) and perception of the Arabidopsis danger signal peptide 2 (PEPR2) which represses root growth⁸⁴.

BIN2 interacts with and phosphorylates TGACG sequence-specific binding proteins 1 and

4 (TGA1 and TGA4), basic leucine-zipper (bZIP) TFs that are highly induced by nitrate⁸⁵ and are known to promote primary and lateral root growth and root hair development⁸⁶. In the nucleus, nonexpressor of pathogenesis-related gene 1 (NPR1) interacts with TGACG sequence-specific binding protein (TGA)4 which increases its DNA-binding activity⁸⁷. Phosphorylation of TGA4 suppresses the interaction of TGA4 and NPR1 and destabilises TGA4. Thus, under mild-nitrate deficiency, brassinosteroids activate TGA4 and potentially TGACG sequence-specific binding protein 1 (TGA1) through inhibition of BIN2⁸⁷, leading to enhanced root growth (fig. 1.3).

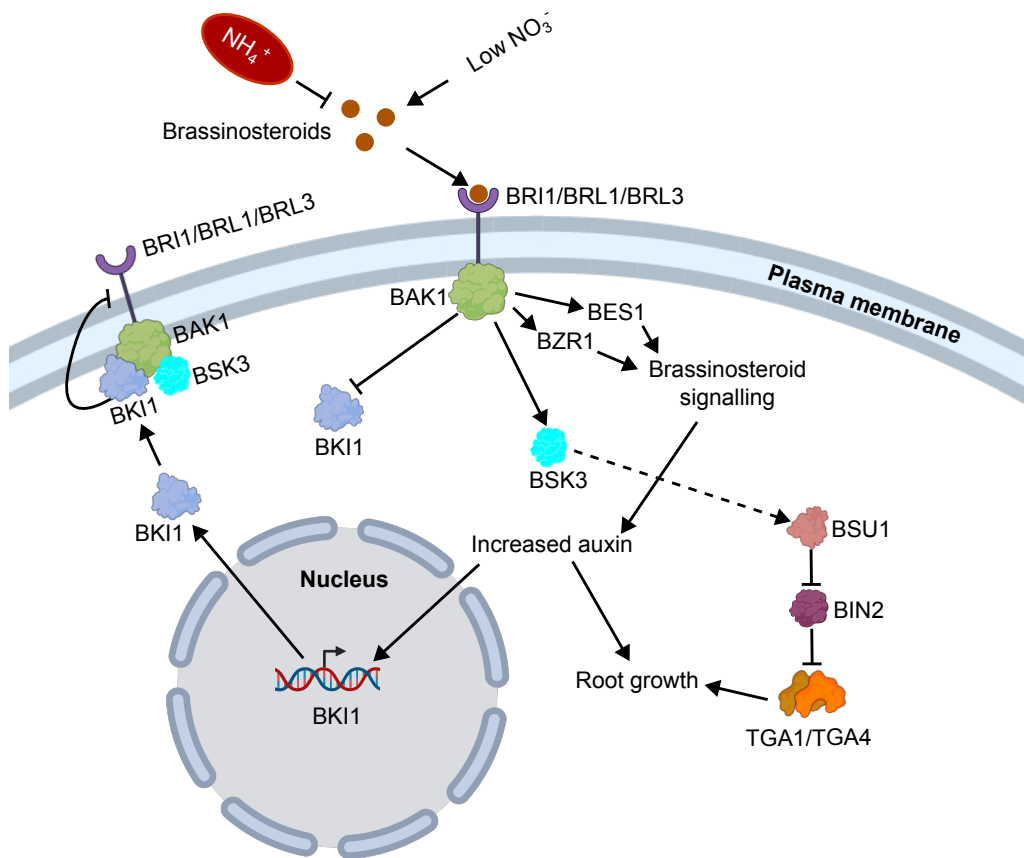


Figure 1.3: **Nitrate and brassinosteroid response mechanisms.** During replete nitrate conditions, BSK1 forms a complex with and inhibits the BAK1 and BRI1 complex. Under mild nitrate deficiency, brassinosteroids bind to BRI1/BRL1/BRL3 receptors and activate BAK1 which phosphorylates BSK3, causing BSK1 to dissociate with the BRI1 complex. BZR1 and BES1 signalling effectors are activated by BAK1, leading to increased auxin biosynthesis. BSK3 is also activated by BAK1, which activates BSU1 through an as yet unknown mechanism, which inhibits BIN2. BIN2 interacts with and phosphorylates TGA1 and TGA4, destabilising them and reducing root growth. Arrows represent positive regulation, blunt lines represent negative regulation, dotted lines represent presumed regulation.

Abscisic acid

Abscisic acid (ABA) regulates many aspects of plant growth, development and stress responses, including the closure of stomata and inhibition of seed germination in response to abiotic stress. Pyrabactin resistance (PYR), pyrabactin resistance-like (PYL) and regulatory component of abscisic acid receptor (RCAR) family proteins combine with clade A protein phosphatases of type 2C (PP2C) to form co-receptors of ABA^{88,89}. PP2Cs form complexes with sucrose non-fermenting 1-related protein kinase 2s (SnRK2s), catalytically inhibiting them⁹⁰. During low nitrate conditions, nitrogen response deficiency 1 (NID1) directly activates the expression of NRT1.1, which causes ABA synthesis and accumulation in root tips through an as yet unknown mechanism⁹¹. At high ABA concentrations, PYR proteins inhibit PP2C activity and derepress SnRK2s including SnRK2.2, 2.3 and 2.6, which phosphorylate downstream effectors and activate stress responses⁹². SnRK2.2, 2.3 and 2.6 also phosphorylate serine 585 in NRT1.1 which results in impaired nitrate transport at both high and low affinities, decreasing nitrate uptake and reduced root growth⁹². During high nitrate conditions, NLP8 directly activates expression of *cytochrome P450 707A2* (*CYP707A2*) which degrades ABA, promoting nitrate-stimulated seed germination⁹³ (fig. 1.2).

Gibberellins

Gibberellins are a group of plant hormones controlling plant growth and development, responding to light, temperature, water and nutrient status⁹⁴. DELLA proteins are GRAS family transcriptional regulators which repress plant growth. Nitrate reduces the abundance of DELLAs by upregulating gibberellin biosynthesis which destabilises DELLAs, increasing Arabidopsis root and shoot growth⁹⁴.

Ethylene

Ethylene was shown to regulate plant N metabolism via constitutive triple response 1 (CTR1) (a negative regulator of the ethylene response) and ethylene insensitive 2 (EIN2)⁹⁵. In the presence of ethylene, the ethylene receptor ethylene response 1 (ETR1) inactivates the CTR1 kinase which causes dephosphorylation and cleavage of EIN2, causing the EIN2 C-terminal fragment to translocate into the cytosol and nucleus. This stabilises the ethylene

insensitive 3 (EIN3) TF which activates ethylene response genes such as ethylene response factors (ERFs)⁹⁶. EIN3 directly represses NRT1.5 in roots⁹⁷, which is involved in root to shoot nitrate transport⁹⁸. Root hair defective 6 (RHD6) interacts with EIN3 and they both co-activate root hair defective 6-like 4 (RSL4) which promotes root hair elongation⁹⁹. Auxin, cytokinin and nutrient starvation were found to also stimulate root hair growth through RSL4, demonstrating that it integrates several signalling pathways⁹⁹. Ethylene biosynthesis is activated by cytokinin¹⁰⁰. Dehydration response element-binding protein 26 (DREB26), also known as ERF012, also increases ethylene biosynthesis in root tips, leading to shorter root tips¹⁰¹ and higher root hair density¹⁰².

Jasmonic acid

Jasmonic acid inhibits plant growth, triggers early reproduction, and induces defence responses¹⁰³. In Arabidopsis, methyl jasmonate induces alternative splicing of NRT1.8¹⁰³. In low nitrate conditions, *anthranilate synthase alpha subunit 1 (ASA1)* is activated by jasmonic acid, promoting auxin biosynthesis and lateral root formation¹⁰⁴.

Strigolactones

Strigolactones regulate plant growth such as root architecture, shoot branching, leaf senescence and reproductive development^{105,106}. Strigolactone biosynthesis is upregulated in the presence of nitrate¹⁰⁷. Strigolactones post-translationally regulate the PIN-formed (PIN) auxin transporters¹⁰⁸ and mediate the plant response to nitrate deficiency¹⁰⁹. In rice, strigolactone activates the *PIN1b* gene which results in root elongation under replete nitrate conditions¹¹⁰.

1.3 Regulation of gene expression

The expression of genes is regulated at several points including transcription, mRNA stability and processing, gene silencing and translation. For many protein-coding genes, the rate of transcription will strongly influence the abundance of the encoded protein¹¹¹. Transcription is controlled by the interactions of certain proteins with DNA regulatory sequences. In eukaryotes, there are two types of DNA regulatory sequences: *cis*-regulatory and *trans*-regulatory sequences.

1.3.1 *Cis*-regulatory DNA sequences

Cis-regulatory sequences are those on the same molecule of DNA as the gene they regulate. The literature varies when defining *cis*-regulatory sequences. For example, Wittkopp and Kalay¹¹² broadly define both promoters and enhancers as *cis*-regulatory elements (CREs). This thesis, in line with Swinnen et al.¹¹³, defines CREs as individual motifs within regulatory sequences that interact with the DNA binding domains of one or more *trans*-acting molecules such as TFs and non-coding RNAs (ncRNAs). Long ncRNAs can contain DNA-binding domains which interact with DNA via triple-helix (triplex) formation¹¹⁴. Therefore, both transcription factor binding sites (TFBSs) and ncRNA binding sites (such as triplex targeting sites) are types of CREs. A *cis*-regulatory module (CRM) is defined as a genomic region encompassing multiple CREs that together regulate an aspect of gene expression pattern¹¹⁵. CRMs include promoters, enhancers, silencers, insulators and locus control regions¹¹⁶. CRMs can be located within exons¹¹⁷, introns¹¹⁸, 5' untranslated regions (UTRs)^{119,120}, or 3' UTRs^{121,122} (Figure 1.4). CRMs can be found tens of kilobases from the core promoter^{123,124}, and are still classed as *cis*-acting if they are found on the same DNA molecule. In this thesis, the term CRM will be used to refer to the sequence of DNA upstream of a start codon encompassing both the promoter region and 5' UTR.

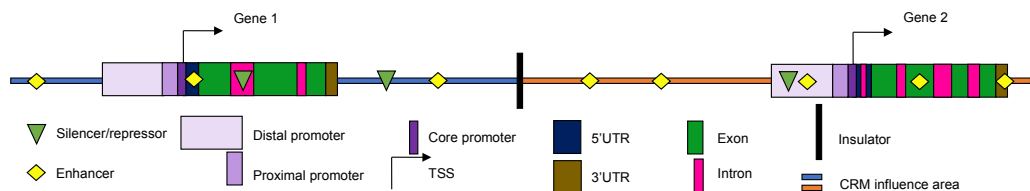


Figure 1.4: ***Cis*-regulatory module (CRM) classification.** CRMs are genomic DNA regions containing multiple *cis*-regulatory elements such as transcription factor binding sites, which together regulate gene transcription. UTR, untranslated region. CRMs include promoters, silencers, and insulators, and can be found within exons, introns, 3'UTRs and 5' UTRs. CRM regulation is combinatorial and can be unique to a specific gene or shared between many. Insulators prevent CRMs from regulating past that region. TSSs are found within the core promoter.

Promoters Promoters are essential in establishing baseline transcriptional capacity through the recruitment of proteins such as TFs, which control the recruitment of the RNA polymerase complex, which for protein-coding genes is typically RNA polymerase

II¹³. Promoters can be divided into two classes depending on when and where they are expressed. Promoters of genes that are expressed in most tissues and developmental stages are classified as constitutive while those that are limited to specific cell-types and conditions are termed variable¹². TFs bind CREs to regulate the spatio-temporal pattern of transcription.

The promoter region can be split into three parts: core, proximal and distal. The minimal DNA sequence required for transcription initiation is called the core promoter and this usually spans from -60 to +40 base pair (bp) relative to the TSS^{125,126}, and can contain CREs such as the initiator, TATA box and B recognition element (fig. 1.4).

Enhancers Enhancers increase promoter activity, improving the spatio-temporal specificity of their target genes. Like promoters, enhancers are DNA sequences that can contain many CREs, recruiting a variety of TFs and ncRNAs. Enhancers help activate transcription regardless of their location, orientation, or distance from the core promoters¹⁴ (Figure 1.4). There are several models explaining how enhancers work: (i) The proteins bound to enhancers and promoters interact with each other by looping^{127–131}. These interactions form the transcription initiation protein complexes. (ii) Some enhancers act by nucleosome remodelling, supercoiling DNA, and altering chromatin structure, facilitating the entry of regulatory proteins to activate transcription¹³². (iii) Both enhancers and promoters share a common regulatory code, with transcription being initiated at both^{133,134}. In this model, RNA polymerase II binds the core promoter where it transcribes DNA in both the sense and antisense directions. This is called bidirectional transcription. At promoters this typically yields long, stable polyadenylated transcripts in the sense direction, and short, unstable non-polyadenylated transcripts in the antisense direction¹³⁵. At enhancers, unstable RNAs called enhancer RNAs are produced in both directions¹³⁶. These ncRNAs, along with cohesin, can stabilise long range enhancer-promoter interactions¹³⁷. In addition, introns containing enhancers can sometimes be more important than the promoter for initiating transcription¹³⁸.

In *Drosophila*, two classes of enhancers were identified that were specific to either constitutively expressed or variably expressed promoters respectively¹³⁹. More than one enhancer can loop to the same promoter, and promoters with more enhancer loop interactions are more conserved than promoters forming fewer enhancer interactions¹⁴⁰.

Silencers Silencers decrease promoter or enhancer activity, reducing expression of their target genes^{141,142}. They are essential for silencing the expression of target genes in certain cell-types or tissues. Like enhancers, silencers contain CREs recruiting a variety of TFs and ncRNAs which can interact with distal promoter or enhancer regions to repress their activity¹⁵.

Insulators Insulators are non-coding DNA sequences which, when located between CRMs and core promoters and bound by certain proteins, prevents the activation or repression of target genes¹⁵. They have two main gene regulatory functions: (i) preventing the spread of repressive or active chromatin along the chromosome and (ii) preventing enhancer-promoter interactions¹⁴³. However, there is limited information on insulators in plants^{15,143,144}, although several animal insulators have been shown to work in plants¹⁴³.

5' untranslated regions In protein coding genes, the 5' UTR is the stretch of DNA found between a TSS and the translation start codon. This region is transcribed into mRNA but is not translated into protein. Plant 5' UTRs can contain a pyrimidine-rich region (5' UTR Py-rich stretch) which improves the transcription rate^{145–147}. The 5' ends of the mRNAs are processed by the replacement of the free triphosphate group with a guanosine triphosphate. Once in place, this 5' cap plays a role in the recognition of the mRNA by the ribosome^{148,149}. The mRNAs of mitochondria¹⁵⁰ and chloroplasts¹⁵¹ are not capped. Once transcribed, the 5' UTR sequence can contain an upstream open reading frame (50% of human 5' UTRs do)¹⁵², a guanine-rich region¹⁵³ and an internal ribosome entry site¹⁵⁴. Alternative 5' UTRs can be produced via alternative splicing or through transcription initiation from a different TSS¹⁵⁵. Because current TSS annotations in the Arabidopsis genome¹⁵⁶ may not be accurate for every gene^{157,158} and the 5' UTR may contain CREs, it is important to include the 5' UTR in regulatory sequence studies to encompass the whole transcription start region (TSR).

1.3.2 *Trans*-acting DNA regulation

A *trans*-regulatory sequence is a DNA sequence that contains a gene encoding for a protein or ncRNA that is used in the regulation of another target gene¹⁵⁹. *Trans*-acting regulation can involve the specific binding of a *trans*-acting molecule such as a TF or ncRNA to a DNA CRE. Another form of *trans*-acting regulation involves non-specific DNA binding by

molecules such as eukaryotic histone proteins in nucleosomes¹⁶⁰.

Sequence specific *trans*-regulation TFs are *trans*-acting proteins with domains that bind to TFBSs in DNA. ncRNAs can contain DNA binding domains which form triple helices with certain *cis*-regulatory elements called triplex target sites (TTS)¹¹⁴. TFs and ncRNAs can contain domains which interact with other proteins to form complexes which can enhance or repress gene transcription. Some TFs can bind DNA as both homodimers and heterodimers, increasing regulatory combinations. DNA regulatory modules such as enhancers can be found on different chromosomes to the gene they are regulating, where they function in *trans*¹⁶¹⁻¹⁶³.

Transcriptional activators and repressors, known as specific TFs, bind to TFBSs in promoters and enhancers to regulate gene expression through protein-protein interactions or by modulating chromatin accessibility. Transcriptional activator proteins facilitate the binding of general transcription factors to the core promoter which form the pre-initiation complex (PIC) via transcriptional co-regulators such as Mediator¹⁶⁴. General transcription factors include TATA-binding protein (TBP), transcription factor II (TFII)A, TFIIB, TFIIE, TFIIF and TFIIH. Binding of TBP to the tata box is the first step in the PIC assembly in TATA-containing promoters¹⁶⁴. General TFs are then sequentially recruited to the core promoter to form the PIC. Independently of the recruitment of general TFs, a transcription co-regulator called Mediator is recruited to enhancer regions via direct interactions with specific TFs bound to these regions¹⁶⁵. Following chromatin looping (see section 1.3.1), Mediator interacts with PIC components and contributes to the recruitment and stabilisation of these components, promoting PIC assembly¹⁶⁵. Once assembled, the PIC is composed of RNA polymerase II and general TFs which are bound to the core promoter, and Mediator which is bound to specific TFs from looping enhancer regions. Mediator stimulates the phosphorylation of the RNA polymerase II carboxy-terminal domain by TFIIH, releasing RNA polymerase II from core promoters which allows for productive elongation across the gene body¹⁶⁵. Once the RNA polymerase II reaches terminator regions at the end of the gene body, it is released from the DNA and the mRNA transcript is released following cleavage at the poly(A) signal¹⁶⁶.

TFs are often part of large families, and many TFs within the same family can bind to highly overlapping sets of binding sites. TFs from different families such as bZIPs and

basic helix-loop-helices (bHLHs) can compete to bind the same TFBS¹⁶⁷. Change in the expression of one TF can result in crosstalk, where the expression of gene targets of other TFs from the same family are also altered¹⁶⁸. Plant TF families have higher expansion rates than many other eukaryotic lineages, not only due to genome duplications, but also due to increased retention of duplications within TF families compared to other plant genes¹⁶⁹.

An important aspect of TF-DNA interactions is the DNA shape. Most DNA-protein interactions occur in the major groove, where bases are most accessible¹⁷⁰, although some occur in the minor groove¹⁷¹. TFs undergo rapid rebinding to ensure the correct orientation is achieved during docking to DNA¹⁷². The spatial arrangement of each base pair and each base in a pair is altered by surrounding bases^{173,174}, and sequence context can influence the minor groove width¹⁷⁵, which can in turn affect the binding affinity of TFs for their DNA targets^{176,177}. DNA shape can be altered by CpG methylation where a bulky methyl group is added, causing widening of the major groove and narrowing of the minor groove¹⁷⁸. TFs can also alter the shape of DNA. Additionally, the overall shape of the DNA molecule can impact the conformation of the TF itself, which could affect its ability to bind to DNA or other proteins¹⁷⁹.

Methylation The DNA base cytosine in CpG dinucleotides can be methylated by DNA methyltransferases that transfer a methyl group from *S*-adenyl methionine (SAM) to the fifth carbon of a cytosine residue, resulting in the formation of 5-methylcytosine¹⁶. By comparing TF binding affinity to TFBSs with and without 5-methylcytosine, it has been suggested that TFs might belong to three main groups: (i) methylation-inhibited, (ii) methylation-insensitive and (iii) methylation-preferred¹⁸⁰. Out of 30 % of Arabidopsis TFs studied, 76 % were sensitive to methylation in their corresponding TFBSs¹⁸¹.

Chromatin Eukaryotic genomes are packed into chromatin, a complex of protein, RNA and DNA¹⁸², made up of repeating units called nucleosomes consisting of a histone octamer wrapped around by 146 or 147 bp of DNA¹⁸³. Many DNA CRMs such as promoters and enhancers are depleted of nucleosomes, and some are nucleosome free¹⁸⁴. Chromatin state influences the accessibility of DNA to TFs and ncRNAs. Several mechanisms such as DNA methylation, histone modifications, chromatin remodelling, and RNA interference can alter

chromatin state¹⁷. Histone modifications are reversible post-translational modifications of histones which include methylation, acetylation, ubiquitination, and phosphorylation. In gene regulatory studies it is important to incorporate data on non-sequence specific interactions between *trans*-acting molecules (*e.g.*, nucleosome histone proteins) and DNA, as they reduce how accessible DNA is to other *trans*-acting molecules. Nucleosomes can inhibit the access of TFs to TFBSs. Chromatin regions highly accessible to DNA-binding proteins are called DNase I hypersensitive sites (DHSs) and these can be mapped following deoxyribonuclease (DNase) I digestion. Pioneer TFs can bind their targets in the presence of nucleosomes, recruiting nucleosome remodellers to open up the chromatin, allowing previously blocked TFs access to the region¹⁸⁵.

1.3.3 Identification of CREs

Gene expression data such as microarray and RNA-sequencing (RNA-seq) data can be used to infer GRNs using the assumption that genes expressed at the same time might perform similar functions. However, these correlation networks cannot confirm transcriptional edges, regulatory interactions between TFs and their targets. To overcome these limitations, data on TF-target interactions can be collected and integrated.

To understand how DNA-protein interactions affect gene function it is necessary to know which proteins bind to specific binding sites in DNA. Earlier methods involved *in vitro* techniques; however, more recent studies identify sites bound by proteins *in vivo*, preserving both context and DNA sequences involved. A recent *in vitro* method called DNA affinity purification sequencing (DAP-seq)¹⁸¹ maps protein-DNA interactions using DNA retaining *in vivo* features. Mapping chromatin accessibility and TFBSs helps inform our understanding of gene regulation, and therefore is the basis of understanding and reconstructing GRNs.

1.4 Identification of chromatin accessible regions

In the past, it was difficult to annotate non-coding DNA regions and, at the genomic scale, it was only possible to detect the presence or absence of ‘regulatory regions’. As detailed in section 1.3.2 (chromatin), CRMs tend to be relatively nucleosome free in which the chromatin has an open structure, exposing the DNA and making it accessible. These

regions are sensitive to digestion by DNase I and are called DNase I hypersensitive sites (DHSs). Many DHS assays have been used to characterise CRMs. In an early assay, DNA that was either untreated or treated with DNase I was fragmented with restriction enzymes, separated by electrophoresis, and transferred to a membrane by Southern blotting. Sequences hypothesised to contain DHSs were then detected by hybridisation to radioactively labelled DNA probes¹⁸⁶⁻¹⁸⁹. This was challenging, limited to short stretches of DNA and not very sensitive¹⁹⁰. DHS methods using polymerase chain reaction (PCR)^{191,192} and later quantitative real-time PCR (qRT-PCR)¹⁹³⁻¹⁹⁵ were developed to improve the identification of chromatin accessible regions. DNase I assays were combined with genome-wide tiling arrays (DNase-chip) either with sequencing^{196,197}, or without sequencing¹⁹⁸⁻²⁰¹ for genome-wide profiling of chromatin accessibility.

As an alternative to DNase assays, a high-throughput assay later called formaldehyde-assisted isolation of regulatory elements (FAIRE) was developed²⁰². In this assay, formaldehyde-crosslinked chromatin and reference non-crosslinked chromatin is extracted using phenol-chloroform. Histone-free DNA is collected in the aqueous phase, and enrichment of histone-free DNA regions from crosslinked DNA are identified by comparing to the non-crosslinked reference on a microarray.

More recent methods combine high-throughput sequencing with previous methods, such as DNase-chip with sequencing (DNase I-seq)²⁰³ and the FAIRE assay with sequencing (FAIRE-seq)²⁰⁴. Although DNase-chip was previously combined with sequencing²⁰⁵, this only identified 20% of DHS sites. High-throughput sequencing allowed production of genome wide DHS maps²⁰³. The combination of DNase I-seq with FAIRE-seq to map open chromatin, however, is more effective than either assay alone²⁰⁶. DNase I-seq and FAIRE-seq require millions of cells as starting material, with time-consuming sample preparations. This means heterogeneity between cells can be drowned out. To improve on this, genome-wide DNase I-seq has since been scaled down to single cells, called single cell DNase I-seq (scDNase-seq)²⁰⁷. The assay for transposase-accessible chromatin using sequencing (ATAC-seq)²⁰⁸ method was also developed. ATAC-seq identifies regions of open chromatin, and analysis of inaccessible unsequenced regions reveals footprints which are typically nucleosome-bound and contain DNA binding proteins. This method builds on the fact that transposons were shown to integrate into active regulatory regions *in vivo*²⁰⁹

and uses Tn5 transposase to integrate its adaptor payload into DNA within chromatin accessible regions. ATAC-seq uses a simple two-step process of Tn5 insertion followed by PCR and sequencing, and has also been demonstrated in single cells^{210,211}

ATAC-seq and DNase I-seq identify regions of open chromatin which correlate to nucleosome occupancy. Nucleosome occupancy itself can be profiled genome-wide using micrococcal nuclease digestion followed by sequencing (MNase-seq)^{212–214}. In MNase-seq, DNA is digested using MNase unless it is protected by nucleosomes. Protected DNA can then be sequenced in a high-throughput manner to identify nucleosome sites. An alternative to MNase-seq called nucleosome occupancy and methylome sequencing (NOME-seq)²¹⁵ uses GpC methyltransferase which methylates GpC dinucleotides not protected by nucleosomes or strongly binding proteins. Native methylation uses CpG dinucleotide cytosine methylation, whereas the added GpC methyltransferase methylates GpC dinucleotide cytosines. Using this information, both nucleosome occupancy and endogenous DNA methylation can be obtained simultaneously.

1.5 Identification of specific DNA-protein interactions

1.5.1 Yeast one-hybrid

The yeast one-hybrid (Y1H) assay is a DNA centred method to identify proteins binding the DNA²¹⁶ (fig. 1.5). It involves a reporter construct where the promoter DNA of interest (termed the bait) is placed upstream of a reporter protein. A TF of interest (termed the “prey”) is fused to a yeast transcription activation domain and is co-expressed with yeast strains containing bait constructs. If the TF binds to the promoter, the yeast activator domain induces expression of the reporter, enabling detection of the interaction. The downside of Y1H assays is that they do not identify whether a TF of interest activates or represses a promoter²¹⁷. Further, false positives can result from transcription initiation by endogenous yeast TFs and false negatives may result from either improper folding, localisation or the lack of post-translational modifications or addition of interacting parts for function. The throughput of Y1H assays were enhanced using a robotic mating platform²¹⁸, and recently, Y1H assays were used to elucidate interactions between plant nitrogen-responsive genes²⁵.

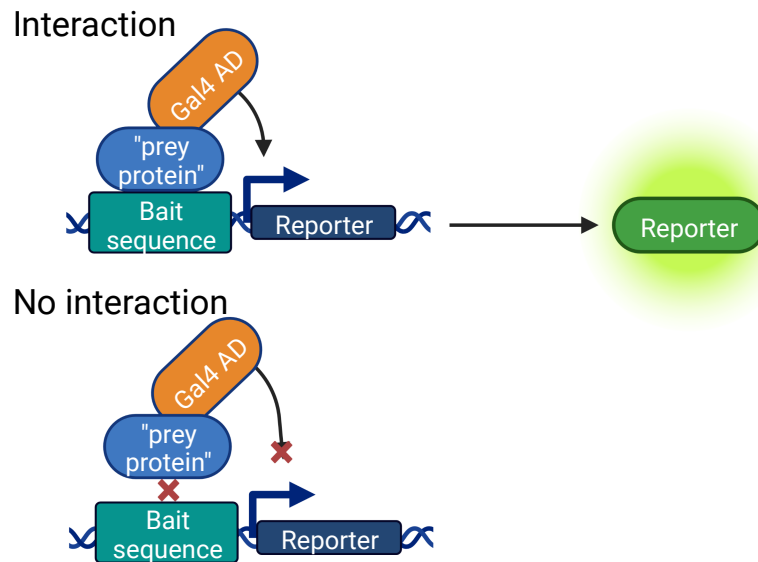


Figure 1.5: **Yeast one-hybrid assay schematic.** The bait DNA (promoter) is placed upstream of a reporter protein. The “prey” TF of interest is fused to a yeast activation domain which induces expression of the reporter upon binding to the bait sequence. The reporter remains inactive in the absence of the “prey” TF. Adapted from Reece-Hoyes and Walhout²¹⁹.

1.5.2 *In planta* expression methods

In planta methods can be used to assess the expression of target genes following TF perturbation by constitutive overexpression, reduction of expression by gene-silencing or the introduction of mutations. However, these methods are time consuming. A technique called transient assay reporting genome-wide effects of transcription factors (TARGET) was developed in plant protoplasts²²⁰ (fig. 1.6). Briefly, the TF of interest is fused to a glucocorticoid receptor (GR) and expressed transiently in protoplasts. The expression plasmid also contains a separate expression cassette with a red fluorescent protein (RFP) enabling fluorescence-activated cell sorting (FACS) of transformed protoplasts²²¹. The GR-TF fusion is retained in the cytoplasm by HSP90-GR binding. Dexamethasone (DEX) is added which disrupts the HSP90-GR binding and therefore enables nuclear localisation. Target gene induction after DEX treatment with and without the presence of a translation inhibitor, cycloheximide (CHX), is measured using RNA-seq analysis. This shows direct and indirect gene regulation. A time-series micro-chromatin immunoprecipitation (ChIP) protocol for use with small cell samples²²² was modified for use with the TARGET assay²²³ allowing both methods to be run on the same samples^{224,225}.

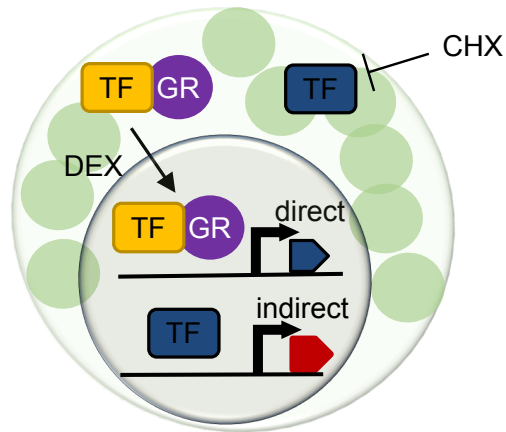


Figure 1.6: **TARGET assay schematic.** TFs are expressed in fusion with a GR-tag that enables translocation to the nucleus following application of dexamethasone. Application of cycloheximide arrests translation, preventing transcriptional cascades, allowing direct regulation to be distinguished.

1.5.3 DNase I

A drawback of using Y1H, TARGET and *in planta* expression techniques is that they do not identify TFBSs. DNase I is used to footprint DNA-protein interactions. *In vitro* DNase I footprinting is where radioactively labelled DNA fragments are incubated with and without the protein of interest, then digested with DNase I and analysed by electrophoresis and autoradiography²²⁶. Bound proteins protect the DNA backbone from digestion by DNase I, allowing identification of specific binding sites of proteins on DNA. In Arabidopsis, recent DNase I-seq data²²⁷ was published, predicting active regulatory regions where many TFs can bind (fig. 1.7).

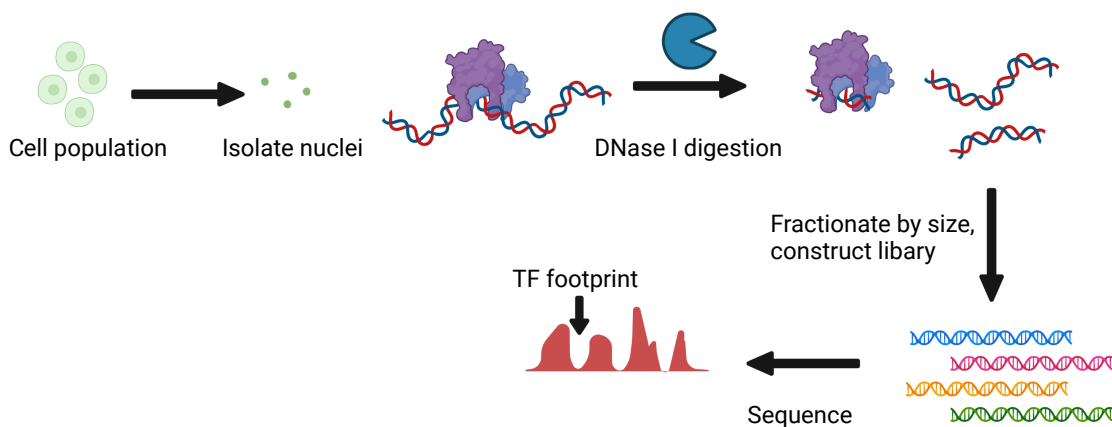


Figure 1.7: **DNase I-seq schematic.** DNase I is used to digest DNA. Regions with bound protein are protected from digestion by DNase I. Unprotected regions are sequenced and aligned to a reference genome and TF footprints are identified allowing identification TFBSs. Adapted from Zeng and Mortazavi²²⁸.

1.5.4 Electrophoretic mobility shift assay

Another method for characterising DNA-protein interactions is the electrophoretic mobility shift assay (EMSA)²²⁹, which quantifies the binding of proteins to DNA. A fragment of DNA containing a putative TFBS is fluorescently or radioactively labelled and is run on a polyacrylamide gel with and without the protein of interest. DNA-protein complexes run more slowly on the gel than DNA alone.

1.5.5 Systematic evolution of ligands by exponential enrichment

Systematic evolution of ligands by exponential enrichment (SELEX) is an *in vitro* assay where libraries of randomly generated nucleic acid ligands of fixed length are flanked by 5' and 3' primers²³⁰. The oligonucleotide library is exposed to a protein of interest, unbound oligos are removed, and then oligos are amplified using PCR. This cycle is repeated several times until the oligo population is enriched for ones which bind protein. These are identified by sequencing and characterised.

1.5.6 Chromatin immunoprecipitation

ChIP was developed to map protein-DNA crosslinking *in vivo*. There are two types of ChIP, differentiated by starting input. XChIP uses chromatin fixed with formaldehyde and sheared by sonication^{231,232}, and NChIP uses native chromatin from nuclei digested by micrococcal nuclease²³³. Protein specific antibodies are used to pull down the protein

of interest bound to DNA. The resulting DNA sequences are purified and analysed using Southern blot, PCR or qPCR²³⁴. ChIP has been combined with sequencing to identify DNA sequences where proteins bind²³⁵, and with DNA footprinting to identify specific TFBSs²³⁶. To increase throughput, ChIP was combined with whole-genome microarrays (ChIP-chip), creating high resolution genome wide maps of DNA-protein interactions^{237–239}.

With the advance of high-throughput sequencing technology, ChIP-seq, where ChIP DNA is sequenced in a high-throughput way has provided greater sensitivity, resolution and specificity than ChIP-chip²⁴⁰. Chromatin immunoprecipitation sequencing (ChIP-seq) is not scalable for high-throughput identification of 1000s of TF-DNA interactions as it is dependent on antibody quality, is difficult for lowly expressed proteins and is challenging to perform²⁴¹.

1.5.7 DNA adenine methyltransferase identification (DamID)

DNA adenine methyltransferase identification (DamID) uses the *Escherichia coli* DNA adenine methyltransferase (Dam) fused to a TF or chromatin protein of interest²⁴². Wherever the protein of interest binds, even transiently, Dam methylates adenine in GATC motifs²⁴³. DNA is extracted and fragmented by the methylation sensitive restriction endonuclease DpnI, followed by amplification and sequencing. DamID is better than ChIP for identifying transient and low-abundance DNA-protein interactions whereas ChIP is good for profiling DNA-protein interactions at short time intervals, as Dam-fusions require expression for several hours to achieve sufficient methylation²⁴⁴.

1.5.8 Protein binding microarray

In the protein binding microarray (PBM), a TF of interest is expressed with an epitope tag which is purified and then applied to a DNA microarray. Any non-bound protein is washed away and then the microarray is stained with a primary antibody specific to the epitope tag. The signal intensity can be quantified as it corresponds to the number of DNA-protein binding events. PBM analysis was utilised in Arabidopsis to identify DNA binding motifs for 63 TFs²⁴⁵.

1.5.9 DNA affinity purification sequencing

The recently developed *in vitro* DAP-seq offers a scalable alternative to ChIP and increased specificity compared to motifs generated by PBMs and SELEX¹⁸¹. DAP-seq uses *in vitro* expressed TFs to find binding locations on genomic DNA retaining its *in vivo* properties (fig. 1.8). Briefly, a genomic DNA library is prepared by attaching a small DNA sequencing adaptor onto purified, fragmented genomic DNA. Separately, affinity-purified TFs are prepared by *in vitro* expression. The DNA library is added to each specific TF, and unbound DNA is washed away. The bound fraction is amplified and sequenced. Reads are mapped to a reference genome and enriched peaks are used to identify TFBSs. As this uses genomic DNA, which contains cell, tissue and organism specific methylation, methylation preferences of TFs can be mapped. This is done using a modified DAP-seq protocol called ampDAP-seq, where a DNA library has any modifications removed by use of a PCR step. This data can be compared to DAP-seq data to assess the methylation preferences of TF binding¹⁸¹. The development of DAP-seq has characterised specific TFBSs from a third of Arabidopsis TFs¹⁸¹. Therefore, TFBS data can now be integrated with the DNase I-seq footprint data to map regulatory regions in Arabidopsis.

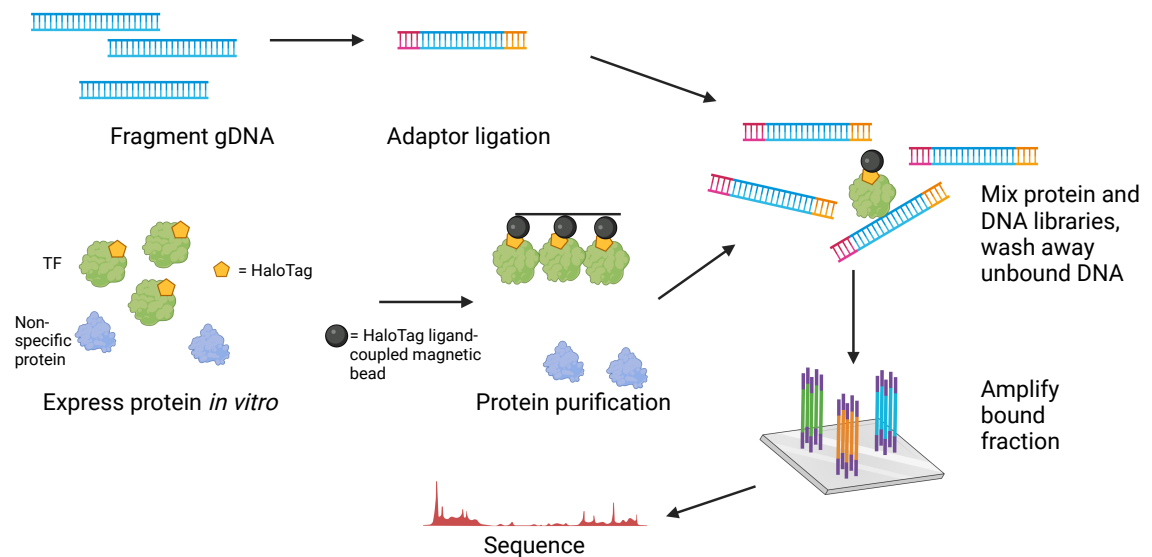


Figure 1.8: **DAP-seq schematic.** DNA affinity purification sequencing allows genome-wide identification of TFBSs. A fragmented DNA library is prepared *in vitro*, and a TF of interest is affinity-purified *in vitro*. The DNA library is added to the TF and unbound DNA is washed away, and the bound fraction is amplified and sequenced. Adapted from Bartlett et al.²⁴⁶.

1.5.10 *In silico* prediction of TFBSs

TFBSs can also be predicted *de novo* with software that can construct position weight matrices (PWMs) describing the binding motifs of TFs, such as MEME²⁴⁷, and the faster STEME²⁴⁸. Motif prediction has a high false positive rate²⁴⁹. Algorithms such as RPM-CMC²⁵⁰ and DREME²⁵¹ offer cofactor motif discovery, which for example can predict if TFs bind in homo or heterodimers. In 2015, convolutional neural network (CNN) models were able to learn filters like PWM motif matching, and these algorithms outperformed PWM algorithms²⁵².

1.6 Engineering regulation

As described in section 1.1, plants mediate their responses to the environment through complex GRNs comprised of suites of genes³. Recent advances in genomics, transcriptomics and tools studying TF-DNA interactions have allowed for increasingly accurate mapping of connections within GRNs. However, the complexity of GRNs means that it is difficult to predict how changes in the network will affect the phenotype of the organism. One aim of synthetic biology is to predictably engineer regulatory systems by tuning existing functions or by adding new functions. Manipulating or engineering GRNs and studying perturbations can provide insights into quantitative network behaviour, including how phenotypes emerge from network functions. There are two main ways to engineer GRNs, either by perturbing existing regulatory systems or by adding new regulatory systems. Various tools can be used to engineer existing regulatory systems, such as genome editing tools. For example, CRISPR-CRISPR associated protein 9 (Cas9) (see section 1.6.1) can be used to disrupt CREs and TF-DNA interactions²⁵³. When adding new regulatory systems such as synthetic circuits, the use of orthogonal parts is desirable to minimise the risk of crosstalk between the new and existing regulatory systems²⁵⁴. These parts can include synthetic TFs and minimal synthetic promoters. Synthetic circuits can convert signals from molecular species to readable outputs. Engineering terminology is often used to describe the different modules in synthetic circuits, such as Boolean logic gates. Biosensors can use various logic gates to detect one or more environmental pollutants. For example, Hg²⁺ and Ag⁺ ions were used as inputs to activate logic gates to provide a fluorescent output²⁵⁵. In this study, two types of logic gates were demonstrated, an AND gate and an

OR gate²⁵⁵ (fig. 1.9). An AND gate is a circuit that converts a signal from two molecular species to a single output. Only when both inputs are present will the output be activated. An OR gate is activated during the presence of either input regardless of the presence of the other input. Another type of logic gate is the NOR gate, which is deactivated when either input is present (fig. 1.9). In yeast, NOR gates have been used to create repression cascades²⁵⁶. In mammalian cells, logic gates were configured to detect and classify micro RNAs to identify the presence of cancer cells and trigger apoptosis²⁵⁷.

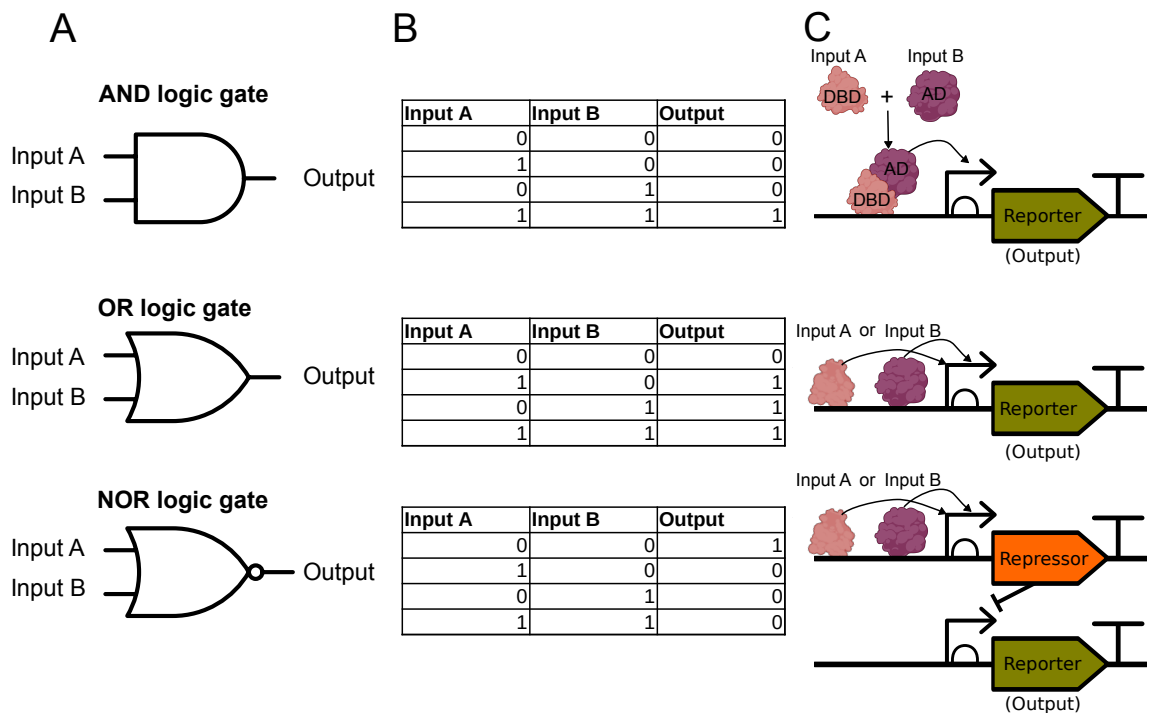


Figure 1.9: **AND, OR and NOR logic gates with two inputs.** A: AND, OR and NOR logic gates with two inputs. B: Logic outputs based on the presence (1) or absence (0) of the two inputs. C: Biological representation of the logic gates. In AND circuit, presence of both inputs is required to activate the output. In OR circuit, presence of either input is sufficient to activate the output. In NOR circuit, presence of either input is sufficient to deactivate the output. DBD, DNA binding domain; AD, activator domain. Adapted from Mendes et al.²⁵⁸.

Synthetic circuits using logic gates can be used to engineer GRNs and network motifs such as feedforward and feedback loops to change network dynamics. Recently, Brophy et al.²⁵⁹ demonstrated the use of many types of logic gates to predictably reprogram root development in plants.

1.6.1 CRISPR-Cas9 gene editing

To modify the DNA sequence or tune the activity of specific genes in GRNs, targeted gene editing tools are required, such as CRISPR-Cas9. CRISPR-Cas systems are prokaryotic adaptive immune systems that confer resistance to foreign nucleic acids such as viral genomes and plasmids²⁶⁰. Bacteria and archaea respond to invading nucleic acids by integrating short fragments (protospacers) into the host genome at the proximal end of the CRISPR array. A CRISPR associated (Cas) protein interacts with guide RNA to recognise and cut the invading DNA.

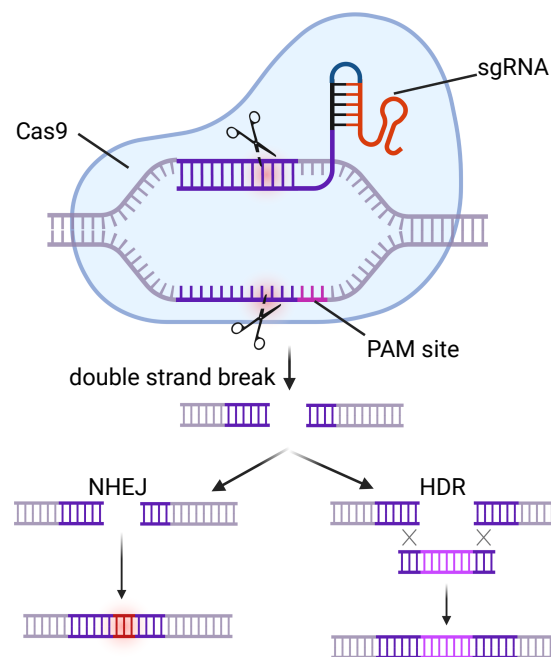


Figure 1.10: **CRISPR-Cas9 gene editing.** A Cas9 nuclease and sgRNA form a complex together and recognise the target spacer DNA sequence when there is a 5'-NGG-3' PAM site located immediately 3' of the protospacer. The Cas9 cleaves the DNA, and DNA is repaired mainly by NHEJ, but also by HDR which can be utilised to insert a desired sequence into the genome.

There are two classes of CRISPR-Cas systems, with multi-subunit effector complexes in class 1, and single-protein effectors in class 2²⁶¹. Many class 2 systems have been repurposed for genome editing, with the type II CRISPR-Cas9 system most widely used²⁶². In this system, a Cas9 nuclease and sgRNA are used, which form a complex together (fig. 1.10). The complex recognises target spacer DNA sequence when there is a 5'-NGG-3' PAM located immediately 3' of the protospacer, and the Cas9 cleaves the DNA. In plants,

double stranded DNA (dsDNA) breaks are repaired mainly by NHEJ, which can lead to insertions and deletions that disrupt the function of genes²⁶³. Alternatively, HDR can be utilised to insert a desired sequence into the genome, though this method has very low efficiency²⁶⁴. Cas9 proteins engineered to have no endonuclease activity (endonuclease deficient Cas9 (dCas9)) can also be fused to an effector domain such as a transcriptional activator or repressor domain to regulate target gene expression²⁶³.

CRISPR-Cas components can be delivered into plants using a variety of methods. Polyethylene glycol (PEG)-mediated transformation of plant protoplasts (cells where the cell wall has been removed) is efficient but regeneration of plants from protoplasts can be difficult²⁶⁵. Although preassembled Cas9-sgRNA ribonuclease complexes can be directly delivered into cells to avoid the introduction of a transgene²⁶⁶, *Agrobacterium*-mediated transformation of plants is the most widely used delivery system for CRISPR-Cas components. Transgenic plants are screened for the presence of the desired edits using PCR and sequencing. In most species, transgenes can be segregated away from the desired mutation in progeny generations.

1.6.2 Development of synthetic promoters for microbes

Many studies have described the assembly and characterisation of promoter libraries for microbes. For example, in *E. coli*, directed evolution was used to engineer a broad range of precise promoter strengths²⁶⁷. Promoters made in this way tend to have a broad host range²⁶⁸. Combinatorial libraries of random promoter architectures have also been achieved in *E. coli*²⁶⁹. They show that gene expression can be regulated by over five orders of magnitude. The location of regulatory elements within a promoter was key to determining the expression level, with those located closer to the TSS more effective than those further away. Some regulatory elements were dominant over others when located in the same promoter²⁶⁹. Synthetic promoters have also been used to study nucleosome occupancy effects on transcription and have enabled the rational design of synthetic promoters with increased transcription²⁷⁰. A massively parallel reporter assay (MPRA) was used to test semi-random 312 bp constitutive and 246 bp responsive promoters in *S. cerevisiae*, where conserved regions containing TFBSs remained the same²⁷¹. Machine learning facilitated the *in silico* design of new promoters of specific expression levels, including promoters with higher expression than any promoters in the test or training datasets. When tested *in vivo*

these designed promoters had similar expression levels to the predicted values. Synthetic minimal yeast promoter libraries were generated to identify minimal upstream activating elements, and this led to the development of a promoter reduced by 80 % in size which still achieved high expression levels²⁷².

1.6.3 Synthetic regulatory elements for plants

The discovery of CRMs and the CREs within them is the first step in the engineering and rewiring of GRNs to improve plant growth and development¹⁵. Promoter studies in yeast²⁷¹⁻²⁷³ and human cell lines²⁷⁴ often rely on high-throughput testing of randomised promoters to identify synthetic promoters with the desired characteristics. High-throughput reporter assays such as MPRA are technically challenging in plants as both the delivery of libraries in a manner that enables each cell to express only a single construct, and the sorting of cells by reporter expression levels is difficult. Despite this difficulty, a large-scale study using self-transcribing active regulatory region sequencing (STARR-seq) measured the activity of 170 bp core promoters from Arabidopsis, maize and sorghum, and demonstrated that promoter elements such as the TATA box, as well as GC content affect promoter strength²⁷⁵. The authors then used key promoter features to design synthetic promoters.

Minimal synthetic promoters are generally composed of a core promoter region containing the TSS as well as contain core motifs such as the TATA-box, GA elements, Y patch and initiator²⁷⁶. In plants, the most used core promoter is the minimal cauliflower mosaic virus *35s* (*CaMV35s*) promoter, consisting of 54 bp containing a 7 bp TATA box and a TSS, which has been shown to drive basal expression in both dicots and monocots²⁷⁷. The core promoter is fused to upstream operator sequences that contain binding sites for the synthetic TF. Recruiting multiple copies of the synthetic TF to the synthetic promoter can further increase the transcriptional effect. This can be done by simply increasing the number of binding sites in the operator region^{259,278}.

Synthetic promoters were used to study flanking sequence changes of tissue-specific CREs in rice²⁷⁹. This identified functional expression pattern changes and led to the design of five synthetic promoters which were specific to different green tissues. Increasing the number of TFBSs in pathogen-inducible synthetic promoters increased transcription in

Arabidopsis²⁸⁰. However, too many repeated *cis*-regions in a promoter led to an increase in background induction of promoters, possibly by increasing basal level of transcription by allowing increased binding of normally low affinity TFs. Increased motif copy number was also shown to increase expression in tobacco²⁸¹ and rice²⁸².

Synthetic promoters can be built from larger regions of natural promoters such as the synthetic nitrate-inducible nitrate-regulated promoter¹ (*NRP*) which contains a 180 bp region from the nitrate reductase 1 (*NIA1*) promoter, a 131 bp region containing the nitrate-responsive *cis*-element (*NRE*) motif from nitrite reductase 1 (*NIR1*), and the *CaMV35s* minimal core promoter¹. Other approaches to building synthetic promoters with desired functions have utilised large-scale experimental datasets. For example, Jores et al.²⁷⁵ combined MPRA with machine learning to optimise native promoter motifs and generate synthetic promoters for plants. In yeast, a deep-learning method trained on genomic and transcriptomic data was used to generate promoters with pre-specified target mRNA levels²⁸³. A library of synthetic transcription activator-like effector (*TALE*)-activated promoters has been developed in plants²⁸⁴, demonstrating a range of promoter strengths. Finally, rational approaches can be applied in which known TFBSs are arranged into the operator region of a synthetic promoter, similar to the approach used to build promoters that attract synthetic TFs^{278,281,285,286}. For these, the type, number and spacing of elements can be used to tune the expression level^{278,281}. Different spacing between motifs can change promoter responses to TFs²⁸⁷. Furthermore, sequences that flank TFBSs have also been shown to affect binding affinity and therefore the strength of promoters²⁸⁵.

1.6.4 Synthetic TFs

The use of synthetic TFs allows control over the expression of transgenes and, more recently, endogenous genes. As the expression of synthetic TFs can be controlled by tissue-specific or constitutive promoters, they can enable both constitutive or tissue-specific control of target genes. Synthetic TFs generally contain an effector domain (a transcriptional activator or repressor) as well as a DNA-binding domain. Inducible synthetic TFs also generally contain a ligand-binding domain that interacts with the inducing molecule²⁸⁸.

The most widely used activation domain is four tandem copies of the VP16 domain (VP64) from herpes simplex virus²⁸⁹, which recruits PIC components of the transcriptional

machinery to the promoter²⁹⁰. Other activation domains used in plant synthetic TFs include Arabidopsis ethylene response factor 2 (ERF2)²⁵⁹ and GAL4^{291,292}. The most used repressor domain is the SUPERMAN repression domain X (SRDX) domain derived from the tobacco ERF-associated amphiphilic repression (EAR) domain^{293,294}, which recruits chromatin remodellers. Other repressor domains used in synthetic plant TFs include the human mSin Interaction Domain (SID) and *Staphylococcus aureus* dCas9-Kruppel-associated box domain (KRAB) domains²⁹⁵. The KRAB domain does not work well in plants, possibly because the mechanism by which it recruits TFs is not conserved²⁹³.

Several DNA binding domains have been used to build orthogonal synthetic TFs. Domains from the yeast GAL4 protein and human LexA protein have been particularly widely used²⁹⁶. The binding sites for these TFs are added to synthetic promoters that are used to drive the expression of the transgene. However, to be able to bind to any promoter in the genome, it is desirable to be able to reprogram the DNA binding domain. Programmable synthetic TFs include TALEs, in which the DNA-binding domain can be reprogrammed to recognise a specific locus²⁹⁷. The use of dCas9-based synthetic TFs also allows for binding to specific sequences by recoding the sgRNA(s) to recognise the desired target sequence²⁹⁸. To improve dCas9-based synthetic TFs, additional synthetic elements such as the SunTag system can be used. In this system, dCas9 is fused to a tandem array of GCN4 peptides that can recruit multiple copies of the tetrameric repeat of the minimal activation domain of herpes simplex virus VP16² (VP64) activator domain^{299,300}, increasing the activation of the target gene (see fig. 2.3). Recently, the CRISPR-Act3.0 system was developed. This couples the SunTag system with an MS2 aptamer to recruit multiple SunTag peptides containing multiple activator domains to target promoters³⁰¹. The CRISPR-Act3.0 system increased activation of target genes to >250 fold when the 2×transcription activator-like effector activation domain motif (TAD) activator domain was used.

To enable synthetic TFs to activate their downstream synthetic promoter, their expression must also be controlled. Expression of synthetic TFs can be controlled by a natural promoter such as a sequence from the host genome. This is often a constitutive promoter, however, promoters known to express in the desired cells or tissues of interest can also be used^{259,297}. Alternatively, synthetic promoters can be used.

1.6.5 Towards engineering regulation in plants

Traditional crop breeding relied on imprecise methods such as physical or chemical mutagenesis and selection, marker assisted breeding, transgenesis, and transfer-DNA (T-DNA) insertion. The advent of CRISPR-Cas9 gene editing has enabled the introduction of precise mutations to target genes in many crop species conferring novel traits^{302–304}.

In addition, more accurate mapping of regulatory regions is providing opportunities for fine tuning of gene regulation. This can be achieved using gene editing tools to make specific changes to CRMs^{305–307}. In addition, synthetic TFs such as TALE- and dCas9-based TFs can be used to tune the expression of endogenous genes^{299,301,308}. Finally, advances in synthetic biology approaches detailed above, including the development of synthetic promoters and TFs³⁰⁹, provides opportunities for more sophisticated perturbations of gene expression^{259,310}. Synthetic regulation of transcription has the potential to enable network rewiring by altering native TF-DNA networks. Previous studies have suggested that the behaviours of even complex plant regulatory networks can be rewired by the introduction of feedback. Various ideas for introducing feedback into biological systems have been raised^{311–314}, including in plants³¹⁵. For example, Foo et al.³¹⁵ propose the use of a synthetic genetic feedback controller to positively regulate plant genes in response to infection. They used Y1H and time-series transcriptomics to predict edges in a pathogen-response GRN and used modelling to suggest that a feedback loop could improve the plant response to pathogens.

1.7 Gaps in knowledge and experimental approaches

The overall goal of the work in this thesis is to identify approaches and methods to rationally engineer plant N-responses with the goal of improving NUE by modulating the expression of TFs involved in nitrate transport, assimilation, and signalling. This is because TFs regulate many genes in a coordinated manner, and so are likely to be more effective targets for engineering than individual genes. This has been demonstrated in previous studies, where modulating the expression of TFs such as NLP4 and dehydration response element-binding protein (DREB)1C improved NUE^{58,316}. A putative network uncovered by Y1H studies identified TFs that interact with the promoter regions of genes involved in nitrogen processes, as well as known nitrate-response TFs²⁵. 21 TFs within this network

were predicted to coordinate plant responses to nitrate. These were investigated further using mutant lines to analyse RSA and shoot development under limiting (1 mM KNO₃) and replete (10 mM KNO₃) nitrate. This GRN showed evidence of network motifs such as feedforward and feedback loops, but these need to be confirmed experimentally. The Y1H assay provides evidence of regulatory interactions but does not provide information about TFBSs within the target gene or information about any regulatory consequences of TF-DNA interactions. The Y1H dataset provided by Gaudinier et al.²⁵ is a promising starting point for engineering the N-response subnetwork, but to be able to predictably manipulate this network, it needs to be elucidated further. Once the network is characterised, it can be engineered either by the addition of new edges or through disruption of existing edges. With knowledge of the existing network, the effects of these manipulations on the expression of downstream genes and, potentially phenotype, can be predicted.

In chapter 3, I describe the characterisation of interactions between TFs in a putative N-response subnetwork identified by Gaudinier et al.²⁵. I identify TFBSs in upstream regulatory regions and characterise their interactions with TFs using *in vitro* methods and transient assays. Using these data, I show that auxin response factor (ARF)9 and 18 are involved in a feedforward loop to NLP7 through NAC domain containing protein 32 (ANAC032) and potentially through DREB26. To disrupt edges in this feedforward loop, in chapter 4 I use CRISPR technologies. CRISPR-Cas9 genome editing can be used to disrupt edges, or even to engineer new recognition sites into target promoters. However, it is difficult to predict the effect of a given mutation on gene expression. The use of CRISPR libraries can introduce novel variation into coding or promoter regions, which can potentially rewire GRNs. I therefore use a library approach to introduce mutations into the promoters and 5' UTRs of *auxin response factor 9 (ARF9)*, *auxin response factor 18 (ARF18)*, *DREB26* and *NLP7*, creating a library of 327 lines with mutations across the promoter regions of the four genes. Five lines containing interesting mutations, some of which overlapped TFBSs, showed alterations to gene expression and root phenotype.

I also explore network engineering by the introduction of synthetic genetic feedback controllers. A synthetic feedback controller was recently engineered and tuned in *E. coli*³¹⁷, however, synthetic feedback controllers have not been demonstrated experimentally in plants. To be able to introduce feedback into the N-response network, synthetic elements

need to be developed and tested, including nitrate-responsive synthetic promoters and synthetic TFs. To explore the development of synthetic promoters that respond to specific TFs, I needed more insight into promoter elements in plants. In chapter 5, I describe the comparison of promoter architectures of four different categories of Arabidopsis genes with different expression patterns. Distinct differences in promoter architecture were found between constitutively expressed, variably expressed, non-specific and tissue-specific genes, such as TFBS density, percentage open chromatin and enrichment for TATA box motifs. Design features learnt from this chapter aided the construction of minimal synthetic promoters for use in synthetic genetic feedback controllers. In chapter 6, I describe the design and construction of synthetic genetic elements that introduce positive and negative synthetic genetic feedback into the N-response subnetwork. Using structural information about the subnetwork (chapter 3) I identified a target TF, ARF18, to activate/repress. Synthetic TFs were designed to bind to the upstream regulatory region of *ARF18*. These were confirmed to either activate or repress its transcription. To allow these synthetic TFs to respond to genes further down the N-signalling hierarchy from ARF18, thus creating a feedback loop, I designed a suite of synthetic minimal N-responsive promoters with binding sites for NLP7. Additional synthetic promoters were built with binding sites for other early responders to nitrate³¹⁸. The best performing synthetic promoters in transient protoplast assays were also confirmed to respond to nitrate in plant roots. Synthetic feedback controllers were assembled using the best minimal synthetic N-responsive promoters controlling the best synthetic TFs. Arabidopsis plants were transformed with the genetic feedback controllers, although analysis of these lines is outstanding.

To summarise, the overall goal of this thesis is to identify approaches and technologies for predictable engineering of plant GRNs using a subnetwork that regulates plant responses to nitrate as an exemplar. This is achieved through the following:

- The characterisation of edges in a regulatory subnetwork for nitrate-associated metabolism and growth including identification of binding sites and regulatory consequences of binding.
- The addition of novel genetic variation into the promoter and 5' UTR regions of four genes encoding TFs involved in regulating root growth in response to nitrate using a CRISPR library.

- The identification of architectural differences in promoters with different expression patterns.
- The design, assembly and quantitative characterisation of synthetic minimal N-responsive promoters, and synthetic transcription factors for use in the construction of genetic feedback controllers.

Chapter 2

Methods

This chapter contains the methods used in chapters 3 to 6. The relevant methods sections are referenced throughout the experimental chapters. All plasmid constructs used in this thesis are found in File S1, and all primers are found in File S2. All gene construct diagrams in this thesis are in the Synthetic Biology Open Language (SBOL) visual standard (v3.0)³¹⁹ (fig. 2.1).

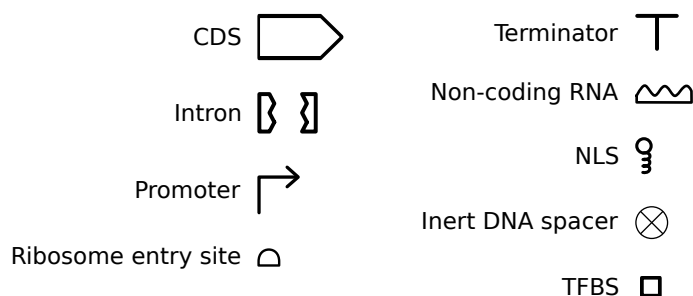


Figure 2.1: **Key of Synthetic Biology Open Language symbols used in diagrams throughout this thesis.** Symbols used in diagrams throughout this thesis are in the SBOL visual standard (v3.0)³¹⁹. CDS, coding sequence. NLS, nuclear localisation signal. TFBS, transcription factor binding site.

2.1 Assembly and validation of plasmid constructs

Plasmids for expression in plant cells were assembled using the Loop modular cloning system³²⁰ (fig. 2.2). Plasmid sequences were designed and edited using Benchling design software (available at <http://www.benchling.com>). New standard parts such as promoters

and coding sequences (CDSs) were cloned into a Universal Level 0 acceptor (pUAP1³⁰⁹, pUPD2³²¹ or pUAP4³²²). Level 0 standard parts were assembled into acceptor plasmids to create Level 1 assemblies, each consisting of a complete transcriptional unit. Where required, Level 1 assemblies were assembled together to create multigene constructs according to the Loop syntax³²⁰.

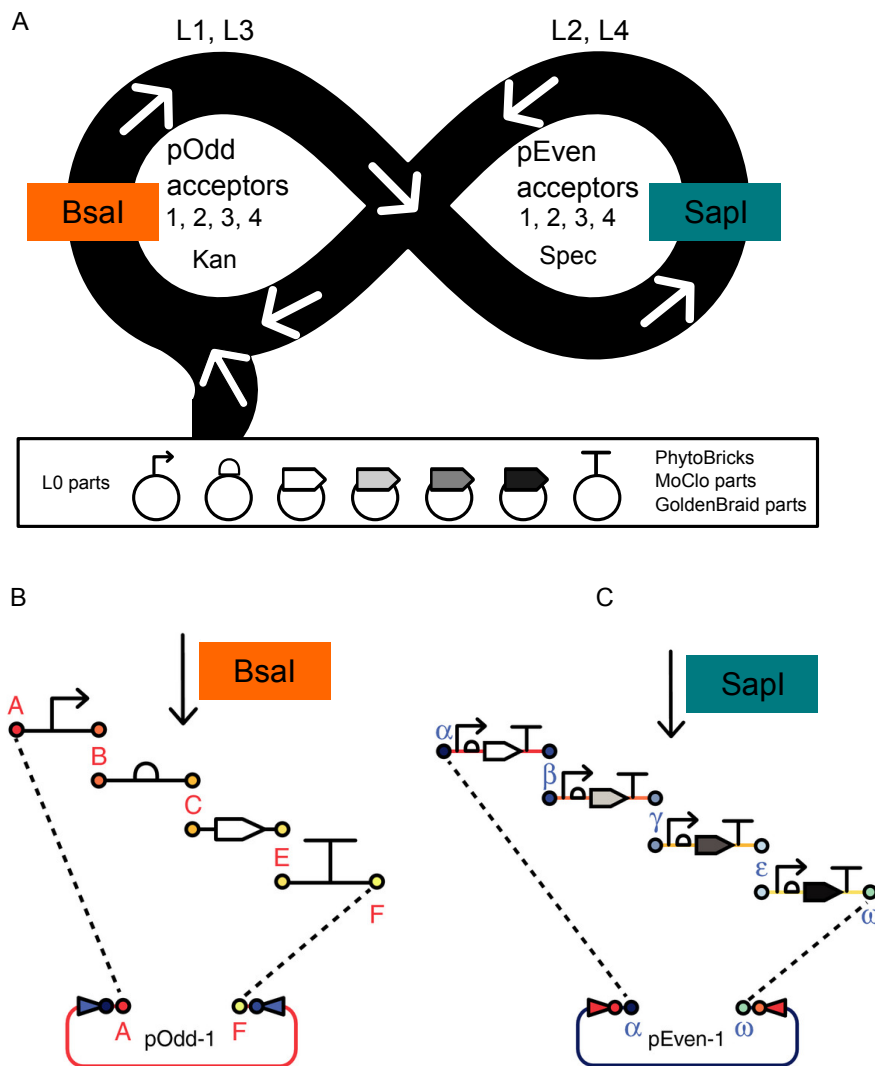


Figure 2.2: **Schematic of loop assembly.** A: Linear level -1 parts such as promoters and CDSs are first cloned into level 0 (L0) acceptor plasmids (PhytoBrick, MoClo or GoldenBraid acceptors). Level 0 (L0) standard parts are assembled into pOdd acceptors using BsaI Type IIS assembly to create Level 1 (L1) constructs, each consisting of a complete transcriptional unit. Level 1 constructs are then assembled together into pEven acceptors to create multigene constructs using SapI Type IIS assembly according to the Loop syntax³²⁰. Four parts can be combined into pOdd or pEven acceptors at each step. B: L0 plasmids are digested with BsaI, creating overlapping overhangs. These are then ligated together to create L1 plasmids in pOdd backbones. C: L1 plasmids are digested with SapI, creating overlapping overhangs. These are then ligated together to create multigene constructs in pEven backbones. Adapted from Pollak et al.³²⁰.

2.1.1 Design and construction of promoter parts

Arabidopsis promoters

Promoter regions of *ANR1* (AT2G14210), *ARF18* (AT3G61830), *ANAC032* (AT1G77450) (mutated BsaI site), *DREB26* (AT1G21910), *NIR1* (AT2G15620), *NLP6* (AT1G64530), *NLP7* (AT4G24020) and *ARF18* (AT3G61830) including ~1000 bp upstream of the TSS and the 5' UTR were cloned by PCR amplification of genomic DNA using primers that introduced restriction enzyme sites for assembly into pUAP1³⁰⁹, pUPD2³²¹ or pUAP4³²². PCR reactions used 0.02 µL Q5 polymerase (New England Biolabs (NEB) M0491), 1 × Q5 reaction buffer (NEB B9027S), 200 µM of each deoxynucleotide triphosphate (dNTP) (Thermo Scientific R0181), 0.5 µM of each primer and <1 ng template DNA, with 3 % v/v dimethyl sulphoxide (DMSO) if necessary. Reactions were cycled using the following conditions: 98 °C 30 s, followed by 30 cycles of 98 °C 10 s, 55–65 °C 20 s and 72 °C 20–30 s kb⁻¹. A final extension of 72 °C for 2 minutes was used.

Amplicons were assembled with pUAP1³⁰⁹ (Addgene 63674), pUPD2³²¹ (Addgene 68161) or pUAP4³²² (Addgene 136079) in a one-step digestion-ligation reaction described in section 2.1.6 to produce Level 0 standard parts (see File S1). To comply with the Phytobrick standard^{309,323}, internal instances of BsaI were removed by the introduction of single point mutations by PCR as in Patron et al.³⁰⁹. Two internal BsaI sites in the *ANAC032* (301 bp upstream of TSS) and *ARF18* (258 bp upstream of TSS) promoters were removed by the introduction of a mutation. Promoter constructs were verified by Sanger sequencing.

Synthetic promoters

A. thaliana TFBS consensus sequences were identified from the plant cistrome database (Release 1)¹⁸¹ (http://neomorph.salk.edu/dev/pages/shhuang/dap_web/pages/browse_table_aj.php). The sequences of TFBSs were added into the variable region of a previously published synthetic promoter consisting of a 19 bp random sequence followed by a 43 bp core promoter sequence from *CaMV35s*²⁷⁸. The sequences of synthetic promoters were synthesised by Integrated DNA Technologies (IDT) and cloned into pUAP1 in a one-step digestion-ligation reaction described in section 2.1.6 to create Level 0 standard parts³²⁴.

The identity and size of spacing between TFBSs was investigated using synthetic promoters.

2.1.2 Luciferase reporter constructs

Standard parts containing promoter sequences were used in assembly reactions with parts containing the CDSs of LucN (pEPYC0CM0133, Addgene #154595²⁷⁸), a C-terminal 3×FLAG® epitope tag (pICSL50007; Addgene #50308³²⁵), and 3' UTR and terminator sequences from *Agrobacterium tumefaciens* nopaline synthase (*AtuNOS*) (pICH41421, Addgene #50339³²⁵). Parts were assembled into Level 1 Loop pCk1³²² (Addgene #136695) backbones in a one-step digestion-ligation reaction described in section 2.1.6. As controls, equivalent plasmids were constructed with the promoter sequences from *AtuNOS* (pEPSW1KN0035), *A. thaliana ubiquitin 10* (*AtUBQ10*) (pEPSW1KN0071), *AtNIR1* (pEPSW1KN0025) and the previously reported *NRP*³²⁶ (pEPSW1KN0092). Calibrator plasmids (pEPSW1KN0034 and pEPSW1KN0072) for ratiometric quantification were assembled from *CaMV35s*:tobacco mosaic virus (Ω TMV)(pICH51277, Addgene #50268³²⁵) or *AtUBQ10* (pICSL12015, Addgene #117506³²⁷), respectively, firefly luciferase (LucF) (pEPAS0CM0008, Addgene #154594²⁷⁸), a C-terminal 3×FLAG® epitope tag (pICSL50007; Addgene #50308³²⁵), and a *AtuNOS* 3' UTR and terminator (pICH41414, Addgene #50337³²⁵).

Synthetic promoters fused to LucN::yellow fluorescent protein (YFP) reporters were co-assembled between FAST-Red^{325,328} and bar phosphinothricin³²⁵ selectable marker genes into Level 2 Loop pCsA backbones for use in transgene luciferase expression assays (see section 2.5.2).

2.1.3 Transcription factor parts

Arabidopsis transcription factors

The CDSs of *ANAC032* (AT1G77450), *ANR1* (AT2G14210), *ARF18* (AT3G61830), *ARF9* (AT4G23980), *NLP6* (AT1G64530), *NLP7* (AT4G24020), *DREB26* (AT1G21910), *TCP20* (AT3G27010), *TGA1* (AT5G65210), *basic leucine-zipper 3* (*bZIP3*) (AT5G15830) and *HRS1 homolog 2* (*HHO2*) (AT1G68670) were cloned, concurrently removing any internal instances of BsaI and SapI with point mutations to introduce synonymous codons with similar Arabidopsis codon usage, into pUAP1, pUPD2, pUAP4 or a reversed pUAP4

backbone (pUAP4r) in a one-step digestion-ligation reaction (section 2.1.6) to create Level 0 standard parts³²⁴ (see File S1). For transient plant co-expression assays these were assembled into Level 1 Loop pCk2³²².

Synthetic transcription factors

For synthetic transcriptional activators, the Cas9-SunTag system was used²⁹⁹ (fig. 2.3). This requires co-expression of (i) an endonuclease deficient Cas9 (pEPSW0CM0371) in translational fusion with NLSGCN4 (pEPSW0CM0370) fusion and (ii) a single chain variable fragment (scFv) fused to superfolder GFP (sfGFP) (a green fluorescent protein variant that folds robustly even when fused to poorly folded proteins³²⁹) (pEPSW0CM0365) and either the VP64 (pEPOZ1KN0250) or two repeats of the TAD³⁰¹ (pEPOZ1KN0252) transcriptional activator domains with a C-terminal simian virus 40 (SV40) nuclear localisation signal (NLS), the B1 domain of B1 domain of streptococcal protein G (GB1) solubility domain, and the human T-cell leukaemia virus type 1 Rex protein (Rex) NLS. The dCas9::NLSGCN4 and scFv::sfGFP::VP64/TAD fusion proteins were assembled into Level 1 transcriptional units with either an *AtUBQ10* promoter (pEPSW1KN0367 containing 1-pTwist_Amp_HighCopy_AtUBQ10p, gratefully received from Siobhan Brady, UC Davis), a synthetic *NRP* promoter³²⁶ (pEPSW1KN0415), a synthetic $4 \times [NRE-TATA]$ (pEPSW1KN0416), or a $4 \times [bZIP3-random]$ (pEPSW1KN0417) promoter. These transcriptional units were co-assembled with an upstream insulator sequence (*Petunia hybrida* transformation booster sequence³³⁰) and one or three sgRNAs based on the scaffold first reported by Chen and co-workers³³¹, each controlled by the *A. thaliana* U6-26 promoter (picsl90002, Addgene #68261³³²). In the final assembly, expression cassettes were flanked by synthetic genes to aid the selection of transgenic lines: an RFP (pICSL70008, Addgene #50336^{325,328}) and a synthetic gene for resistance to phosphinothricin (glufosinate-ammonium) consisting of a $2 \times 35s$ promoter and the ΩTMV 5' UTR (pICH51288, Addgene #50269³²⁵), the phosphinothricin acetyl transferase from *Streptomyces hygroscopicus* CDS (pICH43844, Addgene #50329³²⁵), and the *AtuNOS* terminator (pICH41421, Addgene #50339³²⁵) (see File S1).

For synthetic transcriptional repressors, the *Acidaminococcus* sp. BV3L6 Cpf1 (As-Cpf1) (now known as Cas12a) was used with the SRDX repressor domain²⁹⁴. The CDS of the endonuclease deficient CRISPR associated protein 12a (Cas12a) CDS

was assembled with an N-terminal SV40 NLS, a C-terminal nucleoplasmin NLS, and 3×SRDX. The Cas12a::NLS::3×SRDX fusion protein was assembled into Level 1 transcriptional units with a *AtUBQ10* promoter (pEPSW1KN0395), a synthetic *NRP* promoter³²⁶ (pEPSW1KN0421), a synthetic 4×[*NRE-TATA*] promoter (pEPSW1KN0422), or a 4×[*bZIP3random*] promoter (pEPSW1KN0423). These were co-assembled with one or three CRISPR RNA (crRNA) double hammerhead/hepatitis delta virus ribozyme cassettes and the same insulator and selection genes used in the Cas9-activator constructs. A different transcriptional repressor was also assembled which used the SunTag system²⁹⁹, with a dCas9::NLSGCN4 fusion and a scFv::sfGFP fusion containing 3×SRDX²⁹⁴ repressor domains (see File S1).

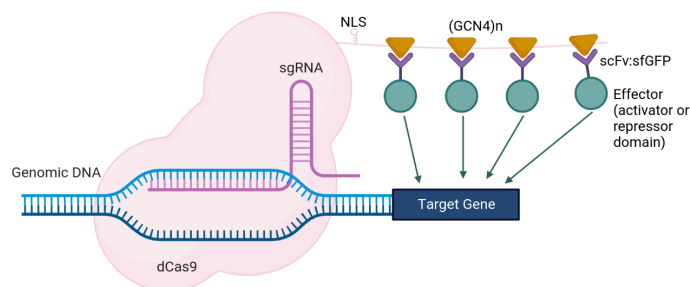


Figure 2.3: **The SunTag system.** A dCas9::NLSGCN4 fusion is co-expressed with a scFv::sfGFP::effector protein fusion. The dCas9::NLSGCN4 is guided to the promoter of the target gene by the sgRNA binding to the protospacer. The chain of GCN4 repeats attached to the dCas9 protein recruit multiple copies of the scFv::sfGFP::effector fusion protein to the promoter, where they activate or repress transcription. NLS, nuclear localisation signal; GCN4, yeast transcriptional activator peptide; scFv, single-chain variable fragment of the anti-GCN4 antibody; sfGFP, superfolder green fluorescent protein.

2.1.4 Transcription factor bacterial expression constructs

Protein expression plasmids were assembled using the Gateway®-ready *E. coli* expression pDEST vector, pH9GW (a modified pET28a(+)) vector (Novagen)³³³. First, TF-HiBiT fusions were created by PCR amplification of the TF CDSs using primers that introduced the Gateway® recombinase sequences attB1 and attB2. The Gateway®BP Clonase™II Enzyme Mix (Invitrogen 11789) was used to clone these amplicons into pDONR207 (attP1-ccdB-attP2) to create pENTR clones (attL1-TF::HiBiT-attL2). Gateway®LR Clonase™II Enzyme Mix was used to combine pENTR plasmids with pH9GW (attR1-ccdB-attR2) to produce *E. coli* expression clones (see File S1).

2.1.5 Transcription factor plant expression constructs

For constitutive overexpression of TFs in plant cells, Level 0 standard parts containing TF CDSs were assembled in a one-step digestion-ligation reaction (see section 2.1.6) into a Level 1 Loop pCk2³²² acceptor together with the $pCaMV35s-\Omega TMV$ promoter/5' UTR (pICH51277) and the *CaMV35s* terminator (pICH41414; Addgene #50337³²⁵). For dexamethasone-inducible expression in plant cells, constructs containing TF CDSs in a translational fusion with a C-terminal GR tag were kindly received from Tufan Oz (Earlham Institute). In these constructs, expression was controlled by a double *CaMV35s*- ΩTMV promoter/5' UTR (pICH51288; Addgene #50269³²⁵) and the *CaMV35s* terminator (pICH41414) (see File S1).

2.1.6 Digestion-ligation reaction

One-step digestion-ligation reactions were carried out, where DNA of interest was mixed with an acceptor vector, restriction enzyme and T4 ligase and incubated on a thermocycler. For cloning of Level 0 DNA parts into pUAP1, pUPD2 and pUAP4 acceptors, and cloning acceptor plasmids into pCk and pCs Loop³²² vectors, the following 10 μL one-step digestion-ligation reaction mixtures were used:

2 nM DNA of interest containing flanking BpiI, BsmBI, SapI or BsaI recognition sequences was used with 1 nM pUAP1, pUPD2, pUAP4, pCk or pCs acceptor vectors respectively depending on recognition sequence. For the pUAP1, pUPD2 and pCk vectors, assembly reactions also contained a $1 \times$ final concentration of T4 DNA ligase buffer (NEB B0202) (final concentration: 50 mM Tris-HCl, 10 mM MgCl₂, 1 mM adenosine triphosphate (ATP), 10 mM dithiothreitol (DTT)), 10 U μL^{-1} T4 DNA ligase (NEB M0202), 0.1 $\mu\text{g} \mu\text{L}^{-1}$ of bovine serum albumin (BSA) (NEB B9000) and 0.25 U μL^{-1} BpiI (NEB R3539), BsmBI (NEB R0739) or BsaI-HF@v2 (NEB R3733) was used.

For the pUAP4 and pCs vectors, a final concentration of $1 \times$ NEB CutSmart® buffer with added ATP (final concentration: 50 mM potassium acetate, 20 mM Tris-acetate, 10 mM magnesium acetate, 0.1 $\mu\text{g} \mu\text{L}^{-1}$ BSA, 1 mM ATP) was used. 10 U μL^{-1} of T4 DNA ligase (Thermo Scientific EL0011) was added, and 0.25 U μL^{-1} of SapI (Thermo Scientific ER1931) was used. Note: *ThermoFisher T4 ligase stock is 5 WeissU μL^{-1} so 1000 cohesive end unit (CEU) μL^{-1} , NEB T4 ligase stock is 400 CEU μL^{-1} . All enzyme units in this*

thesis are $CEU \mu L^{-1}$.

Distilled deionised water was added to reach a reaction volume of 10 μL .

Samples were incubated in a thermocycler with the following program: Assembly: 37 °C for 20 s followed by 26 cycles of 37 °C for three minutes and 16 °C for four minutes. Termination and enzyme denaturation: 50 °C for 5 minutes and 80 °C for 5 minutes.

2.1.7 Transformation of competent *E. coli*

Assembly reactions (0.5 μL) were incubated on ice with 3 μL 5-alpha competent *E. coli* (NEB C2987) (for plasmid production) or *E. coli* BL21(DE3) (NEB C2527) (for protein expression). After 30 minutes, cells were transformed by a brief incubation at 42 °C for 30 seconds. Tubes were placed on ice for 5 minutes and 20 μL lysogeny broth (LB) was added to each tube before incubation at 37 °C for 60 minutes with shaking (250 revolutions per minute; RPM) (New Brunswick Innova® 44 incubator shaker). Cells were spread on selective LB agar plates containing an appropriate antibiotic for the plasmid (35 $\mu g mL^{-1}$ chloramphenicol, 50 $\mu g mL^{-1}$ kanamycin, 50 $\mu g mL^{-1}$ spectinomycin or 50 $\mu g mL^{-1}$ carbenicillin). When acceptor plasmids contained LacZ inserts for blue/white cloning, X-gal (20 $\mu g mL^{-1}$ final concentration) and Isopropyl β -D-1-thiogalactopyranoside (IPTG) (0.1 mM final concentration) were added to the agar. Following overnight incubation at 37 °C, white colonies were selected for further analysis.

2.1.8 Colony PCR

Colony PCR was performed in a 10 μL reaction with final concentrations of 1 \times Promega Green GoTaq® Flexi buffer, 4 mM $MgCl_2$, 0.2 mM of each dNTP (Thermo Scientific R0181), 0.025 U μL^{-1} GoTaq®DNA polymerase, and 0.4 mM forward and reverse primers. Reactions were cycled with an initial incubation at 95 °C for 3 minutes, followed by 26 cycles of: 95 °C, 30 seconds then 60 °C, 30 seconds then 72 °C, 2 minutes.

2.1.9 Preparation and verification of plasmid DNA

Colonies with the expected insert sizes were used to inoculate 5 mL LB with appropriate antibiotics and incubated in a shaking incubator for 16–18 hours at 37 °C, 250 revolutions per minute (RPM). Plasmids were purified using the QIAprep Spin Miniprep Kit (QIAGEN

27106) with a vacuum manifold and the integrity and sequence of plasmids were checked by restriction digestion and Sanger sequencing (Eurofins or Genewiz). For restriction digestion analysis, 300 ng DNA was digested in a 15 μ L reaction volume with 1 \times CutSmart[®] Buffer (NEB B7204) and 0.66 μ L of the appropriate restriction enzyme. Samples were incubated at 37 $^{\circ}$ C for 1 hour and analysed by agarose gel electrophoresis. For protoplast transfection, 100 mL cultures were grown and plasmids were extracted and purified using the Plasmid Plus Midi Kit (QIAGEN 12943) with vacuum manifold. DNA concentration was quantified using a Thermo Scientific NanoDrop One Microvolume UV-vis Spectrophotometer.

2.1.10 Agarose gel

Electrophoresis of DNA samples was performed on 50 mL or 150 mL 0.8–2.2 % w/v agarose gels containing 0.5 \times SYBR[™] Safe DNA Gel Stain (Invitrogen S33102). Gels were made by mixing agarose (Melford A20080) with Tris/Borate/EDTA (TBE) (Severn Biotech 20-6000-50) buffer in a conical flask and microwaving for 1.5 minutes until dissolved. After cooling to 50 $^{\circ}$ C, 0.5 \times SYBR[™] Safe DNA Gel Stain (Invitrogen S33102) was added. Gels were cast and allowed to set before immersion in 0.5 \times TBE buffer. DNA samples were mixed with 1 \times loading dye (NEB B7025), and 0.5 μ L of 166.7 μ g mL⁻¹ DNA ladder mix (1 \times 1kb Plus DNA ladder (NEB N3200), 1 \times loading dye (NEB B7025)) was loaded into the first well of each gel. Gels were run at 80–150 V until the dye had penetrated \sim 75 % of the gel. Gels were visualised with ultraviolet light on a Bio-Rad Imager (ChemiDoc[™] MP or ChemiDoc[™] Touch).

2.2 Plant growth and transformation

2.2.1 Seed sterilisation

Seeds were sterilised in 70 % v/v ethanol for 2 minutes followed by sodium hypochlorite (Honeywell Fluka 71696) 1:2 dilution for 10 minutes. They were washed three times in sterile deionised water. Seeds were stratified in the dark at 4 $^{\circ}$ C for at least 24 hours before sowing.

2.2.2 Arabidopsis growth

For plant transformation and production of protoplasts, plants were grown in soil with an 18-hour day, 6-hour night photoperiod (long day) at 22 °C. For protoplast production, plants were transferred to 8 hour day, 16-hour night photoperiod for at least 3 days.

For assessment of molecular and physiological phenotypes, plants were grown on sterile growth media with an 18-hour day, 6-hour night photoperiod (long day) at 22 °C. Sterile growth media consisting of 4 mM $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ (Fisher M/1050/53), 2 mM KH_2PO_4 (Sigma-Aldrich P5655), 100 μM NaFe-ethylenediaminetetraacetic acid (EDTA) (Sigma-Aldrich EDFS), 1 mM $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (Sigma-Aldrich C7902), 10 mM KCl (Sigma-Aldrich P5405), 0.75 mM 2-(*N*-morpholino)ethanesulphonic acid (MES) (Sigma-Aldrich M8250), 29.21 mM sucrose (1%) (Fisher Scientific S/8600/60), and micronutrients (92.5 μM H_3BO_3 , 18.3 μM - $\text{MnCl}_2 \cdot 4\text{H}_2\text{O}$, 1.61 μM ZnCl_2 , 0.59 μM $\text{CuCl}_2 \cdot 2\text{H}_2\text{O}$ and 0.21 μM $\text{Na}_2\text{MoO}_4 \cdot 2\text{H}_2\text{O}$) was prepared and supplemented with 0, 1 or 10 mM KNO_3 (Sigma-Aldrich P8291) and solidified with 7.5 g/L Phytagel (Sigma-Aldrich P8169). Media was poured into 120 mm square culture plates (Greiner Bio-One 688161). Seeds were sieved to a size range of 250–300 μm before sterilisation and stratification. Plates were incubated in a vertical orientation with long day (18-hour days, 6-hour nights) cycles at 22 °C with a light intensity of approximately 70 $\mu\text{mol m}^{-2} \text{s}^{-1}$ in a Panasonic MLR-352-PE growth chamber. Plants were grown for 9 days following germination. For assessment of RSA, on one half of each plate, ~8 Columbia-0 (Col-0) seeds were sown and, on the other half, ~8 seeds from the genotype to be assessed. Three replicates of each plate were used. For root luciferase assays, 9 T2 seeds were sown per plate per parental line. For RSA phenotyping and root luciferase assays, seeds were distributed equally in a single row ~0.5 cm from the top edge of the plate. For RNA extraction, ~150 seeds from a single T3 parental line were distributed equally in two rows 0.5 cm and 6 cm from the top of each plate. Three biological replicates were used, each consisting of a pool of roots from ~50 seedlings, and each qRT-PCR was performed in duplicate.

2.2.3 Transformation of electrocompetent *A. tumefaciens*

Agrobacterium tumefaciens (*Agrobacterium radiobacter* subsp. *tumefaciens*³³⁴) strain GV3101 cells were thawed on ice. Sterile 2 mm cuvettes (GeneFlow E6-0060) and chamber

slides were placed on ice. 20 μL of bacteria and 0.9 μL of 110 $\text{ng } \mu\text{L}^{-1}$ DNA were added to each cuvette. Cuvettes were placed in the chamber slide on a Bio-Rad Gene Pulser Xcell and were electroporated with one pulse using the Agro program (2.5 kV). 500 μL of terrific broth³³⁵ was added to each cuvette and the contents were transferred to 1.5 mL microcentrifuge tubes. *Agrobacterium* was incubated for 2 hours at 28 °C, 220 RPM. Tubes were centrifuged at 10000 relative centrifugal force (RCF) for 1 minute and supernatant was removed leaving 40 μL with the cells. Cells were resuspended in the remaining media and 20 μL was transferred to flat-bottom 6 well plates (CytoOne CC7672-7506) containing LB agar with 50 $\mu\text{g mL}^{-1}$ spectinomycin, 50 $\mu\text{g mL}^{-1}$ rifampicin and 25 $\mu\text{g mL}^{-1}$ gentamicin selection. Glass beads were used to spread the plates and plates were incubated for 48–60 hours at 28 °C. Medium sized colonies were selected and used to inoculate 5 mL liquid LB with 50 $\mu\text{g mL}^{-1}$ spectinomycin, 50 $\mu\text{g mL}^{-1}$ rifampicin and 25 $\mu\text{g mL}^{-1}$ gentamicin selection at 28 °C with 220 RPM shaking. 20% glycerol stocks were taken and stored at –70 °C.

2.2.4 Arabidopsis transformation

Col-0 Arabidopsis plants were grown in soil with long days until flowering (~4 weeks old). *Agrobacterium* strains containing the plasmids of interest were grown overnight in 5 mL liquid LB with 50 $\mu\text{g mL}^{-1}$ spectinomycin, 50 $\mu\text{g mL}^{-1}$ rifampicin and 25 $\mu\text{g mL}^{-1}$ gentamicin selection at 28 °C with 220 RPM shaking. Cells were centrifuged at 3428 RCF for 10 mins and resuspended in fresh 5% w/v sucrose solution to OD_{600} ~0.8. For the CRISPR library, 96 *Agrobacterium* strains (each containing one of the 96 CRISPR constructs described in section 2.6.1) were mixed together equally (3 mL each). Silvet L-77 was added to 0.05% v/v and *Agrobacterium* was transferred to 9 × 30 mL spray bottles. Arabidopsis plants were placed inside an autoclave bag and the flowers were sprayed with the *Agrobacterium*. Autoclave bags were taped closed and placed inside black bin liners for 24 hours. Plants were grown with long days (18-hour days, 6-hour nights) until seed collection, where a paper seed bag was taped over the inflorescence of plants for collection.

2.2.5 Selection of transgenic seeds

T1 seeds were sown on media containing 10 $\mu\text{g mL}^{-1}$ glufosinate-ammonium (Merck 45520) selection. Alternatively, for constructs containing the FAST-Red selection cassette^{325,328},

they were first checked under a Zeiss Lumar V12 stereo microscope with red fluorescence (filter set 43 HE Cy3) to ensure they were RFP positive. Wild type Col-0 seeds were sown without kanamycin selection. After 2–3 weeks, seedlings were transferred to soil.

2.2.6 Protoplast preparation

Protoplast extraction was carried out as previously described³³⁶, with some alterations. Plant tissue was added to a petri dish and sliced into ~4 mm strips. Leaf strips were submerged in enzyme digestion solution (0.4 M mannitol, 20 mM KCl, 20 mM MES pH 5.6, 10 mM CaCl₂ and 0.1 % w/v BSA) containing 1.5 % w/v cellulase and 0.3 % w/v macerozyme for 4 hours at 22 °C with shaking. During filtering (75 µL filter) 2 mL W5 (2 mM MES (pH 5.7) containing 154 mM MgCl₂ and 5 mM KCl) solution (2 mM MES pH 5.6 containing 154 mM NaCl, 125 mM CaCl₂ and 5 mM KCl) was used to wash the protoplasts. After filtering, protoplasts were centrifuged at 200 RCF for 2 minutes. After leaving on ice to settle for 30 minutes, supernatant was removed and 1 mL of MMG (4 mM MES pH 5.6 containing 0.4 M mannitol and 15 mM MgCl₂) solution was added. Using a cut pipette tip, leaf protoplasts were loaded to a 15 mL centrifuge tube. This was centrifuged for 10 minutes at 90 RCF at 4 °C. The supernatant above the cloudy protoplast layer was removed. The protoplast layer was transferred to another 15 mL centrifuge tube on ice. Protoplasts were diluted to $\sim 2 \times 10^5$ mL with MMG solution on ice.

2.2.7 Protoplast transfection

700 µL of protoplast solution (10^5 mL⁻¹) was slowly added to 5 mL microcentrifuge tubes, using cut pipette tips to avoid damaging cells. 2500 fmol of plasmid DNA was added to the protoplast solution. 700 µL plus the volume of the plasmid solution used of PEG 4000 (Sigma 81240) solution was added to each tube, and the tubes were covered and inverted 20 times to mix. 1.2 mL of W5 solution was added, and tubes were inverted 20 times to mix. The tubes were centrifuged at 100 RCF for 2 minutes, and supernatant was removed. Leaf protoplasts were resuspended with 120 µL W5. 200 µL of 0.1 % w/v BSA (Sigma A4503) solution was added to wells in a spectrophotometer plate. After 2 minutes, BSA was removed, and protoplasts were added to wells in the plate. For leaf protoplasts the spectrophotometer plate was incubated overnight for 18 hours with light intensity of approximately $70 \mu\text{mol m}^{-2} \text{s}^{-1}$ at 22 °C.

2.2.8 Assessment of root system architecture

A photograph of each plate containing Col-0 and mutant plant lines (see section 2.2.2 for growth conditions) was taken on a black felt background next to a measuring ruler using a Panasonic DC-TZ90 on intelligent auto mode with 98 mm zoom. The distance from the camera lens to the plate was 48 cm. A custom Jupyter Notebook³³⁷ using Python code (available at https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/preprocess_images.ipynb) was used to read in raw .RW2 Panasonic image files using the RAWPY (v0.16.0) package. The PILLOW (v9.2.0) package was used to rotate 180 degrees, convert to greyscale, invert and convert to .png format. Primary and lateral roots were traced and labelled using the SMARTROOT plugin (v4.21)³³⁸ for (Fiji Is Just) ImageJ (v2.3.0)³³⁹. Some plants from line plntEPSWT30130-4 lacked chlorophyll pigment. Measurements from these plants were excluded from analyses. A custom Jupyter Notebook³³⁷ using Python code (available at https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/analyse_rsa.ipynb) was used to analyse the SMARTROOT data. First, the .csv SMARTROOT output files were concatenated into a single PANDAS (v1.1.5)³⁴⁰ DataFrame. A new DataFrame was created containing only primary roots. Columns were added for primary root length, lateral root number, lateral root length (which included both first and second order lateral roots), average lateral root length (primary root length / lateral root length), total root length (primary root length + lateral root length), lateral root density (lateral root number / primary root length) and percentage of lateral root length contributing to total root length (lateral root length / total root length). The PANDAS DataFrame was exported as a .csv file. For each plant line and variable, quantile-quantile (Q-Q) plots of residuals were generated using the PINGOUIN (v0.13.2)³⁴¹ package to check for normality. Levene's test was performed using the SCIPY (v0.5.2)³⁴² package to check whether variances were equal. Data were log-transformed and analysed using a type I two-way analysis of variance (ANOVA) which considered the interaction effects of genotype:nitrate concentration. If the genotype:nitrate concentration interaction was significant, nitrate concentrations were analysed separately using one-way ANOVA. If the genotype:nitrate concentration interaction was not significant, the model was refitted without the interaction term and a type II two-way ANOVA was used to test the main

effects using the `BIOINFOKIT` package (v2.0.8)³⁴³. If significant, Tukey’s HSD test was used to compare means³⁴⁴. Boxplots were generated using the `SEABORN` (v0.11.2)³⁴⁵ package and significance annotations were added using the `STATANNOTATIONS` package (v0.4.4; <https://github.com/trevismd/statannotations>).

2.3 Protein expression, purification and analysis

2.3.1 Recombinant protein expression in *E. coli*

E. coli BL21(DE3) (NEB C2527) starter cultures were grown in LB liquid in a shaking incubator (37 °C, 220 RPM) with 50 µg mL⁻¹ kanamycin selection overnight. A second culture was inoculated by diluting 40 µL of starter culture in 4 mL of liquid LB media. At 0.6–0.8 OD₆₀₀, cultures were moved to 18 °C for 5 minutes to acclimatise and IPTG was added to a final concentration of 0.5 mM and incubated overnight at 18 °C, 600 RPM. The next morning, the cultures were centrifuged for 10 minutes at 3428 RCF and pellets were stored at –70 °C.

2.3.2 Protein purification

Lysis wash buffer (10 mM imidazole, 1 mg mL⁻¹ lysozyme, 1 × His-tag protease inhibitor cocktail (Cephem 10475), 50 mM NaH₂PO₄, 0.05 % v/v TWEEN® 20 (Sigma P9416), 300 mM NaCl and 10 % v/v glycerol) was prepared and 1 mL was added to defrosted cell pellets. Pellets were resuspended and transferred to 1.5 mL microcentrifuge tubes and incubated for 30 minutes at 30 °C, 700 RPM. Tubes were snap frozen in liquid nitrogen before incubation at 37 °C for 15 minutes. 10 µL benzyl mix (8.5 mM MgCL₂ and 0.0075 U µL⁻¹ Benzonase® Nuclease (Sigma E1014)) was added to the tubes to digest DNA. Tubes were incubated for 2 minutes at 37 °C and centrifuged for 10 minutes at maximum speed at 4 °C. 50 µL of supernatant was collected for analysis in separate tubes. The remaining supernatant was transferred to new microcentrifuge tubes. INDIGO Ni-Agarose resin beads (PureCube Biotech 75103) were gently resuspended, added to a new microcentrifuge tube and kept on ice to settle. The bead solution (50 µL per sample) was diluted 6 × with equilibrium buffer (10 mM imidazole, 50 mM NaH₂PO₄) and mixed by inverting the tubes. The beads were centrifuged for 2 minutes, 700 RCF at 4 °C and the supernatant was removed. The beads were washed three times before being resuspended in equilibrium

buffer (1:1 ratio of beads to equilibrium buffer). 50 μL of resin bead solution was added to each lysate sample using cut pipette tips. Samples were incubated at 4 $^{\circ}\text{C}$ on a Sunflower Mini-shaker (Grant-bio PS-3D) for 45 minutes before centrifugation for 2 minutes at 700 RCF at 4 $^{\circ}\text{C}$. 50 μL of supernatant was stored for later analysis, and the remaining supernatant was discarded. Samples were resuspended in 125 μL wash buffer (20 mM imidazole, 50 mM NaH_2PO_4) and centrifuged for 2 minutes at 700 RCF at 4 $^{\circ}\text{C}$. They were washed 3 times before resuspension in 25 μL elution buffer (50 mM NaH_2PO_4 , 300 mM NaCl, 10 % v/v glycerol and 500 mM imidazole). Samples were stored on ice for 5 minutes and tapped occasionally to mix. Samples were centrifuged for 2 minutes at 700 RCF at 4 $^{\circ}\text{C}$. The supernatant elution was transferred to a new microcentrifuge tube and kept on ice. A second elution was performed on beads in the same way and that supernatant was mixed with the first elution. 5 μL of elution, 5 μL of lysate and 5 μL of flow-through were analysed by gel electrophoresis. The remaining samples were frozen at -70°C .

2.3.3 Protein quantification

Purified protein was quantified using the Nano-Glo[®] HiBiT Extracellular Detection System (Promega N2420) on a CLARIOstar Plus plate reader. The purified protein eluate was diluted 1000 \times in 1 \times passive lysis buffer (Promega E1941). To each well on a white 96-well plate (4titude 4ti-0273), 4 μL of diluted purified protein eluate was mixed with 36 μL of 1 \times passive lysis buffer (Promega E1941), 10 μL Nano-Glo[®] HiBiT Extracellular Buffer, 0.1 μL LgBiT protein and 0.2 μL Nano-Glo[®] HiBiT Extracellular Substrate. Seven different concentrations (0.01 mM, 0.1 mM, 0.25 mM, 0.5 mM, 0.75 mM, 1 mM, 2 mM) of HiBiT protein standards and a passive lysis buffer 'blank' were included in separate wells so that a standard curve could be calculated. The samples were measured after mixing on the plate reader using a luminescence kinetic program (setting time 0.2 s, 1 kinetic window, 30 cycles, measurement interval time 0.8 s, cycle time 40 s, time to normalise 0, number of multichromatics 1, emission no filter, well scan none, top optic). A timepoint was selected after quantification had plateaued and did not increase any more. Protein concentrations were quantified using the calibration curve.

2.3.4 SDS-PAGE

For sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE), 5 μ L protein sample was mixed with 5 μ L sample buffer ($2 \times$ Laemmli sample buffer (Bio-Rad 1610737) and 1.5 μ L 2-Mercaptoethanol (Sigma M3148)) and incubated at 95 °C for 10 minutes on a thermomixer (Eppendorf Thermomixer C) with no shaking. Samples were cooled to room temperature for 5 minutes and loaded on Mini-PROTEAN TGX Stain-Free Precast Gels (Bio-Rad 4568025) in $1 \times$ Tris/glycine/sodium dodecyl sulphate (SDS) buffer (Bio-Rad 1610732). 5 μ L protein ladder (Precision Plus Protein Dual Color 1610374) was loaded in the first well. 10 μ L of each sample was loaded into the other wells. The gel was run at 120 V for 1 hr 15 mins.

For visualisation of bands, gels were incubated in a square, 120 mm plate in 20 mL Instant-Blue® Coomassie Protein Stain (Abcam ISB1L) overnight with gentle shaking on a Labnet MiniLabRoller. The stained gel was imaged with white light on a Bio-Rad imager.

2.3.5 Western blot

Protein was transferred to membranes using a Bio-Rad Trans-blot Turbo transfer system. The Trans-blot Turbo pack (0.2 μ m nitrocellulose) was opened, and the anode stack and membrane were added to the Trans-blot Turbo transfer system. The gel was laid on top followed by the top cathode stack. Transfer was obtained using the mixed molecular weight program. The top cathode stack and gel were removed, and the blotting membrane transferred to a square 120 mm well plate with tris-buffered saline solution (TBST; 20 mM tris, 150 mM NaCl, 0.1 % v/v Tween® 20 detergent (Sigma P9416)). The Nano-Glo® HiBiT Blotting System (Promega N2410) blotting buffer was prepared ($1 \times$ Nano-Glo® HiBiT Blotting Reagent, $1 \times$ LgBiT Protein). TBST was removed from the blotting membrane after 10 minutes and the membrane was washed twice with distilled deionised water. The membrane was submerged in Nano-Glo® Blotting buffer overnight at 4 °C. The Nano-Glo® Luciferase Assay Substrate was diluted 500X into the LgBiT/buffer solution containing the blotting membrane, mixed by rocking and incubated for 5 minutes. The blotting membrane was imaged on a ChemiDoc™ MP using the chemiluminescence setting.

2.4 Transcription factor relative binding affinity

A plate-based method was used that quantifies the relative affinity of TFs for candidate binding sites (fig. 3.2). In this method, short dsDNA probes are bound to the plate. The plates are incubated with low quantities of recombinant TFs with nine amino acid HiBiT tags. After washing to remove unbound protein, the luciferase activity of the HiBiT tag is reconstituted by adding the large subunit and luminescence is quantified and reported relative to the background affinity for random dsDNA probes.

All probes were ~80 bp long and included a 48–50 bp central region including the TFBS of interest, flanked by random sequence (5' TAGCGAAGTACGATCCC and 3' GGCCATCACGCAGTA). Probes were made by obtaining complementary single stranded DNA (ssDNA) oligos (IDT). Lyophilised DNA was resuspended in distilled deionised water, diluted 10X and annealed by heating to 94 °C for 2 minutes and gradually cooled. Probes were diluted to 40 ng μL^{-1} . Positive control probes were designed based on TFBS consensus sequences from the DAP-seq O'Malley et al.¹⁸¹ database (release 1). Negative control probes consisted of ~80 bp of random sequence with roughly equal ATGC ratios.

In each well of a 96-well black F-bottom medium binding microplate (Greiner Bio-One 655076), 200 ng of DNA probe was mixed with 50 μL Reacti-bind DNA coating buffer (Thermo Scientific 17250). The plate was covered and incubated at room temperature overnight in the dark with gentle shaking (PS-3D, Grant-Bio 144204). The next morning, DNA coating buffer was discarded, and wells were washed three times with 200 μL phosphate-buffered saline (PBS) (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 , 1.8 mM KH_2PO_4) per well. PBS was removed and 270 μL PBS containing 3% v/v BSA (Sigma A4503) was added to each well. Plates were incubated at room temperature for 1 hour and the PBS-BSA buffer was discarded. Wells were washed with 200 μL PBS. Protein buffer was prepared containing 25 mM pH 8 Tris-HCl, 100 mM KCl, 2 mM DTT, 1 mM EDTA, 0.1 mg μL^{-1} BSA, 0.5 $\mu\text{g} \mu\text{L}^{-1}$ Poly(deoxyinosinic-deoxycytidylic), 5% v/v glycerol and 0.05% v/v IGEPAL® CA-630. 25–100 fmol of protein sample (5 μL) was mixed with 45 μL protein buffer in each well. Protein samples and buffer were excluded from some wells on the plate for use as calibrator wells to quantify DNA concentration. The plate was incubated for 3 hours with gentle shaking at room temperature. Protein buffer was discarded,

and wells were washed 3 times with 200 μL of transcription factor relative binding affinity measurement (TRAM) wash buffer (25 mM pH 8 Tris-HCl, 100 mM KCl, 10 % v/v glycerol, 0.05 % v/v IGEPAL[®] CA-630). Wash buffer was discarded, making sure to remove all bubbles. 50 μL of Nano-Glo[®] HiBiT Extracellular lysis mix (40 μL of 1 \times passive lysis buffer (Promega E1941), 10 μL Nano-Glo[®] HiBiT Extracellular Buffer, 0.1 μL LgBiT protein and 0.2 μL Nano-Glo[®] HiBiT Extracellular Substrate) was added to each well and the plate was incubated for 10 minutes in the dark at room temperature. Luminescence was quantified in a plate reader (top optic, 0.2 s setting time, 10 s measurement interval time, no filter, no well scan, number of multichromatics 1, time to normalise 0 s, shaking time 10 s). Auto focus with fixed gain of 3600 was used. DNA concentration was measured by adding 1 \times Quant-iT[™] PicoGreen[™] dsDNA reagent (Invitrogen P7581) to the DNA-only wells. After incubation at room temperature for 2 mins in the dark, the fluorescence was quantified using the PicoGreen[™] endpoint program (number of multichromatics 1, top optic, setting time 0.2 s, no. of flashes 40, PicoGreen[™] excitation 483-15, auto 502.8, emission 530-30, shaking time 10 s). Autofocus was used with a fixed gain of 1200. The HiBiT luminescence reading was normalised to the DNA concentration for each respective DNA probe.

2.5 Expression analysis

2.5.1 Ratiometric quantification of luciferase expression

To account for differences in transfection efficiency within and between experiments, a ratiometric system was used in which LucN luminescence of the test promoter (pTest) is normalised to a co-transfected experimental calibrator ($p_{CaMV35s}$ or p_{AtuNOS} -LucF), followed by a batch calibrator (p_{AtuNOS} -LucN/ $p_{CaMV35s}$ -LucF) (fig. 2.4). Normalised luminescence is calculated as follows:

$$\text{Normalised luminescence (a.u.)} = \frac{\text{Test}_{LucN}/\text{EC}_{LucF}}{\text{NOS}_{LucN}/\text{35s}_{LucF}}$$

where EC is the experimental calibrator and $\text{NOS}_{LucN}/\text{35s}_{LucF}$ is the batch calibrator (BC).

After protoplast transfection with plasmids of interest and calibrator (see section 2.2.7),

luciferase activity was measured using the Promega NanoGlo® Dual-Luciferase® Reporter Assay System (Promega E1910) according to the manufacturer’s instructions. During protoplast transfection, 1000 fmol of plasmids containing test promoters controlling LucN with a C-terminal 3×FLAG® epitope tag (pICSL50007) and the *AtuNOS* terminator (pICH41421) were added. For positive control wells, *pAtuNOS*-LucN (pEPSW1KN0035) or *pCaMV35s*-LucN (pEPSW1KN0070) were used, all with the C-terminal 3×FLAG® epitope tag and *AtuNOS* terminator. For TF-promoter co-expression assays, a total concentration of 1000 fmol of *pCaMV35s-ΩTMV-TF-CaMV35s_t* was used. When testing combinations of different TFs, plasmids were added at equimolar ratios totalling 1000 fmol. All samples were run in triplicate. 18 hours after protoplast transfection, protoplasts were removed from the illuminated incubator and centrifuged for 2 minutes at 200 RCF. A lysis buffer–protease inhibitor mix was made, containing 1 × passive lysis buffer (Promega E1941) and 1 × protease inhibitor cocktail (Sigma P9599). Supernatant was removed, and 30 µL of lysis buffer protease inhibitor mix was added to each well, pipetting up and down to mix. The plate was put on ice for 15 minutes and centrifuged at 3428 RCF for 10 minutes at 4 °C.

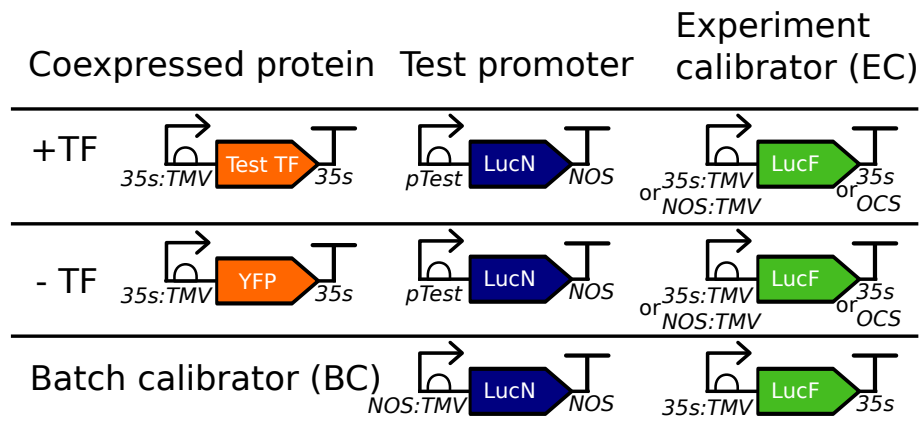


Figure 2.4: **Protoplast co-expression ratiometric luciferase assay.** To assess changes in luminescence resulting from the expression of each promoter tested (pTest) in response to the expression of a given transcription factor (TF), luminescence values relative to an experiment calibrator were compared from samples with and without plasmids expressing each TF. To maintain equal transcriptional loads, in the absence of the TF, a control plasmid expressing YFP from the same promoter was included. The relative value for pTEST was normalised to a batch calibrator to account for variation between protoplast batches providing a normalised value in arbitrary units (a.u.).

30 µL of lysed plant cells were added to each well on a 4titude (4ti-0273; now Azenta Life

Sciences) 96-well white polystyrene microplate solid bottom plate reader plate, with 3 technical replicates. The NanoGlo® Dual-Luciferase® Reporter Assay System (Promega N1610) was used. 30 µL of ONE-Glo™ EX Reagent was added to each well on the white plate, and the solution was mixed by pipetting three times.

Luminescence was measured using a CLARIOstar Plus plate reader. Focus was adjusted automatically, and each well was measured for ten seconds at 3600 gain, settling time one second. This measured firefly luciferase luminescence. NanoDLR™ Stop & Glo® Reagent was prepared by diluting the substrate 100X in the buffer, and 30 µL was added to each well. The plate was incubated in the dark at room temperature for five minutes, and the plate was read in the same way as before. This measured the NanoLuc® luciferase luminescence.

2.5.2 Quantification of transgene luciferase expression

Root tissue from 6–10 pooled plants per replicate was harvested in 1.5 mL microcentrifuge tubes and frozen immediately in liquid nitrogen. Frozen root tissue was homogenised using tungsten–carbide beads (3 mm, QIAGEN Ltd) and a TissueLyser II (QIAGEN) at frequency 20 Hz for 30 seconds. A lysis buffer–protease inhibitor mix was made, containing 1 × passive lysis buffer (Promega E1941) and 1 × protease inhibitor cocktail (Sigma P9599). 150 µL of 1 × lysis buffer was added to each sample, mixed and kept on ice. Tubes were centrifuged at 14000 RCF for 15 minutes at 4 °C and the supernatant was transferred to new microcentrifuge tubes. 30 µL of sample was added to wells in triplicate on a 96-well plate. The plate was incubated in the dark at room temperature for five minutes and LucN luminescence was measured as above (section 2.5.1).

Protein was quantified using Pierce™ 660 nm Protein Assay Reagent (Thermo Scientific™ 22660) in 96-well F-bottom clear microplates (Greiner Bio-One 655101). 10 µL sample was mixed with 150 µL reagent, incubated for 5 minutes in the dark, and absorbance was quantified on a plate reader with 660 nm wavelength, with 2000 gain.

LucN luminescence activity was normalised to protein absorbance at 660 nm.

2.5.3 Root RNA extraction

Roots from 12-day old plants were harvested at ~10:00 in the morning in 2 mL Eppendorf DNA LoBind® microcentrifuge tubes, two 3 mm tungsten-carbide beads were added, and samples were immediately frozen in liquid nitrogen. Three pooled root samples were taken (three biological replicates) from each square culture plate (Greiner Bio-One 688161). A blade and forceps were used to remove the roots from the plants. A new blade was used for each plate, and forceps were washed in 70 % v/v ethanol and 1:2 sodium hypochlorite (Honeywell Fluka 71696) solution followed by Blitz ribonuclease (RNase) Surface Decontaminant Removal Spray (Severn Biotech Ltd.) after each sample was collected. Frozen tissue was ground on a TissueLyser II (QIAGEN) at 20 Hz for 30 seconds until it was turned to powder. Samples were stored at -70°C until RNA extraction.

RNA was extracted using the Spectrum Plant Total RNA Kit (Sigma STRN250) manufacturer's protocol A. A DNase I digestion was performed after the RNA binding step using the RNase-Free DNase Set (QIAGEN 79254) to remove the DNA from the samples. 10 μL of DNase I and 70 μL of DNase digestion buffer were added to each column followed by a 15-minute incubation at room temperature. The columns were washed using 500 μL of wash solution 1. RNA was quantified on a NanoDrop One spectrophotometer (Thermo Scientific).

2.5.4 cDNA preparation

Complementary DNA (cDNA) was prepared by mixing $\sim 300\text{ ng } \mu\text{L}^{-1}$ of RNA with 0.5 μg of Oligo(dT)¹²⁻¹⁸ Primer (Invitrogen 18418012) and 0.83 mM each dNTP (Thermo Scientific R0181) in a 12 μL reaction volume. The reaction was incubated at 65°C for 5 minutes and quickly chilled on ice. 1.05 \times First Strand Buffer (Invitrogen 28025013), 10.5 mM DTT (Invitrogen 28025013) and 40 U RNaseOUT™ Recombinant Ribonuclease Inhibitor (Invitrogen 10777019) were added, making a total volume of 19 μL . The reaction mixture was incubated at 37°C for 2 minutes. 200 U M-MLV Reverse Transcriptase (Invitrogen 28025013) was added. The reaction was incubated at 37°C for 50 minutes followed by inactivation at 70°C for 15 minutes. cDNA was diluted 10X with half Tris-EDTA (TE) buffer (5 mM tris(hydroxymethyl)aminomethane (Tris), 0.5 mM EDTA, pH 8) and stored at -20°C .

2.5.5 Quantitative real-time PCR

Primers that annealed to *ARF9*, *ARF18*, *DREB26*, *NLP7*, *NIR1* and *elongation factor 1 alpha (EF1 α)* genes were obtained from Dr. Tufan Oz, Earlham Institute (see File S2). QRT-PCR was performed using a QuantStudio™ 6 Pro Real-Time PCR System (Applied Biosystems A43182) using 384-well plates. Amplifications were performed in 10 μ L reactions containing 1 \times SYBR® Green JumpStart™ Taq ReadyMix™ (Sigma S4438), 0.2 μ M each primer and 6 ng cDNA template. No template and no reverse-transcriptase controls (pooled RNA from all samples) were included in each run. Analysis was done for three biological replicates and two or three technical replicates. A QuantStudio™ 6 Pro 384-well standard, relative quantification with melt program (comparative Ct with melt) was used with the following parameters: 94 °C for 2 minutes then 40 cycles of 94 °C for 15 seconds, 94 °C for 1 minute. A melt curve was performed at the end of the run to confirm the specificity of the amplification.

For analysis, a normalized reporter value (ΔR_n) value of 19310 was used in ThermoFisher Design and analysis software. A .csv file was exported containing the raw data. A custom Jupyter Notebook³³⁷ containing Python code (https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/qPCR_06.09.22.ipynb) was used to calculate the relative expression of each gene in each sample. The .csv files were read in as PANDAS (v1.1.5)³⁴⁰ DataFrames, samples which did not amplify were filtered out, and samples with a quantification cycle (Cq) of >40 were filtered out. Outliers were removed as in Maussion et al.³⁴⁶ by calculating Cq standard deviation (SD) of the technical replicates of a given sample and removing the replicate furthest from the sample mean when the SD was greater than a cut-off of 0.3. If the Cq SD was higher than 0.3 but the absolute (mean-median)/median was less than 0.1, the replicates were not removed. This was useful in cases where there was no clear outlier. A tab-separated values (TSV) file was saved containing marked outliers. For expression analysis, data were normalised using the $\Delta\Delta C_q$ method, first to the EF1 α housekeeping gene Cq mean (ΔC_q). For analysis of plants grown in different nitrate concentrations, expression relative to samples grown on 1 mM KNO₃ Cq mean ($\Delta\Delta C_q$) was also calculated. Additionally, samples for each gene were normalised to the 1 mM KNO₃ Col-0 biological replicate Cq mean for that gene. The inverse log transformation was calculated. Normality of data

was calculated using the Shapiro–Wilk normality test³⁴⁷, and Levene’s test was used to test for homogeneity of variance. When variances were equal, independent t-tests were used for significance testing³⁴⁸. When variances were not equal, Welch’s t-tests were used³⁴⁹. Barplots were generated using the SEABORN (v0.11.2)³⁴⁵ package and significance annotations were added using the STATANNOTATIONS package (v0.4.4; <https://github.com/trevismd/statannotations>).

2.6 CRISPR library

2.6.1 Design and assembly of constructs for a targeted CRISPR library

The *cis*-regulatory regions of four TF genes, *ARF9* (AT4G23980), *ARF18* (AT3G61830), *DREB26* (AT1G21910) and *NLP7* (AT4G24020), were analysed to identify PAM sites recognised by SpCas9 in regions of open chromatin. To identify open chromatin, ATAC-seq data³⁵⁰ (see section 2.7.7 Open chromatin coverage) were superimposed over the non-coding regions upstream of the ATG start codon of each gene. A region of ~600–1000 bp falling within open chromatin was identified and SpCas9 NGG PAM sites were identified using CRISPOR³⁵¹ (the Arabidopsis information resource (TAIR) 10¹⁵⁶ genome assembly, Ensembl Plants³⁵² version 76). Using a custom Jupyter Notebook³³⁷ using Python³⁵³ code (available on GitHub: https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/Choose_guides.ipynb), protospacer sequences with on-target efficiency³⁵⁴ of <40 % were filtered out and 30 protospacers per gene were selected. Protospacers were organised into pairs roughly 90–110 bp apart, with each protospacer being allocated to at least one pair. Virtual cloning of protospacers to the scaffold first used by Chen and co-workers³³¹ were created using PyDNA (v3.0.2³⁵⁵) and DNA Cauldron (v2.0.1; <https://github.com/Edinburgh-Genome-Foundry/DnaCauldron>). From this assembly plan, 122 guide scaffolds were produced by PCR amplification of a sgRNA template (pSLQ1661-sgMUC4-E3; Addgene #51025³³¹) using primers to introduce the spacer sequence. These amplicons were used to produce 122 sgRNA expression cassettes by co-assembly with the *AtU6-26* promoter (picsl90002, Addgene #68261³³²). Finally, these were used to assemble 96 constructs for plant transformation, each containing a kanamycin resistance cassette (pEPSW1KN0114; p*AtuNOS*- Ω *TMV*-neomycin phosphotransferase II (NPTII)-*Agrobacterium tumefaciens* octopine synthase (*OCS*)_t), a Cas9 expression cassette

(pEPSW1KN0333; YAO AT4G05410 promoter, Cas9 with potato intron IV2 as described in Castel et al.³²⁷) and a pair of sgRNA expression cassettes.

2.6.2 Growth and selection of a population of edited plants

100 Col-0 plants were transformed with the CRISPR library plasmids. T1 seeds were collected and ~3000 were sown on 0.8 % w/v LB 3 % w/v sucrose agar containing Murashige and Skoog (MS) nutrients with 100 $\mu\text{g mL}^{-1}$ kanamycin selection. At the same time, wild type Col-0 seeds were sown to provide seeds of equivalent generations, grown in the same conditions as controls. After 2–3 weeks, 200 T1 transgenic seedlings were transferred to soil and leaf samples were frozen on dry ice and stored at -70°C . 50 T1 parental lines were selected randomly and 10 T2 seeds from each of these 50 lines were sown on 0.8 % LB 3 % sucrose agar containing MS nutrients with no selection. Leaf samples were frozen in Phire buffer (Thermo Scientific F160) on dry ice and stored at -70°C . Plants were transplanted to soil and grown in a controlled environment room at 22°C with long day cycles (18-hour days, 6-hour nights) and T3 seeds collected.

2.6.3 Polymerase chain reaction for amplicon genotyping

Samples were forcefully crushed with a pipette tip. After a 3-minute incubation at room temperature, samples were centrifuged at 3428 RCF for 30 s. The supernatant was transferred to a well in a new plate and diluted $5 \times$ in distilled deionised water and used as a template in a 20 μL PCR reaction containing 0.25–0.5 μL Phire extract, 0.5 μM each primer, $1 \times$ Phire master mix and, if necessary, 3 % v/v DMSO. The following cycles and conditions were used: 98°C 5 minutes, then 40 cycles of 98°C 5 s, $55\text{--}63^\circ\text{C}$ 5 s and 72°C 20 s kb^{-1} .

2.6.4 Generation of amplicon sequencing library

Leaf samples were taken from T2 CRISPR plants, were extracted in Phire buffer (Thermo Scientific F160) and PCRs were performed as above section 2.6.3 using primers specific to the promoter target of interest (*ARF9*, *ARF18*, *DREB26* or *NLP7*). The forward primer added a 5' 'TCCTCTGTCACGGAAGCG' sequence, and the reverse primer added a 3' 'TTTAGCCTCC-CCACCGAC' sequence to each amplicon. These sequences were specific to a second set of primers used in a final PCR reaction. PCR products were

checked by agarose gel electrophoresis (see section 2.1.10) to make sure no unwanted amplicon sizes were present. When yield was low, a nested PCR was performed using the first PCR reaction with Q5 polymerase as a template with primers internal to the first primer pair. This improved the specificity of the reaction. 96 unique forward and 96 unique reverse barcodes were designed using this webpage: https://www.bioinformatics.org/sms2/random_dna.html. This allowed for a unique pair for each well of a 96-well plate. A final PCR reaction with Q5 polymerase using primers specific to the flanking sequences on the product from the first or nested PCR reactions was performed. This final PCR added the unique forward and reverse 8 bp DNA barcodes to the amplicons, allowing them to be pooled for multiplex amplicon sequencing. 6 amplicon libraries were generated by pooling 200 fmol of each sample. Each library contained 384 pooled samples apart from library 6 which contained 104 samples. Each barcode pair was multiplexed four times in each library, one time for each target amplicon. For example, well A1 in each of the four 96-well plates in a library contained the same unique forward and reverse barcodes surrounding an amplicon from *ARF9*, *ARF18*, *DREB26* or *NLP7*. 120 μL of each library was purified using 90 μL of AMPure XP magnetic beads (Beckman A63880) (0.75 \times ratio beads to DNA) in 1.5 mL microcentrifuge tubes. Sample-bead mixture was mixed and incubated for 5 minutes. Beads were pulled down using a magnet for 2 minutes to separate the beads from the solution. The cleared solution was discarded. 200 μL of 70 % v/v ethanol was added and samples were incubated for 30 seconds at room temperature. Beads were pulled down using the magnet and ethanol was discarded. The ethanol step was repeated. Samples were dried at room temperature for 3 minutes. 80 μL of pH 8 10 mM tris-HCl was mixed with the beads followed by a 5-minute incubation at room temperature. Beads were pulled down using the magnet and the supernatant was transferred to a new tube. Sample concentrations were quantified using a NanoDrop One spectrophotometer (Thermo Scientific). 1 μg of DNA per library with minimum concentration of 20 $\text{ng } \mu\text{L}^{-1}$ was provided for PacBio Sequel II single molecule real time (SMRT) sequencing (8M, v2, 15-hour movie) (Genomics Pipelines, Earlham Institute).

2.6.5 Demultiplexing of samples

4049544 raw reads of mean length 39721 bp were obtained. Genomics Core Bioinformatics Group (Earlham Institute) performed circular consensus sequencing (CCS) analysis using

the PacBio SMRT Link (10.2.0.133434) pipeline. This resulted in 2313978 high fidelity (HiFi) reads with yield of 2907046832 bp with median read quality of Q44. During CCS base-calling, a phred-like algorithm assigns a quality score, Q , to each base in the read, and the quality score is used to determine whether a base is accurate or not, defined by the following equation³⁵⁶:

$$Q = -10 \log_{10} P$$

where P is the probability of the base being incorrect. A quality score of 20 (Q20) indicates an error rate of 1 in 100, while a score of Q30 means an error rate of 1 in 1000, with a read accuracy of 99.9%. The PacBio barcodes were demultiplexed using the CCS demultiplex barcode application. A `.fasta` file was generated containing the HiFi reads from each library. The `.fasta` file was aligned against the four reference genes (promoter and 5' UTR region) to demultiplex by gene. This was done using `exonerate` (v2.2.0)³⁵⁷ with the `affine:local` model with an 80% maximal query score. This generated an output file containing a list of reads assigned to a single gene once, and a file containing a list of reads assigned to multiple genes. Reads assigned to single genes were demultiplexed using the 96 forward and reverse barcodes using `Cutadapt` (v3.2.0) with a reference file containing a list of adaptor/barcode pairs (the common adaptor sequence from the first amplicon PCR along with the second unique 8 bp barcodes)³⁵⁸. Error rate was set to 0, and a minimum overlap of 24 bp was used in the forward direction, and a minimum overlap of 25 bp was used in the reverse direction, allowing 1 or 2 bp of the 8 bp barcode to be missing. Neither barcode was anchored since some reads were missing part of the barcode. The adaptor sequence and barcodes were removed (`--action trim`), no insertions or deletions were allowed (`--no-indels`) and both the read and its reverse complement were checked for adaptor matches (`--revcomp`). After demultiplexing by gene and barcode, there was a `.fasta` file for each amplicon (*ARF9*, *ARF18*, *DREB26* and *NLP7*) from each plant line (500 plants lines). Dr. David Swarbreck (Earlham Institute) performed the demultiplexing of samples.

2.6.6 Variant calling

The `.fasta` files (after demultiplexing) were converted to `.fastq` files using a custom Jupyter Notebook³³⁷ using Python code (https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/fasta2fastq.ipynb) involving `SEQIO` from `BIOPYTHON` (v1.79)³⁵⁹. A fixed quality score of 40 was used. `CRISPRessoBatch` from `CRISPResso2` (v2.2.7)³⁶⁰ was run on the `.fastq` files for each amplicon using manually created `.batch` files containing lists of `.fastq` files for each amplicon. Several amplicon sequences of different lengths were provided for each gene since slightly different initial primers were used depending on the plant line due to troublesome samples. The `--skip_failed` flag and a window size of 7 (`--quantification_window_size 7`) was used. Guide names and sequences were provided with the `--guide_seq` and `--guide_name` flags. The `--suppress_plots` and `--write_detailed_allele_table` flags were used. `CRISPRessoAggregate` was performed to aggregate the runs.

2.6.7 Mutation analysis

A reference `.fasta` file (`genes_longest_region.fasta`) was created containing the longest amplicon region for each gene. Using an Arabidopsis genome (TAIR 10) `.fasta` file as the index, `bowtie2` (v2.4.5) was run with the `genes_longest_region.fasta` input `.fasta` file, to generate a sequence alignment map (SAM) file (`genes_longest_region.sam`)³⁶¹.

The `.sam` file was converted to a browser extensible data (BED) file using `BEDTools bamToBed` (v2.30.0)³⁶². A custom Python script was run to process the `CRISPResso2` post-analysis (https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/pacbio_analyse_variantcall.py). This script identified the promoter genomic positions along with the TSS using the `genes_longest_region.bed` file and a promoter `.bed` file containing all promoters in the Arabidopsis genome. The alleles frequency text files from the `CRISPResso2` output along with a `.fasta` file containing the promoter regions of interest were used to check each CRISPR protospacer location for mutations. For each amplicon from each plant line, the `CRISPResso2` output directory was scanned for the alleles frequency text file. The 40 bp aligned sequence at a CRISPR protospacer site was compared to the reference sequence to determine if the site was mutated. If a difference was found between the sequences, all insertions, deletions, and substitutions

were recorded along with the relative mutation positions in the 40 bp protospacer window. The CRISPR dsDNA cut site relative locations in the whole promoter and whole genome were recorded. The distance from a mutation to the cut site was recorded, along with the distance of the mutation to the Araport 11 TSS³⁶³. Read number and percentage of total reads were recorded. A plant identity table (`plant_IDs.tsv`) was provided to label plant lines based on their primer pairs, library number and gene target. Some protospacer locations were in multiple allele frequency tables due to other nearby protospacer sites. These duplicate values were merged. Duplicates within a single allele frequency table were also merged, taking the sum of the read numbers and read percentages for duplicates. An output `.tsv` file was created containing the information described above.

Since the output file was large, it was split into smaller files for further analysis. The headers from the `.tsv` file were removed using the `tail BASH GNU coreutils (v8.32)` command followed by the `split` command which split the `.tsv` file into separate files, each 100 lines long.

A custom Python script (https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/pacbio_analyse_overlapping_TFBSs_part_1.py) which added TFBSs overlapping the mutations was run on each 100-line file. A mapped motifs `.bed` file was provided containing TFBS motifs in all Arabidopsis promoters and 5'UTRs scanned using FIMO³⁶⁴ (v5.1.1) (see section 2.7.2). Mutation genomic positions were extracted from the `.tsv` files, and `PYBEDTOOLS (v0.9.0)` `intersect` was used to find out which TFBSs overlapped mutations. Unique overlapping TFBS families and TFBS AGI (Arabidopsis Genome Initiative) codes were recorded and saved to a `.part` file.

The output `.part` files were concatenated together into one `.tsv` file using the `cat BASH` command. All `.split` and `.part` files were removed.

Another custom Python script (https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/pacbio_analyse_overlapping_TFBSs_part_2.py) was run on the output `.tsv` file for each target gene. This script genotyped the plant lines based on the mutations present. Mutations were classed as homozygous if 70% of the reads or more showed only one mutation at a locus. Mutations accounting for 10% or less of the reads were ignored. Mutations accounting for between 10 and 70% of reads were classified as heterozygous. If there were two different mutations at a locus

and there were no wild type reads, the locus was classified as biallelic. If the number of different mutations at a specific locus was more than 2 then the plant line was classified as chimeric. For each gene, an output `.tsv` file (`'_TFBSoverlapping_genotyped.tsv'` suffix) was created containing the genotyped plant lines and their mutations. The identity of the guide pairs on a single CRISPR plasmid was compared to the mutated guide sites in each plant line to determine if the mutations were a result of a single CRISPR plasmid or multiple different CRISPR plasmids. The number of different *Agrobacterium* strains delivered into each plant line was estimated. Another output `.tsv` (`'_TFBSoverlapping_genotyped_only_mutated.tsv'` suffix) was created containing only plant lines which were not wild type.

A custom Python script (https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/pacbio_analyse_flattened_mutations.py) was run to combine adjacent mutations and determine their length. As some mutations at the same locus were represented more than once in `'_TFBSoverlapping_genotyped.tsv'` files, duplicate mutations in each plant line were removed. An output `.tsv` file (`'_TFBSoverlapping_genotyped_only_mutated_flattened.tsv'` suffix) was created for each gene target containing a single row per plant line with a list of mutations, their genomic positions, their genotypes and which, if any, TFBSs overlapped them. ARF18 guide 14 locus was filtered out due to the region being very repetitive leading to false chimeric assignment.

2.7 *Cis*-regulatory module analysis

All data analyses and plotting were done using Python 3.7.8³⁵³. The Shapiro–Wilk normality test³⁴⁷ and Levene’s homogeneity of variance³⁴⁸ were used to test assumptions for parametric tests. For non-parametric analyses, the Kruskal–Wallis H test³⁶⁵ was used to test for differences between promoter categories using the `SCIKIT-POSTHOCS` package (v0.6.4)³⁶⁶. If necessary, Dunn’s post hoc tests³⁶⁷ were used with Bonferroni adjustment for multiple comparisons using `PINGOUIN.KRUSKAL` (v0.3.7)³⁴². All plots were created using `SEABORN` (v0.10.1)³⁴⁵.

Significance annotations were added to plots using `STATANNOT` (v0.2.3; <https://github.com/webermarcolivier/statannot>).

2.7.1 Extraction of *cis*-regulatory modules

Promoters were extracted from the Arabidopsis TAIR 10¹⁵⁶ genome assembly and Ensembl Plants³⁵² annotation (.gff3 release 47 date 08/03/2020) using a custom Python script (`extract_promoter.py`, available at <https://github.com/samwitham/PromoterArchitecture>). Only promoters from genes on Arabidopsis chromosomes 1–5 were extracted. Using the same Python script, promoters were extracted from protein coding genes that did not overlap other protein coding genes using PYBEDTOOLS (v0.8.1)³⁶⁸. Promoters were extracted 1000 bp upstream of the longest annotated transcript TSS or until the nearest annotated protein coding gene, and 5' UTRs were included downstream of the TSS up until the closest annotated CDS ATG start codon. Genes where the whole promoter overlapped a protein coding gene, leaving only part of the 5' UTR non-overlapping, were flagged and filtered out. In chapter 3, a 2000 bp upstream region was used instead of 1000 bp to extract promoters.

2.7.2 Transcription factor binding site identification

The resulting promoter annotations were transformed to .bed format using BEDOPS (v2.4.39) `gff2bed`³⁶⁹, and BEDTools (v2.29.2) `getfasta`³⁶² was used to extract promoter sequences from the reference genome. Promoters were scanned for DAP-seq TFBS motifs¹⁸¹ using FIMO³⁶⁴ (v5.1.1) with a zero-order background model created using `fasta-get-markov`³⁷⁰. A *p*-value threshold of $\sim 10^{-4}$ and max stored sequences ~ 5000000 was used, and the output was filtered using a Benjamini and Hochberg³⁷¹ *q*-value threshold of 0.05. Arabidopsis gene IDs for promoters and the TFs binding them were recovered for further analysis.

2.7.3 Gene selection

To investigate the stability of gene expression, Czechowski et al.³⁷² analysed microarray data from *A. thaliana* Col-0 across 79 different tissues, organs, and developmental stages. These data enable genes to be ranked according to stability of expression across tissues using CV values. The Tau tissue-specificity³⁷³ was calculated for each gene in the Czechowski et al.³⁷² dataset and the 100 or 300 most tissue-specific and 100 or 300 most non-specific

genes were selected. The Tau (τ) index is defined as^{373,374}:

$$\tau = \frac{\sum_{i=1}^n (1 - \hat{x}_i)}{n - 1}; \hat{x}_i = \frac{x_i}{\max_{1 \leq i \leq n} (x_i)}$$

where n is the number of tissues, and \hat{x}_i is the expression level of the gene in tissue i , normalized by the maximal value.

Only genes which had at least one TFBS found in their promoters using FIMO (see section 2.7.2) were ranked according to CV and Tau.

Recreating the methodology used by Czechowski et al.³⁷², only genes expressed in 80% of conditions and developmental stages based on mas5 calls were included when calculating CV. This step filtered out 10764 genes leaving 12046 genes. CV was used to select the 100 or 300 most constitutively expressed genes from raw expression data generated in their study. Alongside this, the 100 or 300 most variably expressed genes were selected. 100 or 300 control genes were randomly selected from the central distribution of the expression CV ranked genes using the Tau ranking. This ensured that the selected genes were present in both CV and Tau datasets. To ensure even sampling across the distribution of CV/Tau values, 10 or 30 genes were selected randomly from each of 10 bins covering the range of Tau values between the non-specific and tissue-specific gene sets.

2.7.4 Sliding window creation

Promoters were split into 100 bp sliding windows with a 50 bp step size using a custom Python script ([rolling_window.py](https://github.com/samwitham/PromoterArchitecture), available at <https://github.com/samwitham/PromoterArchitecture>). Windows with fewer than 100 promoters extending to that location were removed. For each sliding window plot, the median percentage GC content, TFBS coverage, open chromatin coverage or TF diversity was calculated. 95% confidence intervals were estimated using 10000 bootstrap iterations using SEABORN (v0.10.1).

2.7.5 GC content

Percentage GC content of promoters and each promoter window was determined using Python (https://github.com/samwitham/PromoterArchitecture/blob/master/src/data_sorting/promoter_GC_content.py) to test the hypothesis that constitutive genes have a higher GC content than variable genes and the hypothesis that non-specific genes

have a higher GC content than tissue-specific genes.

2.7.6 Transcription factor binding site coverage

To test the hypotheses that the CRMs of variable genes will have a lower percentage of base pairs covered by at least one TFBS than CRMs of constitutive genes, and that tissue-specific CRMs will have a lower percentage of base pairs covered than non-specific CRMs, the BEDTools (v2.29.2) `coverage` tool was utilised³⁶². The number of base pairs covered by at least one motif in a given sequence was calculated using the BEDTools `coverage` tool.

2.7.7 Open chromatin coverage

Negative control (treated with 5 μ M NaOH) ATAC-seq data were downloaded from Potter et al.³⁵⁰ for root and shoot tissues. Individual `.bed` files for each replicate were concatenated and BEDTools (v2.29.2) `merge`³⁶² was used to combine overlapping peaks. An intersection for root and shoot open chromatin was created using BEDTools `intersect`³⁶². To test the hypothesis that the CRMs of variable genes will have a lower proportion of open chromatin than constitutive CRMs, and that non-specific CRMs will have a lower proportion of open chromatin than tissue-specific CRMs, BedTools `coverage` tool³⁶² was used. The number of base pairs covered by root, shoot or the root-shoot intersect open chromatin was calculated using BedTools `coverage`.

2.7.8 TF diversity

The unique TF count for each promoter and promoter window was calculated. This meant that if TFBSs for a TF were found several times within a promoter, that TF was only counted once. TFs were only classed as present in a promoter window if the centre of the TFBS was inside the window. To test the hypothesis that constitutive CRMs will have a more diverse TFBS profile than variable CRMs, and that non-specific CRMs will have a more diverse TFBS profile than tissue-specific CRMs, the Shannon diversity was calculated. The mapped motif annotations were analysed using the `SKBIO.DIVERSITY.ALPHA.SHANNON` Python module (v0.5.6; <https://github.com/biocore/scikit-bio>) to calculate the Shannon diversity of individual TFs and also TF families binding each promoter or promoter window. The Shannon diversity, unique TFBS counts and raw TFBS counts were

analysed comparing constitutively expressed promoters to variably expressed promoters.

As documented in [TF_diversity_plots_wholeprom.ipynb](https://github.com/samwitham/PromoterArchitecture) (available at <https://github.com/samwitham/PromoterArchitecture>), a table was created containing each promoter on a different row with each TF family in a different column. The numbers in each cell represent the number of times TFs belonging to a particular TF family are predicted to bind to a certain promoter. A principal component analysis was run where 95 % of the variation was contained within with 22 components. Hierarchical clustering was used (Python code from <http://www.nxn.se/valent/extract-cluster-elements-by-color-in-python>) to estimate the number of clusters, K, to be used in Kmeans clustering. The number of clusters was predicted using the silhouette method³⁷⁵ and used as K in Kmeans clustering using the `SKLEARN.CLUSTER.KMEANS` Python module (<https://github.com/scikit-learn/scikit-learn/tree/master/sklearn/cluster>).

2.7.9 TATA box enrichment

15 bp TATA box locations were downloaded from Eukaryotic Promoter Database (EPD) (release: At_EPDnew_004)³⁷⁶ present between -50 to 0 relative to the EPD TSS. Genomic Association Tester (GAT)³⁷⁷ (v1.3.6) was used to compare enrichment of TATA boxes in constitutive and variable genes. This enabled testing of the hypotheses firstly that variable genes are enriched for TATA boxes over constitutive genes, and secondly that tissue-specific genes are enriched for TATA boxes over non-specific genes. The 15 bp TATA boxes were used as segments of interest. Constitutive and variable promoter annotations were separately tested for enrichment of TATA boxes compared to the background workspace file containing all 200 promoters of interest from both categories. Non-specific and tissue-specific promoter categories were tested in the same way.

2.7.10 GO-term analysis

A background gene set of all protein coding genes remaining after the filtering steps mentioned in section 2.7.1 (Extraction of *cis*-regulatory modules) was used so that all promoters contained at least one TFBS when scanned using FIMO³⁶⁴. Gene ontology (GO) terms (go-basic.obo; 2020-08-11 release) were downloaded from the Gene Ontology Consortium site (<http://current.geneontology.org/products/pages/downloads.html>). GOATOOLS³⁷⁸ (v1.0.6) was used for GO enrichment analysis using Fisher's

exact test³⁷⁹ and Benjamini/Hochberg false discovery rate (FDR) correction³⁷¹. The CLUSTERPROFILER³⁸⁰ (v3.14.0) R³⁸¹ package was used for Kyoto Encyclopedia of Genes and Genomes (KEGG) gene set enrichment analysis³⁸² with Benjamini/Hochberg FDR correction³⁷¹.

2.7.11 Visualisation of TFBSs in promoters

TFBSs for NLP6/7, DREB26 and TGA1 were identified using FIMO (v5.4.1) using TFBSs from DAP-seq¹⁸¹ with a p-value cut-off of 10^{-4} , with a zero-order background model created using `fasta-get-markov`³⁷⁰. Root and shoot ATAC-seq data³⁵⁰ (see section 2.7.7 Open chromatin coverage) were used to annotate open chromatin regions. Bigwig files from the NaOH treatment were downloaded from this webpage: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116287>. Genbank files containing the FIMO scanned and manually annotated TFBSs and open chromatin regions were downloaded from Benchling. A custom Jupyter Notebook³³⁷ using Python code (https://github.com/samwitham/PromoterArchitecture/blob/master/src/TRAMP/candidate_TFBSs_for_manuscript_1000bp%20downstream.ipynb) was used to create figures. Genbank files were read in as seqrecords using SEQIO from BIOPYTHON (v1.79)³⁵⁹, and the genes were shortened to 1000 bp downstream of the ATG. A custom graphic record was generated using a custom BiopythonTranslator class from the DNA_FEATURES_VIEWER (v3.0.3) package³⁸³. This specified feature colours, and which features to include in the figure. Open chromatin regions were added above each promoter using the PYGENOMETRACKS (v3.0) package³⁸⁴. The MATPLOTLIB (v3.4.3) package was used to combine the promoter and open chromatin tracks together and create a custom legend³⁸⁵. For figures showing the location of CRISPR mutations in promoters (in chapter 4), mutants were aligned to Col-0 using BIOTITE.APPLICATION.MUSCLE Python module, and figures made using the BIOTITE.SEQUENCE.GRAPHICS MODULE³⁸⁶ (see https://github.com/samwitham/PromoterArchitecture/blob/master/src/CRISPR_library/mutations_above_promoter.ipynb).

Chapter 3

Validation of a regulatory subnetwork for nitrate-associated metabolism and growth

3.1 Preface

My contribution to this chapter includes the identification of TFBSs within promoters/5' UTRs. I designed DNA probes with Dr. Yaomin Cai and Dr. Tufan Oz for use in a plate-based, *in vitro* assay for measuring protein-DNA binding affinity (TRAM). I expressed and purified recombinant *AtARF18* protein while Dr. Cai expressed and purified *AtANAC032*, *AtDREB26*, *AtNLP6*, *AtNLP7* and *AtTGA1* for use in the TRAM assays. Dr. Yaomin Cai performed the TRAM assays. I assembled all plasmids for protoplast co-expression and performed and analysed all co-expression luciferase assays.

3.2 Introduction

Plants need to survive in variable environmental conditions without the ability to relocate. To do this, they modulate their growth and metabolism in response to environmental signals. These changes are often coordinated by GRNs comprised of sets of genes that interact with each other. GRNs contain nodes representing genes, and edges that represent the regulatory relationships between nodes. Knowledge of GRN topology can provide

insights into network function as well as an understanding of the mechanisms that underlie phenotypic robustness⁵. Such knowledge of GRNs allows the identification of strategies for engineering network response⁴. GRN topologies can change depending on the time point or developmental stage that a transcriptomic snapshot is taken following treatment with a stimulus³⁸⁷. Network motifs such as feedforward and feedback loops can cause delayed or increased activation or repression of target genes, contributing to network complexity and dynamics³⁸⁸.

GRNs can be inferred from gene expression data including microarray or RNA-seq data using the assumption that genes expressed at the same time might contribute to phenotypes exhibited at or around the same time point. However, these correlation networks cannot confirm regulatory interactions between TFs and their targets. To overcome these limitations, connections in the network can be validated by obtaining data to support these interactions. This can include evidence that TFs can bind to cognate sequences (TFBSs) in the regulatory regions of target genes and/or evidence that TFs cause a change in the expression levels of target genes.

TFBSs can be identified using genome-wide methods such as ChIP-seq (see section 1.5.6) or DAP-seq (see section 1.5.9).

Other methods for the identification of TFBSs include systematic evolution of ligands by exponential enrichment followed by sequencing (SELEX-seq) (see section 1.5.5) and PBM (see section 1.5.8). TFBSs can be identified from the DNA which bound the TF of interest, and PWMs can be derived³⁸⁹.

Once TFBSs are known for a given TF and PWMs have been derived, motif-finding software can be used to predict candidate TFBSs in genomes using pattern matching^{364,390}. Because DNA-binding domains are often conserved between protein families it can be difficult to determine which specific family members bind to a given site in a target gene³⁸⁹. Due to the conservation of DNA-binding domains, candidate TFBSs can be identified in genomes for which primary datasets have yet to be generated.

Other methods can be used to experimentally confirm interactions between TFs of interest and fragments of DNA such as promoter regions. These include Y1H, where a TF fused to an activator domain is introduced to a yeast strain along with a reporter construct

containing a test promoter controlling a reporter gene²¹⁶ (see section 1.5.1). Electrophoretic mobility shift assay (EMSA) is a more limited technique that can test binary interactions between TFs and candidate regulatory regions, and can also, with the use of mutated or competitor probes, identify and characterise specific TFBSs³⁹¹. Protein-DNA mixtures are separated by gel electrophoresis, with protein-DNA complexes migrating more slowly than the free DNA³⁹¹.

However, evidence of TF-DNA interaction is not always informative of regulatory consequences. The functional consequences of interactions between TFs and candidate target genes can be inferred from transcriptomics data or, more reliably, they can be tested using specific assays. One common approach is to constitutively or inducibly overexpress a given TF and investigate if and how this affects the expression of candidate target genes. Other assays developed for plants include the TARGET method in which protoplasts are transfected with a TF::GR fusion that can be translocated to the nucleus by the application of DEX²²⁰. Application of cycloheximide arrests further translation, allowing the determination of direct regulation. TARGET data has recently been used to build a GRN that is activated upon osmotic stress³⁹² and was also used to build a GRN of the plant response to nitrate³¹⁸. On a smaller scale, reporter assays can be used in which promoters are fused to a reporter gene such as luciferase and co-expressed with the TF of interest³⁹³. However, such reporter assays are unable to distinguish whether changes in expression are consequences of direct or indirect interactions.

Recently, several Arabidopsis nitrate response GRNs have been constructed using different types of data. Varala et al.³⁹⁴ measured the changes in gene expression over time in response to nitrate and ammonium, and used this data to build a GRN of the Arabidopsis response to nitrate; Brooks et al.³¹⁸ integrated data from TARGET using 33 nitrogen-early response TFs into time-series transcriptome data from Varala et al.³⁹⁴; Alvarez et al.³⁹⁵ integrated RNA-seq, RNA polymerase II occupancy and DNase I hypersensitive sites sequencing (DNase-seq) to establish a hierarchical contribution of TFs to the nitrate response. Interestingly, but perhaps expectedly given variations in growth and treatment conditions, time points and methods, there are differences between GRNs constructed in these different studies. Gaudinier et al.²⁵ used a more targeted approach using Y1H to identify TFs that interact with the promoter regions of genes for which there was existing

evidence for involvement in nitrogen processes including transport, assimilation and signalling. This Y1H network for nitrogen-associated metabolism comprised 1660 interactions between 431 genes, 345 TFs and 98 promoters. In this study, 21 TFs were hypothesised to coordinate the responses of the wider network. These were further investigated by analysis of mutant lines in which phenotypic changes to root architecture, shoot development and transition from vegetative to reproductive growth under limiting (1 mM KNO₃) and sufficient (10 mM KNO₃) nitrogen. These 21 TFs are therefore predicted to regulate a wide spectrum of root and shoot developmental processes in response to changes in N-availability. This combination of systems and genetic resources provided a promising dataset for informing engineering strategies to manipulate plant growth in response to nitrate. However, although the Y1H provides some support of regulatory interactions, it cannot provide evidence on the TFBSs within the target gene, limiting the potential to engineer regulatory sequences. Further, Y1H can generate false positive interactions because of transcription initiation by endogenous yeast TFs²¹⁷. In contrast, false negatives may result from improper folding or localisation, or because binding of TFs to TFBSs might require post-translational modifications that only occur *in planta*.

This chapter is focussed on validating and characterising interactions between a subnetwork of TFs identified by Gaudinier et al.²⁵. These TFs were predicted to cross-regulate each other and, directly or indirectly, many downstream genes involved in nitrate metabolism and growth. Interactions between ANAC032, ANR1, ARF9/18, DREB26, NLP6/7 and TGA1 are investigated (fig. 3.1). An important feature of this subnetwork is that it contains a putative feedforward loop through ARF18-DREB26-NLP7 and ARF18-ANAC032-NLP7²⁵, and a key outcome of this chapter will be to characterise this. This network motif is important as it could be a key target for disrupting the network.

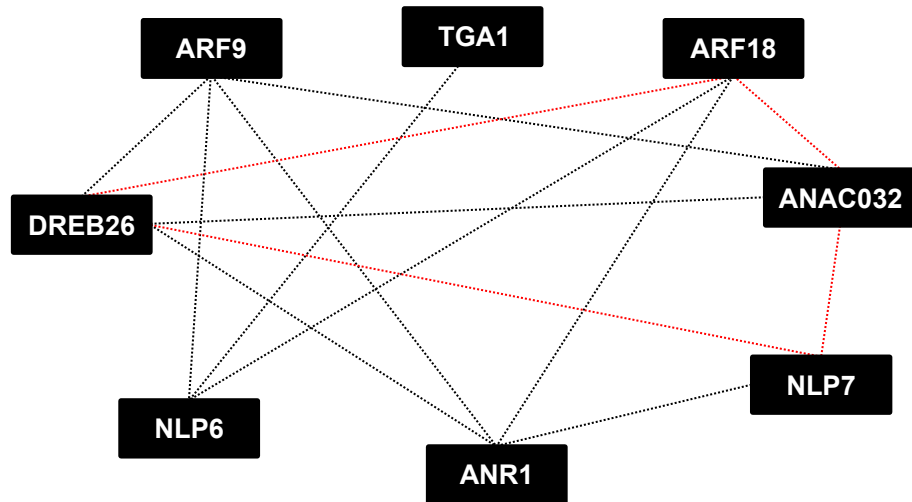


Figure 3.1: **A subnetwork of TFs predicted to cross-regulate each other in Arabidopsis in response to N.** Black lines represent putative connections based on RNA-seq and Y1H data²⁵. A feedforward loop from ARF18 to NLP7 through DREB26 and ANAC032 is shown through red dotted lines.

3.3 Aims

The first aim is to identify and test binding sites in promoters of genes in a regulatory subnetwork for nitrate-associated metabolism and growth. The second aim is to characterise edges in this regulatory subnetwork by determining the regulatory consequences of binding.

3.4 Experimental approach

Existing Y1H data predicted that each TF interacted with the promoter regions of their targets but could not predict where in the promoter they bound to, or whether they activated or repressed their targets. To identify TF binding locations in the promoters and 5' UTRs of each gene in the subnetwork, TFBSs were computationally predicted using DAP-seq¹⁸¹ PWMs. To test the activity of these predictions, I worked with a colleague, Dr. Yaomin Cai, who developed a plate-based, *in vitro* assay for measuring protein-DNA binding affinity (transcription factor relative binding affinity measurement; TRAM) (fig. 3.2). We worked together to design DNA probes to test each candidate binding site using this assay. I expressed and purified recombinant *At*ARF18 protein while Dr. Cai expressed and purified *At*ANAC032, *At*DREB26, *At*NLP6, *At*NLP7 and *At*TGA1. The protein and DNA probes were then used by Dr. Cai in TRAM assays.

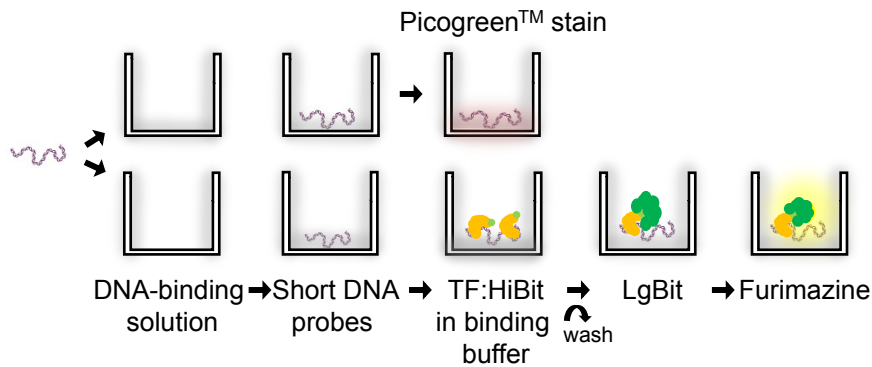


Figure 3.2: ***In vitro* TRAM protein-DNA binding affinity assay.** Short dsDNA probes are bound to the plate. The plates are incubated with low quantities of recombinant TFs with 9 amino acid HiBiT tags. After washing to remove unbound protein, the luciferase activity of the HiBiT tag is reconstituted by adding the large subunit, LgBit. In parallel, PicoGreen™ (1:200 in TE buffer) was added to wells containing dsDNA probes without TF and fluorescence was measured. The quantity of bound protein in each well was calculated by dividing luminescence by fluorescence. Relative binding affinity was reported relative to the background affinity for random DNA probes. See section 2.4 for further details.

Finally, two assays were used to investigate the regulatory consequences of interactions. I conducted reporter assays in which each promoter was fused to a luciferase reporter and co-expressed with each TF. In tandem, Dr. Tufan Oz, a colleague in our research group, investigated the same interactions using a variation of the TARGET method, called transient assay reporting specific effects of transcription factors (TARSET), in which protoplasts are transfected with a TF::GR fusion that can be translocated to the nucleus by the application of DEX²²⁰ with substitution of RNA-seq with qRT-PCR to determine direct and indirect activation or repression of gene targets (fig. 3.3).

At the time of writing, this data has been collated into a collaborative manuscript with equivalent data from tomato data generated in the Brady Lab (UC Davis) showing conservation and rewiring of this regulatory subnetwork across plant lineages.

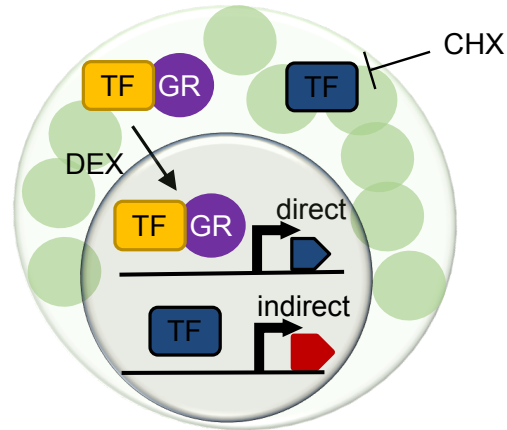


Figure 3.3: **TARSET assay schematic.** TFs are expressed in fusion with a GR-tag that enables translocation to the nucleus following application of dexamethasone. Application of cycloheximide arrests translation, preventing transcriptional cascades, allowing direct regulation to be distinguished.

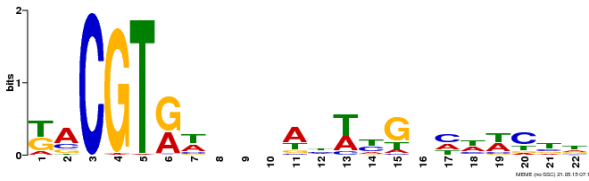
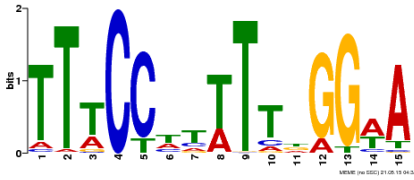
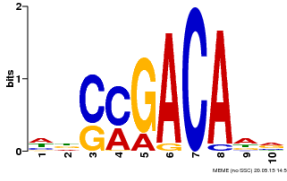
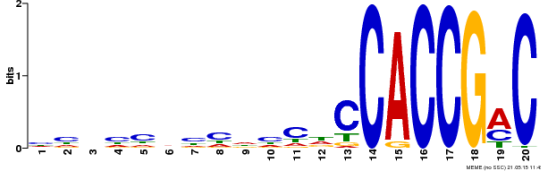
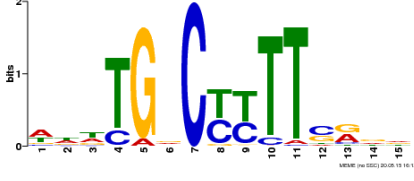
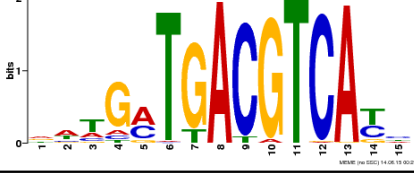
3.5 Results

3.5.1 Identification of candidate binding sites

For each gene (*ANAC032*, *ANR1*, *ARF18*, *DREB26*, *NLP6*, *NLP7*), a promoter region up to 2 kilobases (kb) upstream of the TSS or until the nearest protein coding gene was extracted, along with the 5' UTR. These regulatory regions will be referred to as CRMs from this point on. The CRM of *NIR1* was also selected since the binding of *NLP6/7* to the NRE in its CRM is well studied³⁹⁶. Candidate binding sites were identified using FIMO³⁶⁴ using PWMs from DAP-seq¹⁸¹. The PWMs used for each TF are represented by sequence logos in table 3.1. For *ANAC032*, *ANR1* and *ARF9/18*, TFBSs were identified from closely related TFs as binding site data for these TFs was not present in the DAP-seq or other datasets. The TFBS for Arabidopsis NAC domain containing protein 2 (*ATAF1/ANAC002*) (*AT1G01720*) was used to identify candidate binding sites for *ANAC032*, the TFBS for agamous-like 16 (*AGL16*) (*AT3G57230*) was used for *ANR1*, and the TFBS for auxin response factor 2 (*ARF2*) (*AT5G62000*) was used for *ARF9/18*, as the DNA-binding domains are highly conserved within these TF families. Some well-known TFBSs were not detected by FIMO including the NRE in the *NIR1* promoter, of which binding by *NLP7* has been previously reported³⁹⁶. Therefore, additional potential

TFBSs were identified by using Benchling searches (see <https://help.benchling.com/help/en-us/articles/9684279850765-Search-sequences-for-features>) for the following DAP-seq sequence logos: ANAC032-KACGTR; ANR1-TWMYHAAWDDRGWWW; DREB26-CCRCCGHC; TGA1-TGAYRTMAK; NLP6/7-TGNCYYTT. For ARF9/18 TFBSs, TGTCTC and TGTCGG motifs were used since they were shown to bind ARFs with higher affinity than other auxin response elements²⁸⁶. Additionally, pairs of TGTCNN motifs spaced less than 14 bp apart were annotated as candidate ARF9/18 TFBSs, since pairs of auxin response elements were shown to be enriched and preferentially bound in DAP-seq data^{286,397,398}.

Table 3.1: DAP-seq sequence logos used to identify candidate TFBSs using FIMO³⁶⁴.

TF	Sequence logo
ANAC032 (logo ATAF1/ANAC002)	
ANR1 (logo from AGL16)	
ARF9/18 (logo from ARF2)	
DREB26	
NLP6/7	
TGA1	

The locations of all candidate binding sites identified are shown in fig. 3.4. In addition, publicly available ATAC-seq data³⁵⁰ (see section 2.7.7) were used to identify regions of open chromatin, which are predicted to be more accessible to TFs. Finally, ~80 bp dsDNA probes containing candidate TFBSs flanked by fixed random sequence (5' TAGCGAAGT ACGATCCC and 3' GGCCATCACGCAGTA) were designed together with Dr. Cai and Dr. Oz using Benchling (fig. 3.4). Open chromatin regions were found around or upstream of the TSS of all seven genes fig. 3.4. Roughly half of the dsDNA probes were found to

have binding activity to their respective TFs (fig. 3.4).

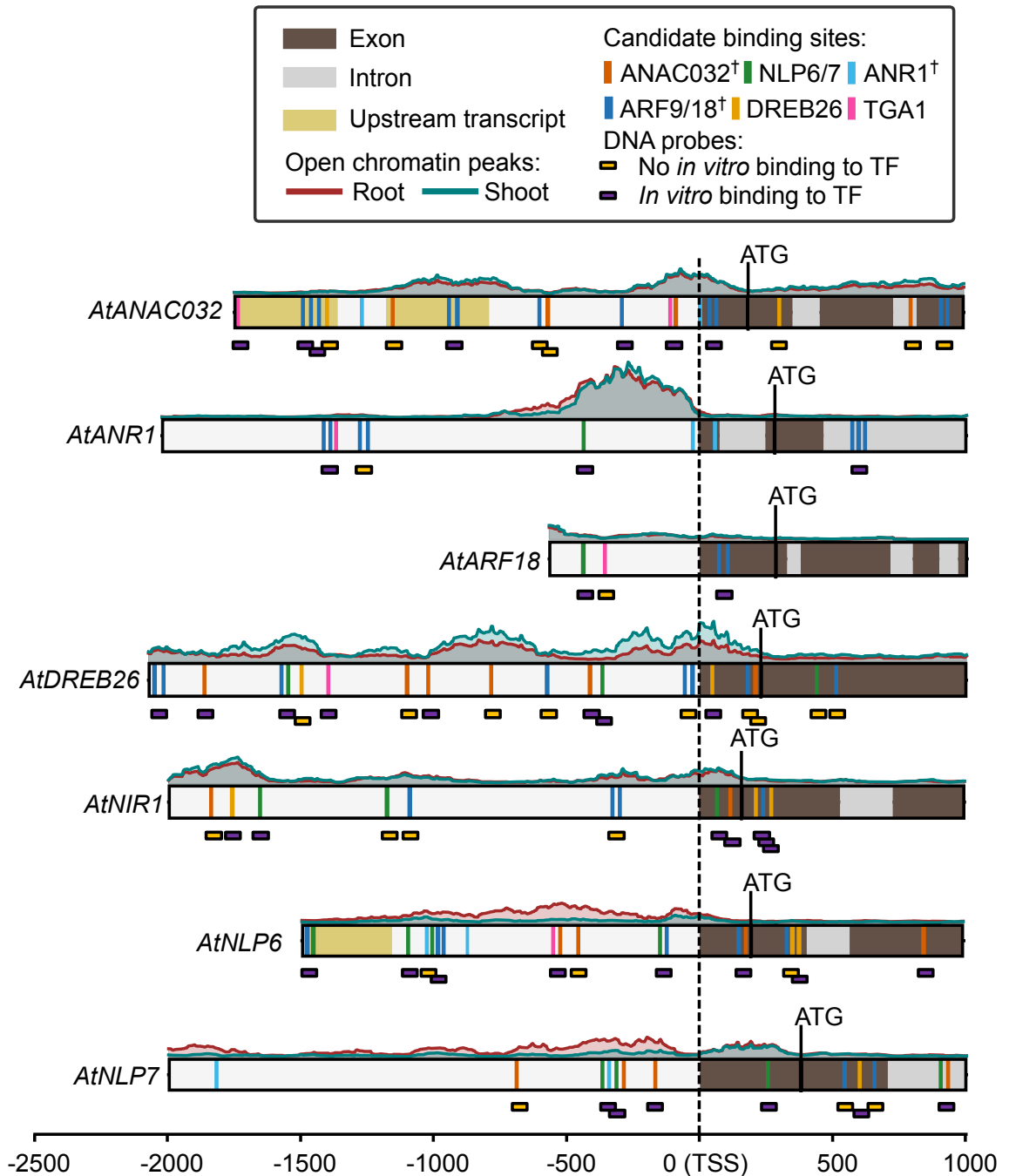


Figure 3.4: **Schematic of the upstream regulatory regions of genes in the nitrogen-response subnetwork showing gene structure and locations of open chromatin and candidate TFBSs.** PWMs were used to identify candidate TFBSs in the 5' regions (up to 2 kb or until the next protein-coding gene upstream and 1 kb downstream of the start of transcription including coding regions) of *AtANAC032*, *AtANR1*, *AtARF18*, *AtDREB26*, *AtNIR1*, *AtNLP6* and *AtNLP7*. The locations of DNA probes used to validate TF-DNA binding are illustrated by yellow and purple boxes beneath each gene. Purple coloured probe boxes indicate that binding of at least one TF was detected using *in vitro* relative affinity binding assays. †, TFBS inferred from closely related TF, see section 3.5.1. Overlapping TFBSs in this figure were offset by 20 bp for clarity. ATAC-seq data³⁵⁰ (see section 2.7.7) were used to annotate open chromatin peaks.

3.5.2 Expression and validation of *At*ARF18 in TRAM assays

A protein expression plasmid containing the ARF18::HiBiT fusion was assembled using the Gateway®BP Clonase™II Enzyme Mix (Invitrogen 11789) (see section 2.1.4). Recombinant ARF18::HiBiT was expressed using *E. coli* BL21(DE3) (see section 2.3.1). The protein was purified (see section 2.3.2) and quantified using the Nano-Glo® HiBiT Extracellular Detection System (see section 2.3.3). The size of the purified protein after SDS-PAGE (see section 2.3.4) was ~69 kDa as expected (69.108 kDa) (fig. 3.5).

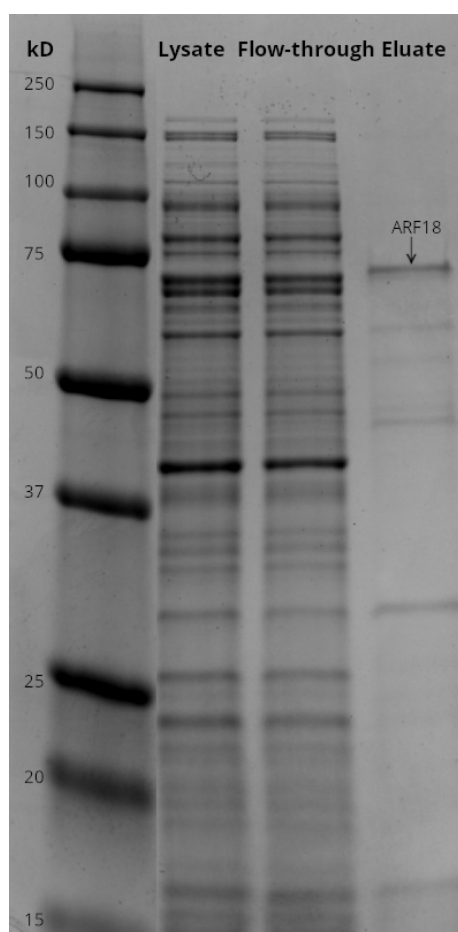


Figure 3.5: **ARF18 lysate, flow-through and eluate on stained gel after SDS-PAGE.** Protein was purified using INDIGO Ni-Agarose resin beads (PureCube Biotech 75103) (see section 2.3.2). Mini-PROTEAN TGX Stain-Free Precast Gel (Biorad 4568025) in 1× Tris/glycine/SDS buffer (Biorad 1610732). 5 µL protein ladder (Precision Plus Protein Dual Color 1610374) was loaded in the first well.

To check that the protein was functional in the TRAM assay, a positive control dsDNA probe containing seven tandem direct repeats of the auxin response element (TGTCTC),

each spaced by 5 bases (DR5; CCTTT)²⁸⁶ was synthesised and compared to a random dsDNA control containing no known TFBSs (see section 2.4). Data suggested that the DNA-binding domain of the protein was functional with a significantly higher relative binding affinity of ARF18::HiBiT to the DR5 probe (7.33 ± 1.36) than to the random control (0.22 ± 0.03 ; Welch's t-tests, $t = -9.07$, $P < 0.05$) (fig. 3.6).

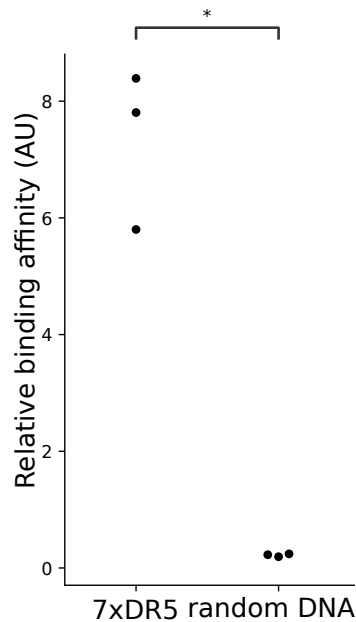


Figure 3.6: **Relative binding affinity of ARF18::HiBiT fusion to auxin response element positive control compared to random DNA probe.** $N = 3$ (technical replicates). P-values were calculated using Welch's t-tests. *, $P < 0.05$.

3.5.3 Validation of TF:DNA interactions within the subnetwork

Dr. Cai used purified protein of *AtARF18* (see section 3.5.2) together with purified protein of TFs *AtANAC032*, *AtDREB26*, *AtNLP6*, *AtNLP7* and *AtTGA1* to quantify relative binding affinity to the probes identified (fig. 3.4 and ??). He was unable to obtain protein for *AtANR1* or *AtARF9*. The data obtained by Dr. Cai are made available in the Appendix (????????????????). Significant binding between TFs and probes was identified in 58 out of 105 interactions. These data supported 29 edges in the subnetwork, which are summarised in fig. 3.7. Out of these 29 edges, 7 were supported by both Y1H²⁵ and TRAM data.

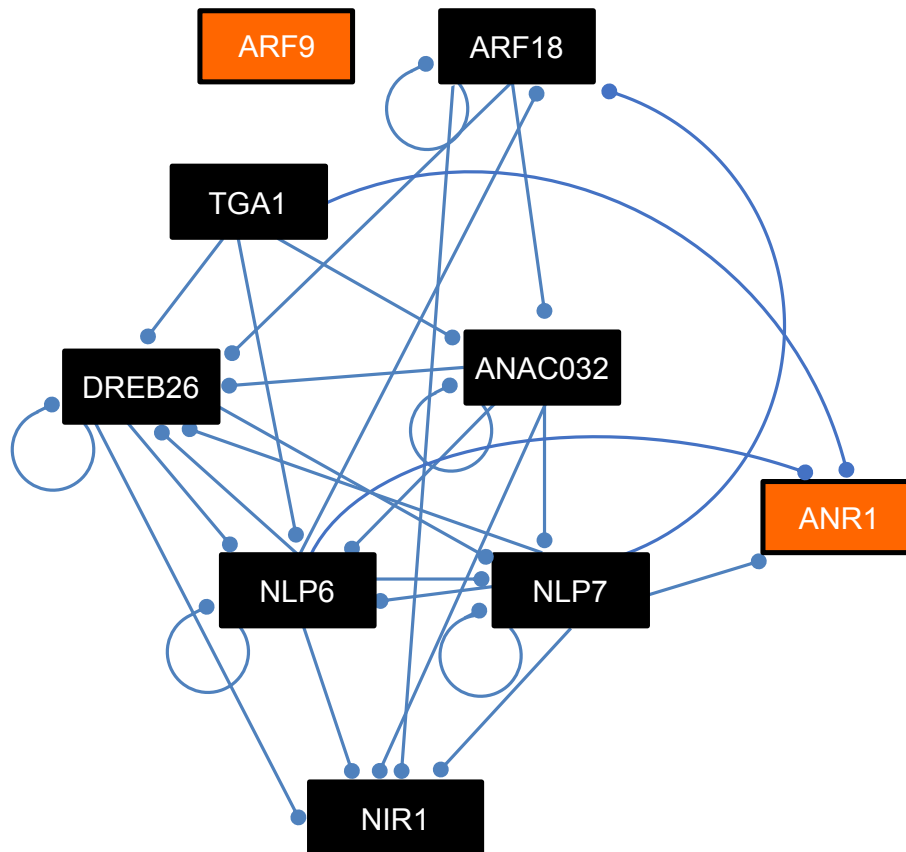


Figure 3.7: **Edges supported by *in vitro* TRAM protein-DNA binding affinity assay.** Protein from orange nodes could not be purified, and outgoing edges were not tested using TRAM. Many TF-DNA edges were supported by TRAM data, represented by blue lines in the direction of the blue dots.

3.5.4 Assembly of plasmids for luciferase assay

Promoters of *ANR1*, *DREB26*, *NLP6*, *NLP7*, and *NIR1*, including 1000 bp upstream of the TSS and 5' UTRs, were amplified from genomic DNA and cloned into pUPD2³²¹, pUAP1³⁰⁹ or pUAP4³²² universal acceptor plasmids in order to create Level 0 standard parts compatible with the Phytobrick assembly standard^{309,323} (see section 2.1.1). During cloning, two internal *Bsa*I sites in the *ANAC032* (301 bp upstream of TSS) and *ARF18* (258 bp upstream of TSS) promoters were removed by the introduction of a mutation. These mutations did not disrupt any candidate TFBSs.

The test promoters were then assembled into pCk1³²² with the coding sequence for LucN²⁷⁸ fused to a C-terminal 3×FLAG® epitope tag³²⁵ and the *AtuNOS* terminator³²⁵.

For positive controls, equivalent plasmids with the *AtuNOS* promoter or *CaMV35s* promoter were constructed.

The CDSs of each TF (ANAC032, ANR1, ARF9, ARF18, DREB26, NLP6, NLP7 and TGA1) were also cloned to produce standard parts which were used in assembly reactions (see section 2.1.3) with $p_{CaMV35s}\text{-}\Omega\text{TMV}$ and $CaMV35s$ terminator.

3.5.5 Optimisation of reporter co-expression assays

To determine the regulatory effect of a given TF on the expression of the candidate target promoter, constructs encoding promoter-LucN were used to transfect mesophyll protoplasts in the presence and absence of constructs encoding $p_{CaMV35s}\text{-TF}$ (see section 2.2.7). One exception was the testing of ARF TFs: several ARF TFs are known to interact with auxin/IAA proteins (aux/IAs) which recruit the co-repressor TOPLESS (TPL), inactivating the ARFs³⁹⁹. Class A activator ARF TFs interact with many aux/IAA proteins (see section 1.2.4). In the presence of auxin, the aux/IAA proteins are degraded, activating class A ARFs³⁹⁹ (see section 1.2.4). ARF9 and 18 are repressors in class B⁴⁰⁰. The relationship between repressor ARFs and aux/IAs is less clear but there is evidence that they interact with a limited number of aux/IAs³⁹⁸. Previous work has shown that ARF18 interacts with IAA33⁴⁰¹. IAA33 negatively regulates the auxin response by competing with canonical IAs, stabilising ARF18 in the presence of auxin. Therefore, $p_{CaMV35s}\text{-ARF9}$ and $p_{CaMV35s}\text{-ARF18}$ were also delivered in combination with a construct encoding $p_{2\times CaMV35s}\text{-IAA33}$ (pEPOZ1KN0222; kindly supplied by Dr. Tufan Oz) and with synthetic auxin (1 μM 1-Naphthaleneacetic acid (NAA)).

To determine if co-expression of the $p_{CaMV35s}\text{-TF}$ construct would reduce the expression of any test construct by competing for transcriptional machinery, an initial experiment was conducted in which luminescence from $p_{CaMV35s}\text{-LucF}$ was detected with or without co-expression of $p_{CaMV35s}\text{-YFP}$. Results confirmed that luminescence was indeed reduced when the $p_{CaMV35s}\text{-YFP}$ construct was co-expressed (284.0 ± 5.7) compared to without the $p_{CaMV35s}\text{-YFP}$ construct (519.3 ± 4.9) (Welch's t-tests, $t = 0.54$, $P < 0.001$) (fig. 3.8).

To ensure that the transcriptional load on protoplasts was equal with or without co-expression of the TF, $p_{CaMV35s}\text{-YFP}$ was included as a control plasmid that was co-transfected in place of $p_{CaMV35s}\text{-TF}$. To account for differences in transfection efficiency within and between experiments, a ratiometric system was used in which LucN luminescence of the test promoter is normalised to a co-transfected experimental calibrator ($p_{CaMV35s}$

or p_{AtuNOS} -LucF), followed by a batch calibrator (p_{AtuNOS} -LucN/ $p_{CaMV35s}$ -LucF) (fig. 2.4; see section 2.5.1).

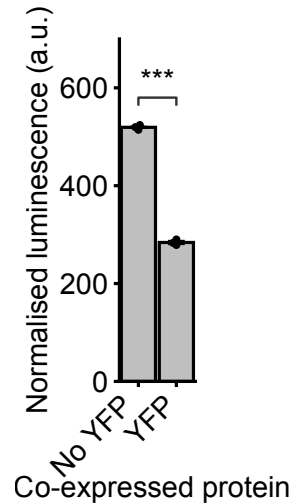


Figure 3.8: **Relative luminescence from $p_{CaMV35s}$ -LucN is reduced when co-expressed with $p_{CaMV35s}$ -YFP.** Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ (technical replicates). P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.6 Testing and selection of calibrators

The $CaMV35s$ promoter is known to be regulated by TGA1^{402–404}. To determine if the TFs to be tested would also affect the expression of the experiment calibrator, preliminary assays were conducted in which the $p_{CaMV35s}$ -LucN was co-transfected with each $p_{CaMV35s}$ -TF construct.

Welch's t-tests showed that, compared to the no TF control (2857.5 ± 30.1), luminescence from $p_{CaMV35s}$ -LucN was significantly increased when co-expressed with $p_{CaMV35s}$ -ANAC032 (6249.4 ± 70.9 ; $t = -76.2$ $P < 0.001$), $p_{CaMV35s}$ -ANR1 (5479.0 ± 28.3 ; $t = -109.8$ $P < 0.001$), $p_{CaMV35s}$ -ARF18 (3146.6 ± 49.0 ; $t = -8.7$ $P < 0.01$), $p_{CaMV35s}$ -DREB26 (5517.3 ± 116.6 ; $t = -38.3$ $P < 0.001$), $p_{CaMV35s}$ -NLP6 (4509.2 ± 96.4 ; $t = -28.3$ $P < 0.001$), $p_{CaMV35s}$ -NLP7 (4052.3 ± 42.17 ; $t = -39.9$ $P < 0.001$) and $p_{CaMV35s}$ -TGA1

(4073.6 ± 102.0 ; $t = -19.8$ $P < 0.01$) (fig. 3.9). Luminescence from $pCaMV35s$ -LucN was significantly decreased when co-expressed with $pCaMV35s$ -ARF9 (2747.3 ± 28.7 ; Welch's t-test $t = 4.59$ $P < 0.05$) compared to the no TF control (2857.5 ± 30.1) (fig. 3.9).

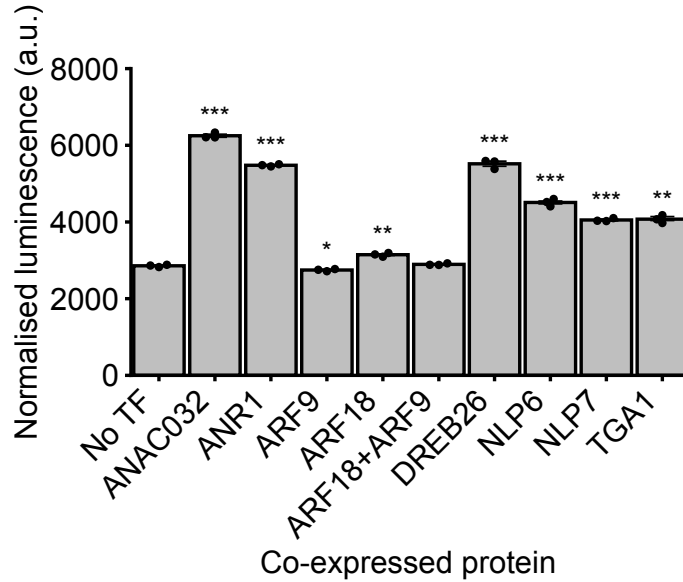


Figure 3.9: **Relative luminescence of $pCaMV35s$ -LucN following co-expression with Arabidopsis TFs.** No TF = $pCaMV35s$ -YFP; Experimental calibrator = $pCaMV35s$ -LucF. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ (technical replicates). P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

These results showed that several TFs (ANAC032, ANR1, ARF9, ARF18, DREB26, NLP6/7, TGA1) affect the expression of $CaMV35s$. To mitigate this, the same assay was repeated using an alternative calibrator vector (pEPYC1CB0003) containing $pAtuNOS$ - Ω TMV-LucF and the OCS terminator.

Welch's t-tests showed that, compared to the no TF control (530.0 ± 13.3), luminescence from $pAtuNOS$ -LucN was significantly increased when co-expressed with $pCaMV35s$ -ANAC032 (1397.1 ± 57.3 ; $t = -25.5$ $P < 0.001$), $pCaMV35s$ -ANR1 (585.8 ± 11.6 ; $t = -5.5$ $P < 0.01$), $pCaMV35s$ -ARF9 (701.3 ± 11.8 ; $t = -16.6$ $P < 0.001$), ARF18 (685.9 ± 7.9 ; $t = -16.6$ $P < 0.001$), $pCaMV35s$ -DREB26 (725.4 ± 12.5 ; $t = -18.5$ $P < 0.001$), $pCaMV35s$ -NLP6 (575.7 ± 8.8 ; $t = -5.0$ $P < 0.05$), and $pCaMV35s$ -TGA1 (833.7 ± 8.8 ; $t = -32.9$ $P < 0.001$) (fig. 3.10).

$pAtuNOS$ -LucN luminescence was significantly decreased when co-expressed with

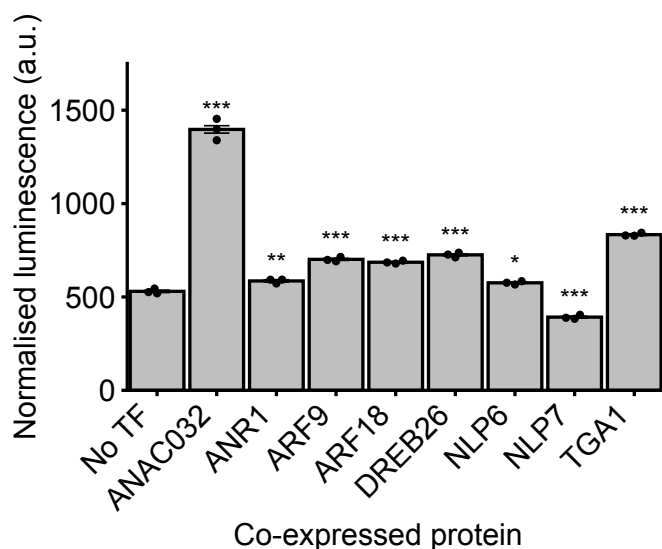


Figure 3.10: **Relative luminescence of p_{AtuNOS} -LucN following co-expression with Arabidopsis TFs.** No TF = $p_{CaMV35s}$ -YFP; Experimental calibrator = p_{AtuNOS} -LucF. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ (technical replicates). P -values were calculated using Welch’s t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

$p_{CaMV35s}$ -NLP7 (392.0 ± 11.1 ; Welch’s t-test $t = 13.8$ $P < 0.001$) compared to the no TF control (530.0 ± 13.3) fig. 3.10.

For each TF, the calibrator that least affected expression was selected (table 3.2).

Table 3.2: **Calibrators used in the TF co-expression assay.**

TF	Experiment calibrator promoter
<i>ANAC032</i>	<i>CaMV35s</i>
<i>ANR1</i>	<i>AtuNOS</i>
<i>ARF9/18</i>	<i>AtuNOS</i>
<i>DREB26</i>	<i>AtuNOS</i>
<i>NLP6/7</i>	<i>CaMV35s</i>
<i>TGA1</i>	<i>AtuNOS</i>

3.5.7 Regulatory effects of TF expression on $p_{ANAC032}$ -LucN

To test the effect of potential direct TF-DNA interactions on the luminescence from $p_{ANAC032}$ -LucN, ANAC032 and $p_{ANAC032}$ -LucN, and ANR1 and $p_{ANAC032}$ -LucN

were co-expressed since their candidate binding sites were found in the *ANAC032* co-expressed region (fig. 3.11A).

Relative luminescence from $p_{ANAC032}$ -LucN was increased by co-expression with $p_{CaMV35s}$ -ANAC032 (1.33 ± 0.02 ; Welch's t-tests, $t = -25.7$ $P < 0.01$; fig. 3.11B) compared to the no TF control (1.00 ± 0.02). In contrast, relative luminescence from $p_{ANAC032}$ -LucN was significantly decreased by co-expression with $p_{CaMV35s}$ -ANR1 (0.48 ± 0.00 ; Welch's t-tests, $t = 33.7$ $P < 0.001$; fig. 3.11C), ARF9 (no NAA: 0.74 ± 0.01 ; Welch's t-tests, $t = 15.0$ $P < 0.001$; $1 \mu\text{M}$ NAA: 0.88 ± 0.02 ; $t = 7.0$ $P < 0.05$; fig. 3.11D), ARF18 (no NAA: 0.67 ± 0.02 ; $t = 17.7$ $P < 0.001$; $1 \mu\text{M}$ NAA: 0.76 ± 0.02 ; $t = 10.4$ $P < 0.01$; fig. 3.11E), ARF18/IAA33 (no NAA: 0.65 ± 0.02 ; $t = 17.6$ $P < 0.001$; $1 \mu\text{M}$ NAA: 0.75 ± 0.02 ; $t = 10.4$ $P < 0.01$; fig. 3.11E) and a combination of ARF9/18 TFs (no NAA: 0.83 ± 0.01 ; $t = 11.2$ $P < 0.01$; $1 \mu\text{M}$ NAA: 0.93 ± 0.02 ; $t = 5.2$ $P < 0.05$; fig. 3.11F) compared to the no TF control (no NAA: 1.0 ± 0.03 ; $1 \mu\text{M}$ NAA: 1.11 ± 0.06) regardless of auxin treatment.

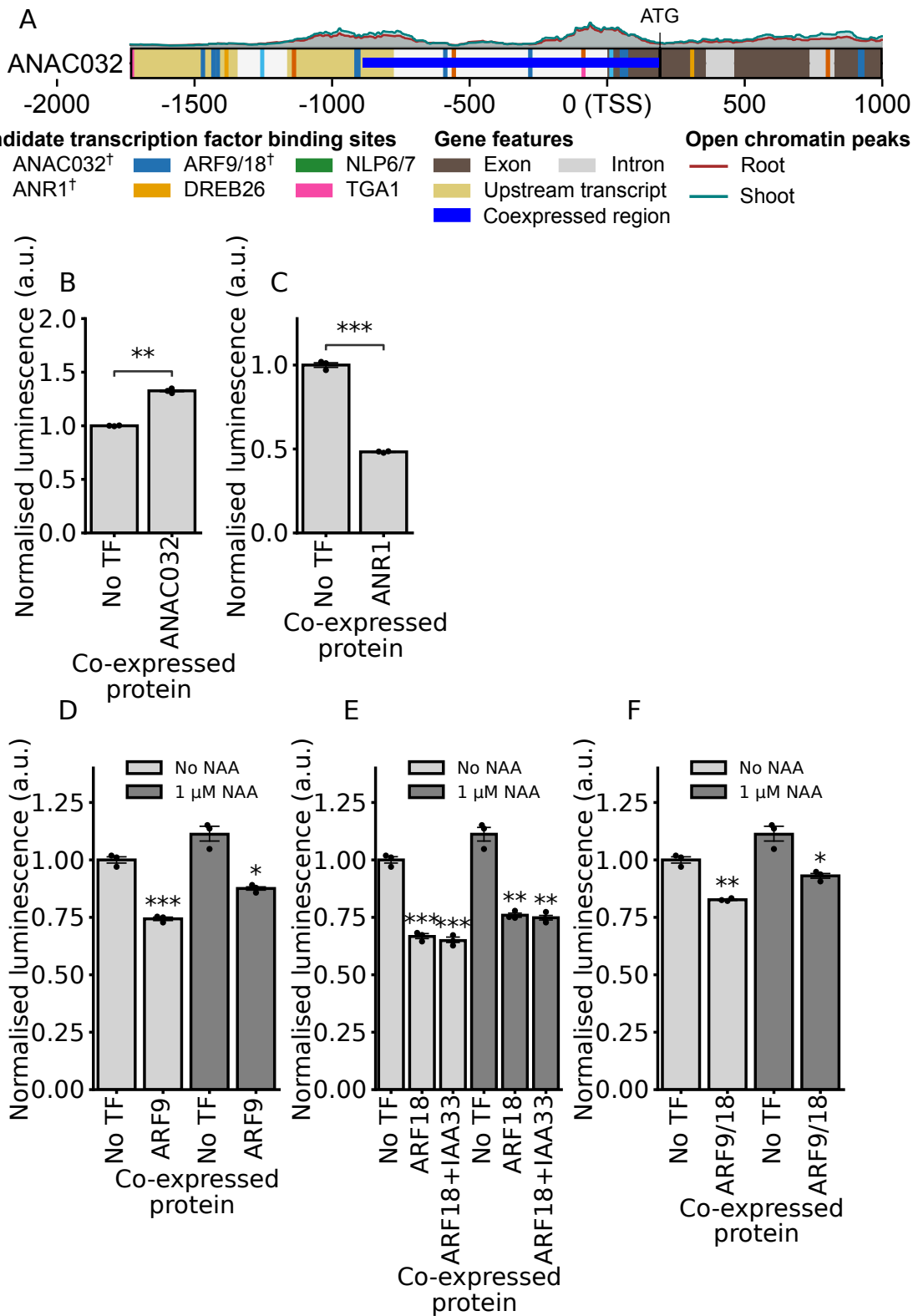


Figure 3.11: Normalised luminescence of $p_{ANAC032}$ -LucN following co-expression with *Arabidopsis* TFs. A: Candidate TFBSs in the *ANAC032* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–C: Normalised luminescence of the $p_{ANAC032}$ -LucN when co-expressed with different TFs. D–F: Normalised luminescence of the $p_{ANAC032}$ -LucN when co-expressed with $p_{CaMV35s}$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ (technical replicates). P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.8 Regulatory effects of TF expression on p *ANR1*-LucN

To test the effect of potential direct TF-DNA interactions on the luminescence from p *ANR1*-LucN, NLP7 and p *ANR1*-LucN were co-expressed, and ANR1 and p *ANR1*-LucN were co-expressed since their candidate binding sites were found in the *ANR1* co-expressed region (fig. 3.12A). To test the potential indirect regulatory effect from ANAC032 to *ANR1* through NLPs, and from ARFs to *ANR1* through NLPs (see figs. 3.16 and 3.17A), ANAC032 and p *ANR1*-LucN, and ARFs and p *ANR1*-LucN were co-expressed. Relative luminescence from p *ANR1*-LucN was significantly decreased when co-expressed with ANAC032 (0.84 ± 0.02 ; $t = 13.1$ $P < 0.01$; fig. 3.12B), ANR1 (0.48 ± 0.01 ; $t = 38.6$ $P < 0.001$; fig. 3.12C), NLP6 (0.34 ± 0.04 ; $t = 40.9$ $P < 0.001$; fig. 3.12D) and NLP7 (0.36 ± 0.08 ; $t = 22.7$ $P < 0.001$; fig. 3.12E) compared to the no TF control (1.00 ± 0.02). The data also showed that p *ANR1*-LucN luminescence was very strongly reduced when co-expressed with ARF9 (no NAA: 0.27 ± 0.02 ; $t = 56.6$ $P < 0.001$; 1 μ M NAA: 0.34 ± 0.00 ; $t = 70.7$ $P < 0.001$; fig. 3.12F), ARF18 (no NAA: 0.21 ± 0.00 ; $t = 62.9$ $P < 0.001$; 1 μ M NAA: 0.25 ± 0.00 ; $t = 79.9$ $P < 0.001$; fig. 3.12G), ARF18/IAA33 (no NAA: 0.20 ± 0.00 ; $t = 63.2$ $P < 0.001$; 1 μ M NAA: 0.17 ± 0.00 ; $t = 87.4$ $P < 0.001$; fig. 3.12G) and ARF9/18 (no NAA: 0.24 ± 0.02 ; $t = 46.9$ $P < 0.001$; 1 μ M NAA: 0.32 ± 0.02 ; $t = 65.9$ $P < 0.001$; fig. 3.12H) compared to the no TF control (no NAA: 1.00 ± 0.02 ; 1 μ M NAA: 1.12 ± 0.02), both with and without auxin.

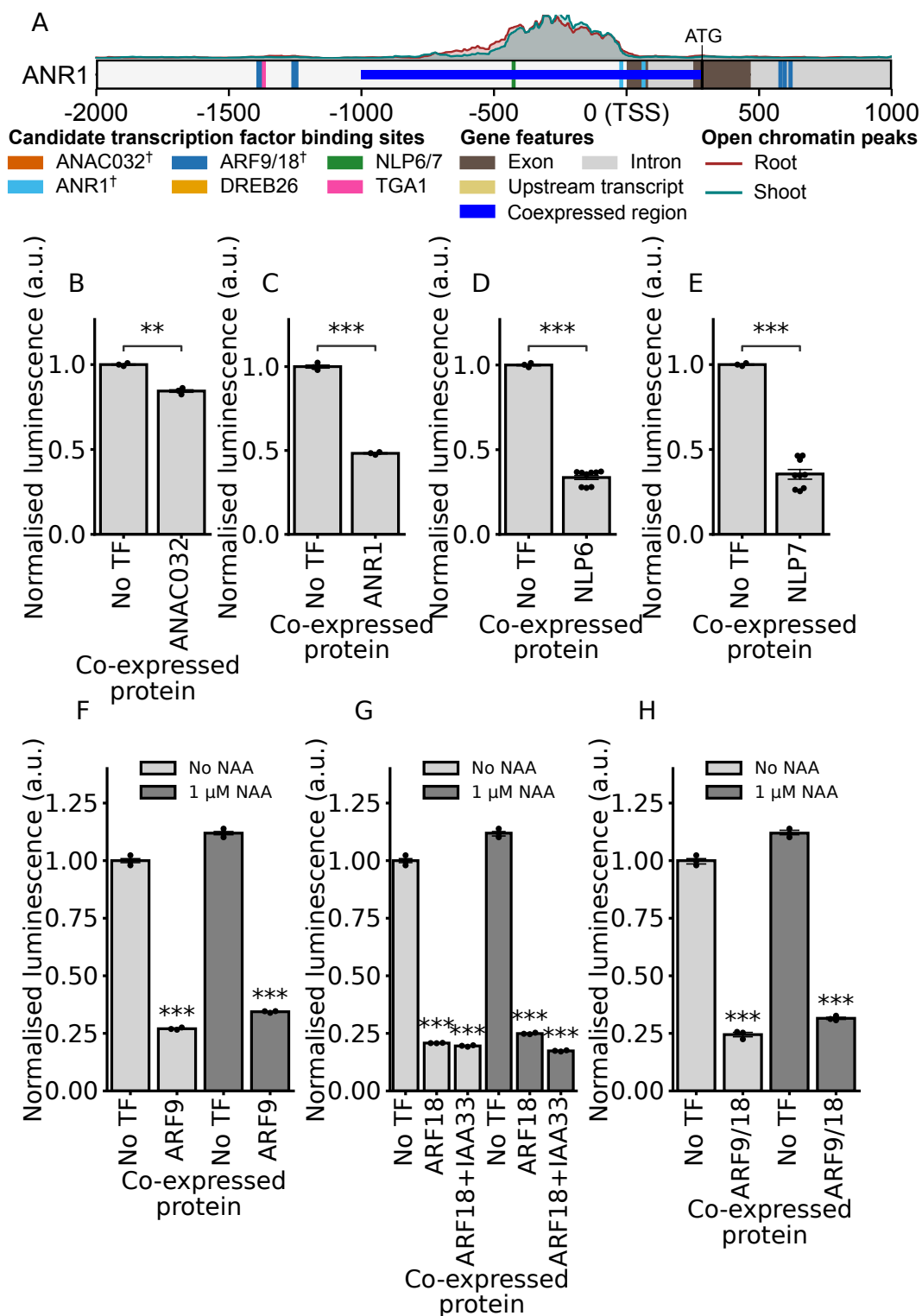


Figure 3.12: Normalised luminescence of p_{ANR1} -LucN following co-expression with *Arabidopsis* TFs. A: Candidate TFBSs in the *ANR1* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–E: Normalised luminescence of p_{ANR1} -LucN when co-expressed with different TFs. F–H: Normalised luminescence of p_{ANR1} -LucN when co-expressed with $p_{CaMV35s}$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3$ –9 (technical replicates). P -values were calculated using Welch's t -tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.9 Regulatory effects of TF expression on p_{ARF18} -LucN

To test the effect of potential direct TF-DNA interactions on the luminescence from p_{ARF18} -LucN, NLP7 and p_{ARF18} -LucN, and ARFs and p_{ARF18} -LucN were co-expressed since their candidate binding sites were found in the *ARF18* co-expressed region (fig. 3.13A). To test the potential indirect regulatory effects from ANAC032 and ANR1 to *ARF18* through NLPs (see figs. 3.16 and 3.17A), ANAC032 and ANR1 were co-expressed with p_{ARF18} -LucN.

The luminescence of p_{ARF18} -LucN was significantly increased following co-expression with ANAC032 (14.61 ± 0.43 ; $t = -54.9$ $P < 0.001$; fig. 3.13B), NLP6 (1.60 ± 0.02 ; $t = -36.9$ $P < 0.001$; fig. 3.13D) and NLP7 (1.83 ± 0.02 ; $t = -54.6$ $P < 0.001$; fig. 3.13E) compared to the no TF control (1.0 ± 0.03). In contrast, p_{ARF18} -LucN luminescence was strongly repressed following co-expression with ARF18 (no NAA: 0.34 ± 0.01 ; $t = 44.3$ $P < 0.001$; $1 \mu\text{M}$ NAA: 0.42 ± 0.01 ; $t = 44.8$ $P < 0.001$; fig. 3.13F), ARF18/IAA33 (no NAA: 0.36 ± 0.02 ; $t = 35.5$ $P < 0.001$; $1 \mu\text{M}$ NAA: 0.44 ± 0.01 ; $t = 45.6$ $P < 0.001$; fig. 3.13F) and ARF9/18 (no NAA: 0.34 ± 0.01 ; $t = 53.1$ $P < 0.001$; $1 \mu\text{M}$ NAA: 0.43 ± 0.01 ; $t = 44.7$ $P < 0.001$; fig. 3.13G) compared to the no TF control (no NAA: 1.00 ± 0.02 ; $1 \mu\text{M}$ NAA: 1.14 ± 0.03 ; fig. 3.13G), regardless of auxin concentration. p_{ARF18} -LucN luminescence was also slightly repressed following co-expression with ANR1 (0.94 ± 0.00 ; $t = 4.9$ $P < 0.05$; fig. 3.13C) compared to the no TF control.

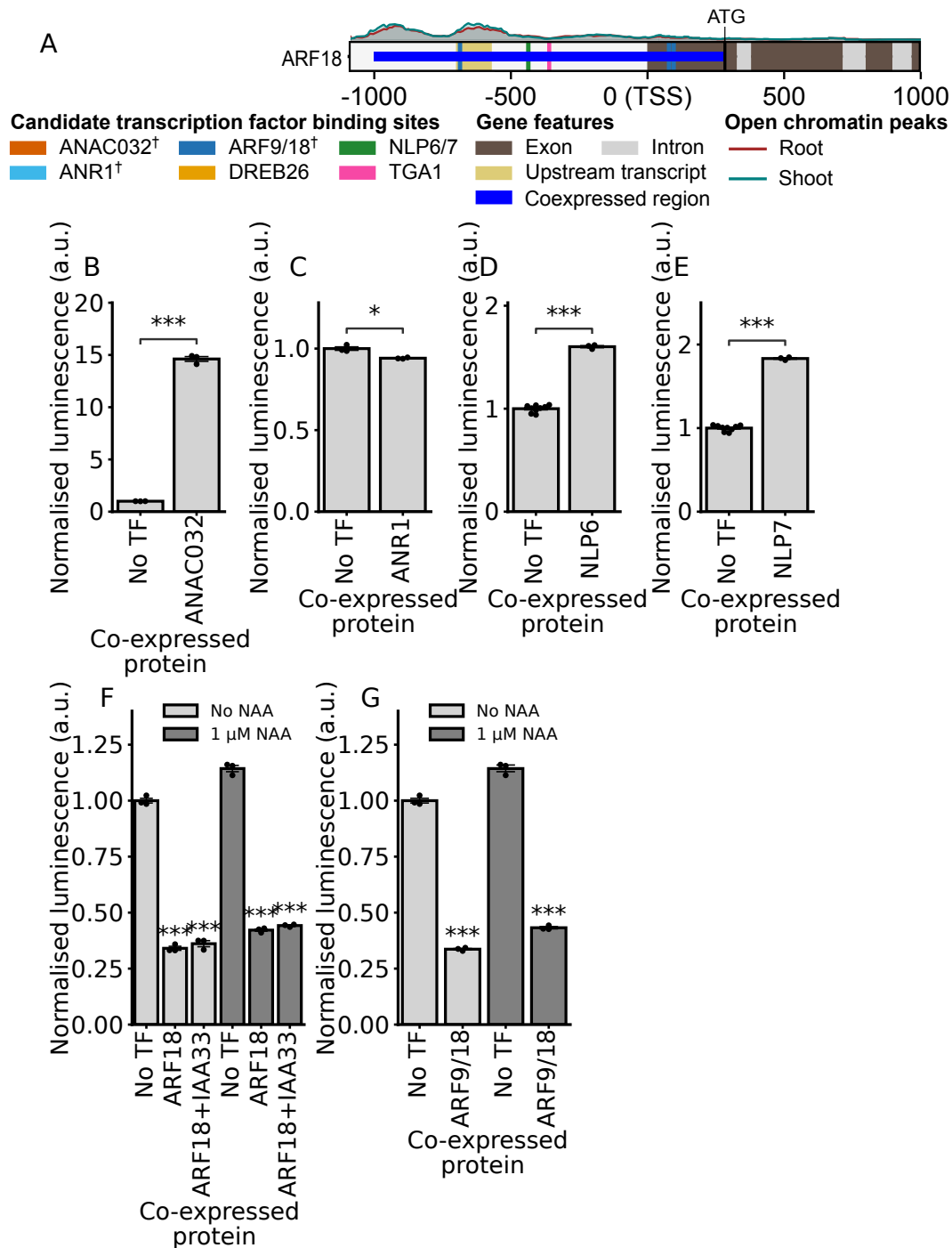


Figure 3.13: **Normalised luminescence of p_{ARF18} -LucN following co-expression with *Arabidopsis* TFs.** A: Candidate TFBSs in the *ARF18* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–E: Normalised luminescence of the p_{ARF18} -LucN when co-expressed with different TFs. F–G: Normalised luminescence of the p_{ARF18} -LucN when co-expressed with $p_{CaMV35s}$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3-9$ (technical replicates). P -values were calculated using Welch's t -tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.10 Regulatory effects of TF expression on p_{DREB26} -LucN

To test the effect of potential direct TF-DNA interactions on the luminescence from p_{DREB26} -LucN, ANAC032, DREB26, NLPs and ARFs were co-expressed with p_{DREB26} -LucN since their candidate binding sites were found in the *DREB26* co-expressed region (fig. 3.14A). To test the potential indirect regulatory effects from ANR1 to *DREB26* through NLP6 (see fig. 3.16A), ANR1 was co-expressed with p_{DREB26} -LucN.

p_{DREB26} -LucN luminescence was increased by all co-expressed TFs apart from ARFs (ANAC032: 7.07 ± 0.13 ; $t = -81.0$ $P < 0.001$; ANR1: 1.28 ± 0.02 ; $t = -13.4$ $P < 0.001$; DREB26: 2.51 ± 0.06 ; $t = -39.1$ $P < 0.001$; NLP6: 1.58 ± 0.05 ; $t = -20.1$ $P < 0.001$; NLP7: 1.88 ± 0.14 ; $t = -10.9$ $P < 0.01$) (fig. 3.14B–F), compared to *DREB26* expression with the no TF control (1.0 ± 0.03). p_{DREB26} -LucN luminescence was significantly reduced in the no auxin treatment with ARF9 (0.86 ± 0.01 ; $t = 9.3$ $P < 0.01$), ARF9/IAA33 (0.79 ± 0.02 ; $t = 10.7$ $P < 0.001$), ARF18 (0.75 ± 0.01 ; $t = 15.7$, $P < 0.001$), ARF18/IAA33 (0.83 ± 0.01 ; $t = 10.7$, $P < 0.01$), ARF9/18 (0.79 ± 0.01 ; $t = 13.9$ $P < 0.01$) and ARF9/18/IAA33 (0.73 ± 0.01 ; $t = 16.8$ $P < 0.001$) compared to the no TF control (1.0 ± 0.3) (fig. 3.14G–I). However, in the presence of auxin, *DREB26* expression significantly increased when co-expressed with ARF9 (1.33 ± 0.04 ; $t = -5.7$ $P < 0.01$; fig. 3.14G), and significantly decreased when co-expressed with ARF9/18/IAA33 (0.95 ± 0.01 ; $t = 5.5$ $P < 0.05$; fig. 3.14I) compared to the no TF control (1.12 ± 0.05).

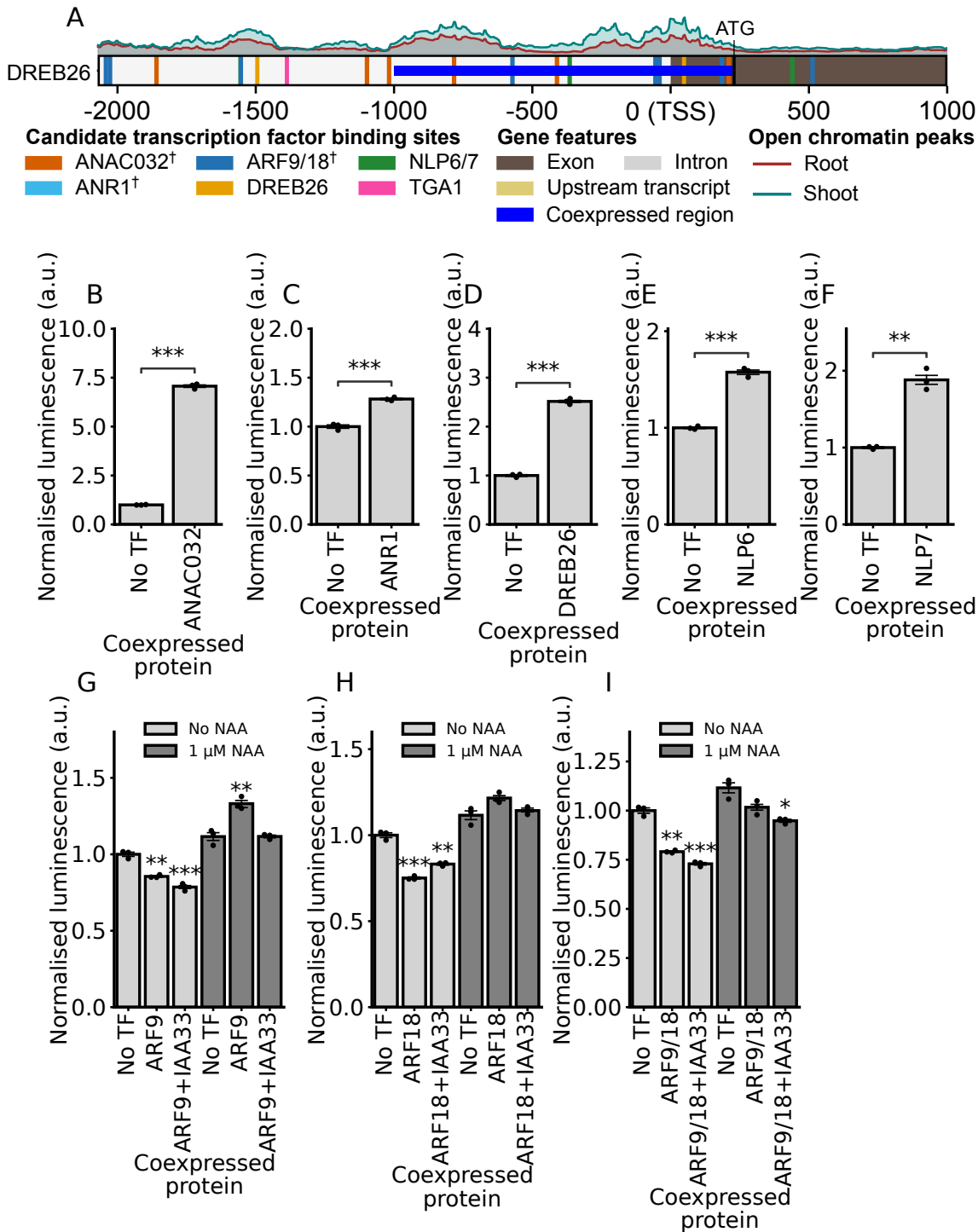


Figure 3.14: Normalised luminescence of $pDREB26$ -LucN following co-expression with Arabidopsis TFs. A: Candidate TFBSs in the *DREB26* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–F: Normalised luminescence of $pDREB26$ -LucN following co-expression with different TFs. G–I: Normalised luminescence of $pDREB26$ -LucN following co-expression with $pCaMV35s$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $pCaMV35s$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3$ (technical replicates). P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.11 Regulatory effects of TF expression on p_{NIR1} -LucN

To test the effect of potential direct TF-DNA interactions on the luminescence from p_{NIR1} -LucN, ANAC032, ARFs and NLPs were co-expressed with p_{NIR1} -LucN since their candidate binding sites were found in the *NIR1* co-expressed region (fig. 3.15A). To test the potential indirect regulatory effects from ANR1 to *NIR1* through NLP6 (see fig. 3.16A), ANR1 was co-expressed with p_{NIR1} -LucN.

The luminescence of p_{NIR1} -LucN was significantly increased following co-expression with ANAC032 (1.59 ± 0.02 ; $t = -43.6$ $P < 0.001$), ANR1 (1.86 ± 0.02 ; $t = -69.7$ $P < 0.001$), NLP6 (6.51 ± 0.03 ; $t = -325.7$ $P < 0.001$) NLP7 (9.76 ± 0.04 ; $t = -343.4$ $P < 0.001$) compared to the no TF control (1.0 ± 0.01) (fig. 3.15B–E). However, the luminescence of p_{NIR1} -LucN was significantly decreased following co-expression with ARFs (fig. 3.15F–H) compared to the no TF control (1.0 ± 0.01). ARF18 (no NAA: 0.48 ± 0.00 ; $t = 67.4$ $P < 0.001$; 1 mM NAA: 0.51 ± 0.001 ; $t = 94.3$ $P < 0.001$), ARF18/IAA33 (no NAA: 0.63 ± 0.00 ; $t = 87.8$ $P < 0.001$; 1 mM NAA: 0.53 ± 0.001 ; $t = 61.0$ $P < 0.001$) and ARF9/18 (no NAA: 0.47 ± 0.00 ; $t = 71.6$ $P < 0.001$; 1 mM NAA: 0.59 ± 0.001 ; $t = 89.8$ $P < 0.001$) reduced p_{NIR1} -LucN luminescence more than ARF9 (no NAA: 0.92 ± 0.00 ; $t = 10.5$ $P < 0.01$; 1 mM NAA: 0.92 ± 0.00 ; $t = 29.0$ $P < 0.01$) compared to the no TF control (no NAA: 1.0 ± 0.01 ; 1 mM NAA: 1.06 ± 0.01), regardless of auxin treatment.

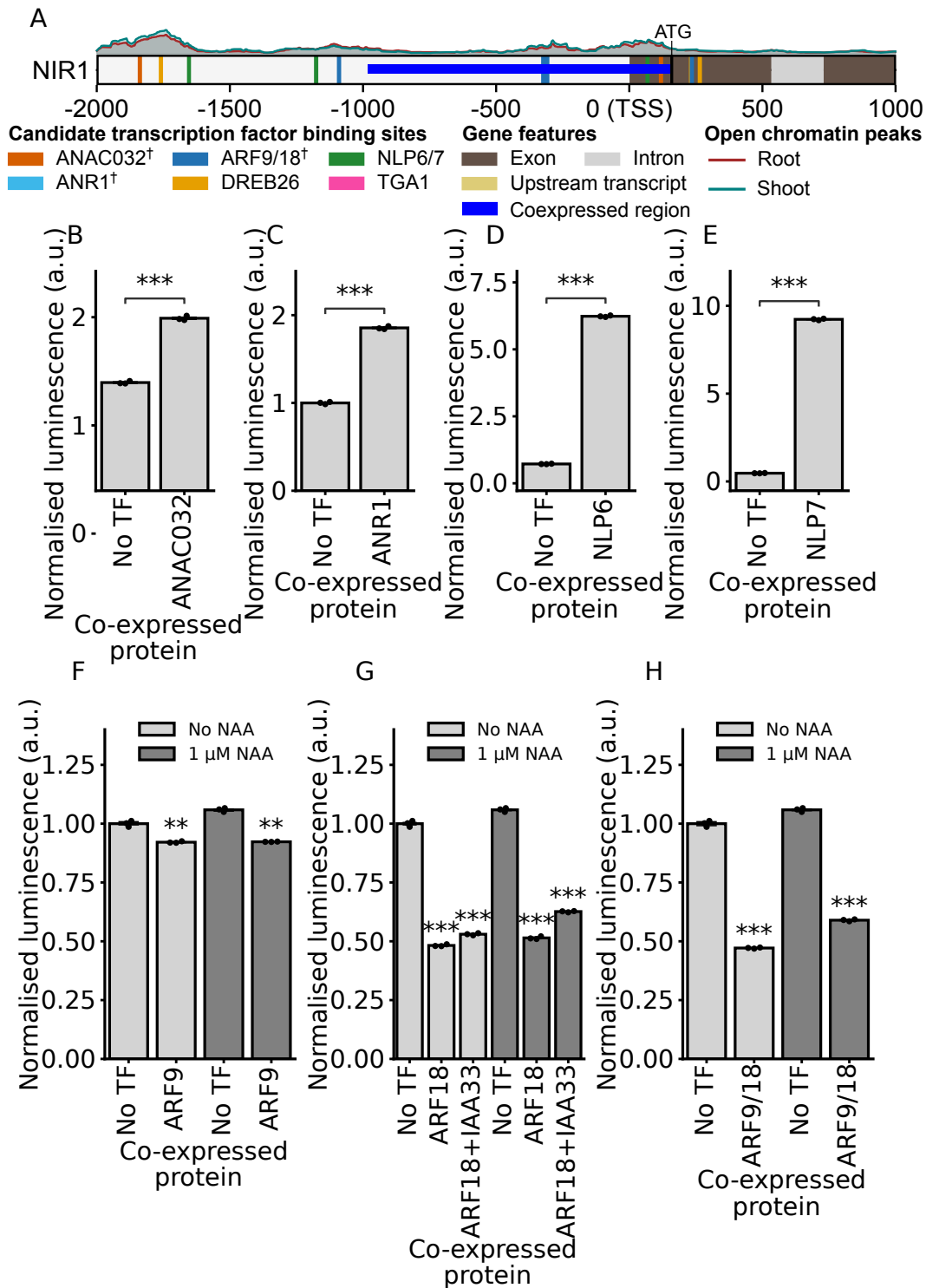


Figure 3.15: Normalised luminescence of p_{NIR1} -LucN following co-expression with *Arabidopsis* TFs. A: Candidate TFBSs in the *NIR1* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–E: Normalised luminescence of p_{NIR1} -LucN following co-expression with different TFs. F–H: Normalised luminescence of p_{NIR1} -LucN following co-expression with $p_{CaMV35s}$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ (technical replicates). P -values were calculated using Welch's t -tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.12 Regulatory effects of TF expression on $pNLP6$ -LucN

To test the potential effect of direct interactions on the luminescence of $pNLP6$ -LucN, ANAC032, ANR1, NLPs and ARFs were co-expressed with $pNLP6$ -LucN since their candidate TFBSs were found within the *NLP6*.

The luminescence of $pNLP6$ -LucN was activated by ANAC032 (1.94 ± 0.01 ; $t = -77.8$ $P < 0.001$), ANR1 (1.26 ± 0.00 ; $t = -157.2$ $P < 0.001$), NLP6 (1.19 ± 0.02 ; $t = -12.8$ $P < 0.001$) and NLP7 (2.99 ± 0.05 ; $t = -71.7$ $P < 0.001$) (fig. 3.16B–E) compared to the no TF control (1.0 ± 0.01).

ARF9 (no NAA: 0.67 ± 0.00 ; $t = 141.6$ $P < 0.001$; 1 mM NAA: 0.99 ± 0.02 ; $t = 21.5$ $P < 0.001$; fig. 3.16F), ARF18/IAA33 (no NAA: 0.68 ± 0.03 ; $t = 19.1$ $P < 0.01$; 1 mM NAA: 0.92 ± 0.01 ; $t = 38.3$ $P < 0.001$; fig. 3.16G) and ARF9/18 (no NAA: 0.83 ± 0.01 ; $t = 35.1$ $P < 0.001$; 1 mM NAA: 0.90 ± 0.01 ; $t = 33.5$ $P < 0.001$; fig. 3.16H) reduced $pNLP6$ -LucN luminescence with and without auxin, while ARF18 only reduced luminescence with auxin (1.15 ± 0.04 ; $t = 7.3$ $P < 0.01$; fig. 3.16G) compared to the no TF control (no NAA: 1.0 ± 0.00 ; 1 mM NAA: 1.33 ± 0.02).

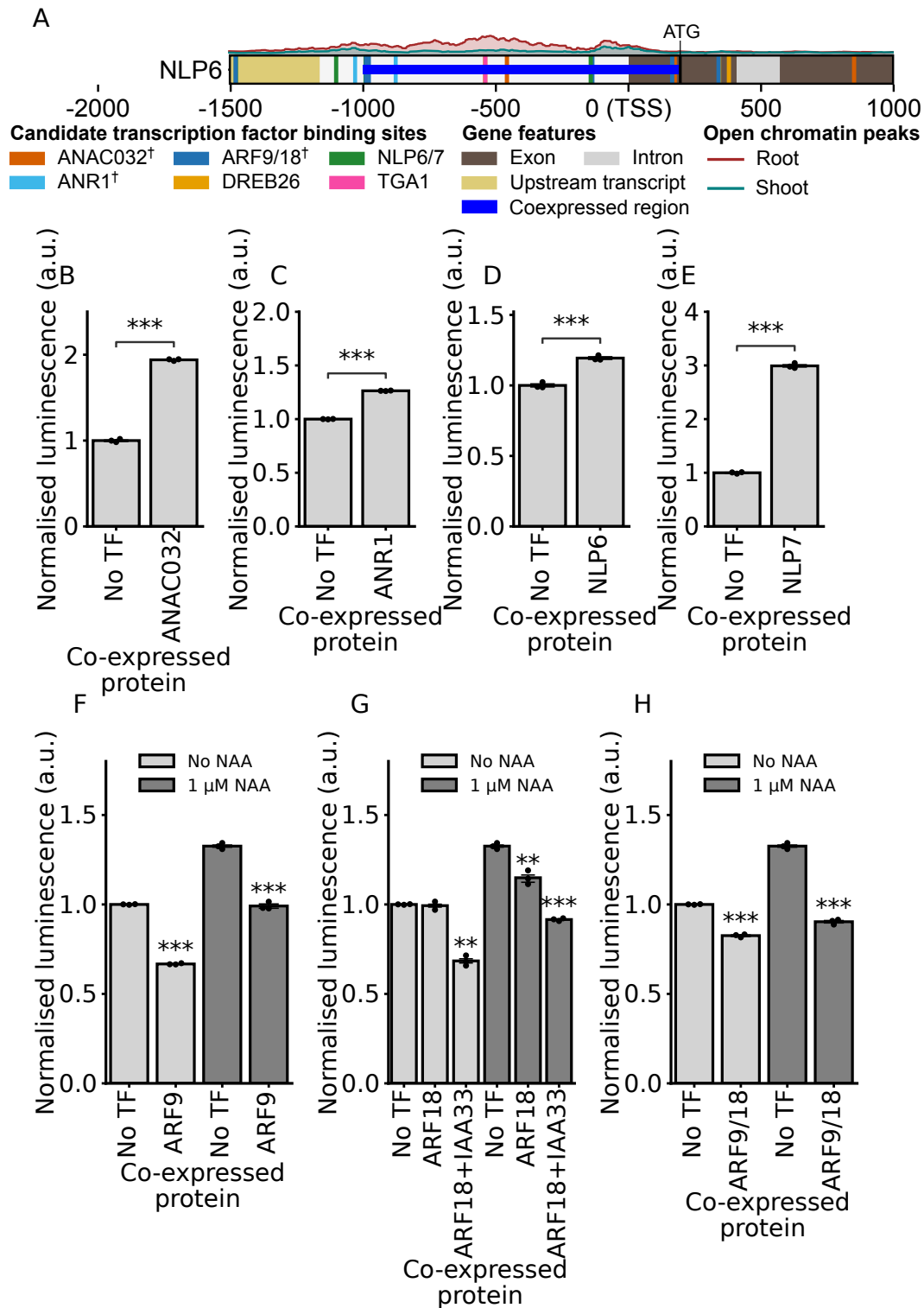


Figure 3.16: Normalised luminescence of p_{NLP6} -LucN following co-expression with *Arabidopsis* TFs. A: Candidate TFBSs in the *NLP6* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–E: Normalised luminescence of p_{NLP6} -LucN following co-expression with different TFs. F–H: Normalised luminescence of p_{NLP6} -LucN when co-expressed with $p_{CaMV35s}$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3$ (technical replicates). P -values were calculated using Welch's t -tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.13 Regulatory effects of TF expression on $pNLP7$ -LucN

Similar to $pNLP6$ -LucN luminescence, $pNLP7$ -LucN luminescence was also increased following co-expression with ANR1 (1.26 ± 0.01 ; $t = -18.9$ $P < 0.001$; fig. 3.17C), NLP6 (1.35 ± 0.02 ; $t = -9.2$ $P < 0.001$; fig. 3.17D) and NLP7 (1.33 ± 0.03 ; $t = -8.3$ $P < 0.001$; fig. 3.17E) compared to the no TF control. Additionally, $pNLP7$ -LucN was increased following co-expression with ARF9 with (1.50 \pm 0.02; $t = -15.6$ $P < 0.01$) and without NAA (1.50 \pm 0.02; $t = -33.0$ $P < 0.001$) (fig. 3.17F) compared to the no TF control (1 mM NAA: 1.27 \pm 0.01; no NAA: 1.0 \pm 0.02). $pNLP7$ -LucN luminescence was increased without auxin but reduced in the presence of auxin by ARF18 (no NAA: 1.20 \pm 0.04; $t = -8.1$ $P < 0.01$; 1 mM NAA: 1.07 \pm 0.01; $t = 20.2$ $P < 0.001$; fig. 3.17G), ARF18/IAA33 (no NAA: 1.26 \pm 0.02; $t = -16.7$ $P < 0.001$; 1 mM NAA: 1.03 \pm 0.01; $t = 24.0$ $P < 0.001$) and ARF9/18 (no NAA: 1.04 \pm 0.02; $t = -2.8$ $P < 0.05$; 1 mM NAA: 1.12 \pm 0.02; $t = 9.4$ $P < 0.01$; fig. 3.17H) compared to the no TF control.

In contrast to $pNLP6$ -LucN luminescence, the luminescence of $pNLP7$ -LucN was repressed by ANAC032 (0.53 ± 0.01 ; $t = 17.1$ $P < 0.01$; fig. 3.17B) compared to the no TF control (1.0 ± 0.02).

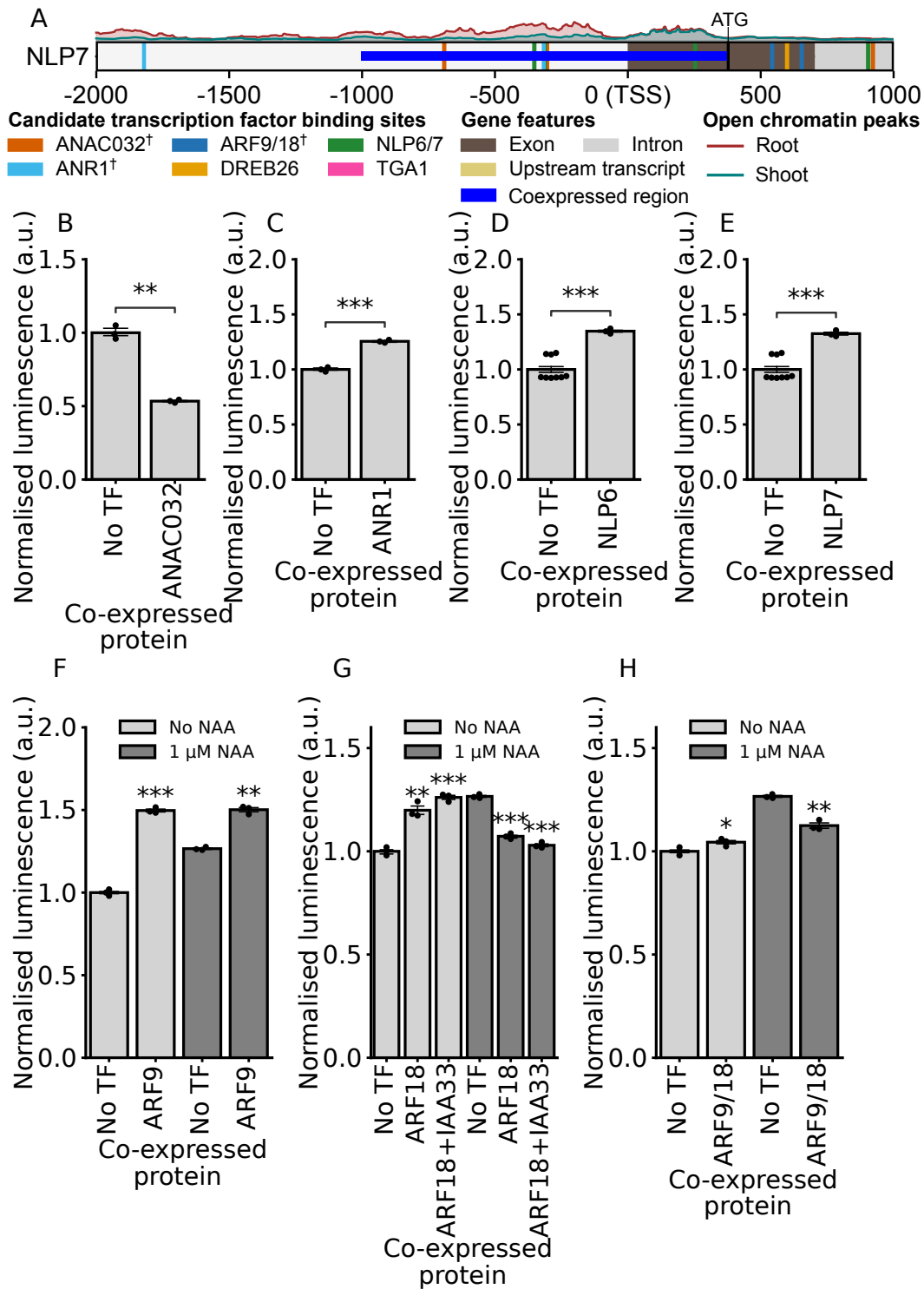


Figure 3.17: Normalised luminescence of $pNLP7$ -LucN following co-expression with *Arabidopsis* TFs. A: Candidate TFBSs in the *NLP7* promoter/5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–E: Normalised luminescence of $pNLP7$ -LucN when co-expressed with different TFs. F–H: Normalised luminescence of $pNLP7$ -LucN following co-expression with $pCaMV35s$ -ARF TFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $pCaMV35s$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3$ –9 (technical replicates). P -values were calculated using Welch's t -tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.14 Regulatory interactions between TGA1 and target genes

The $p_{CaMV35s}$ -TGA1 expression construct was also co-expressed with each candidate promoter-LucN construct to determine changes in expression. However, since specific technical issues were encountered, these results are presented separately: TGA1 was found to increase the normalised luminescence from all promoter-LucN constructs (fig. 3.18). This was true regardless of whether there was a candidate TFBS in the promoter (*ANAC032*, *ARF18*, *NLP6*) or not (*ANR1*, *DREB26*, *NIR1*, *NLP7*). This is inconsistent with data from previously published TARGET assays used in Brooks et al.³¹⁸ in which TGA1 was observed to be a negative regulator of *NLP6*, *ANAC032*, *NIR1* and *ARF18*. There are several differences between the TARGET assay and this reporter assay: the TARGET assay tests expression of endogenous genes while the co-expression assay tests expression of exogenous genes; in co-expression assays only a proportion of the promoter is present; in TARGET assays the TGA1 is fused to a GR tag.

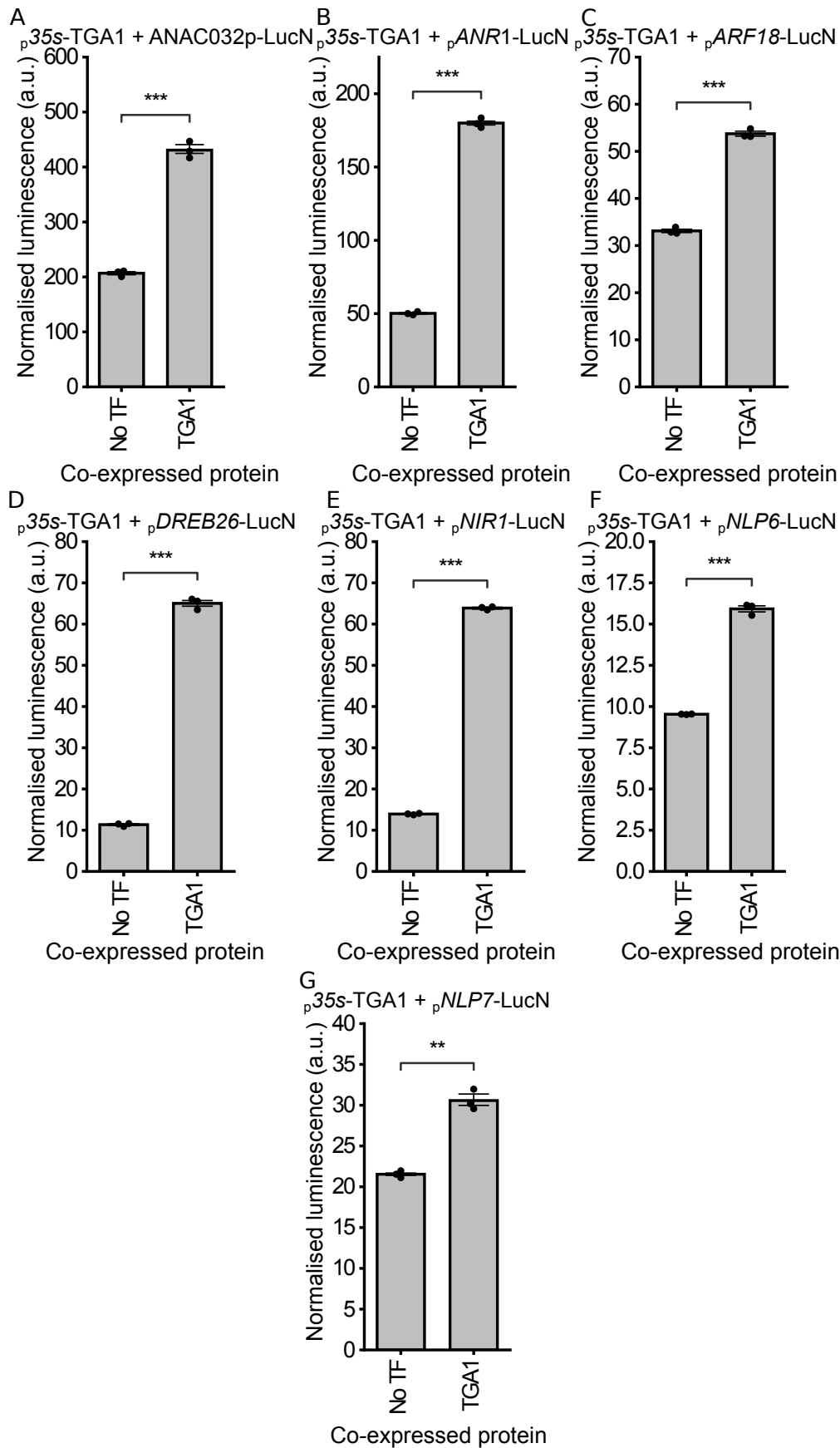


Figure 3.18: Normalised luminescence of promoter-LucN reporters following co-expression with ${}^pCaMV35s\text{-TGA1}$. No TF control = ${}^pCaMV35s\text{-YFP}$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ (technical replicates). P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.¹³⁴

While it is technically challenging to change the context and length of the promoter, it was possible to investigate whether the GR might alter the activity of TGA1. To do this, a $p_{CaMV35s}$ -TGA1::GR fusion was constructed and used in the reporter co-expression assay, with 10 μ M DEX added for three hours after two hours of 20 mM KNO₃ and 20 mM NH₄KNO₃ treatment (mimicking the conditions used in the TARGET assay)³¹⁸.

Relative luminescence from $p_{ANAC032}$ -LucN and p_{NIR1} -LucN was increased following co-expression with $p_{CaMV35s}$ -TGA1::GR (fig. 3.19A, C). Treatment with DEX led to moderate increases for the former (DEX: 73.00 ± 1.56 ; $t = 4.5$ $P < 0.05$; No DEX: 68.31 ± 0.91 ; fig. 3.19A) but substantial increases for the latter (DEX: 6.01 ± 0.21 ; $t = 18.8$, $P < 0.001$; No DEX: 3.42 ± 0.11). In contrast, relative luminescence from p_{ARF18} -LucN was reduced following co-expression with $p_{CaMV35s}$ -TGA1::GR and further reduced following treatment with DEX (DEX: 6.13 ± 0.07 ; $t = -25.6$ $P < 0.001$; No DEX: 8.67 ± 0.16) (fig. 3.19B). Relative luminescence from p_{NLP7} -LucN was increased by co-expression with $p_{CaMV35s}$ -TGA1::GR but reduced by application of DEX (DEX: 9.67 ± 0.15 ; $t = -8.3$, $P < 0.01$; No DEX: 11.00 ± 0.23) (fig. 3.19D). In general, these data suggest that the presence of the GR-tag may affect the activity of TGA1.

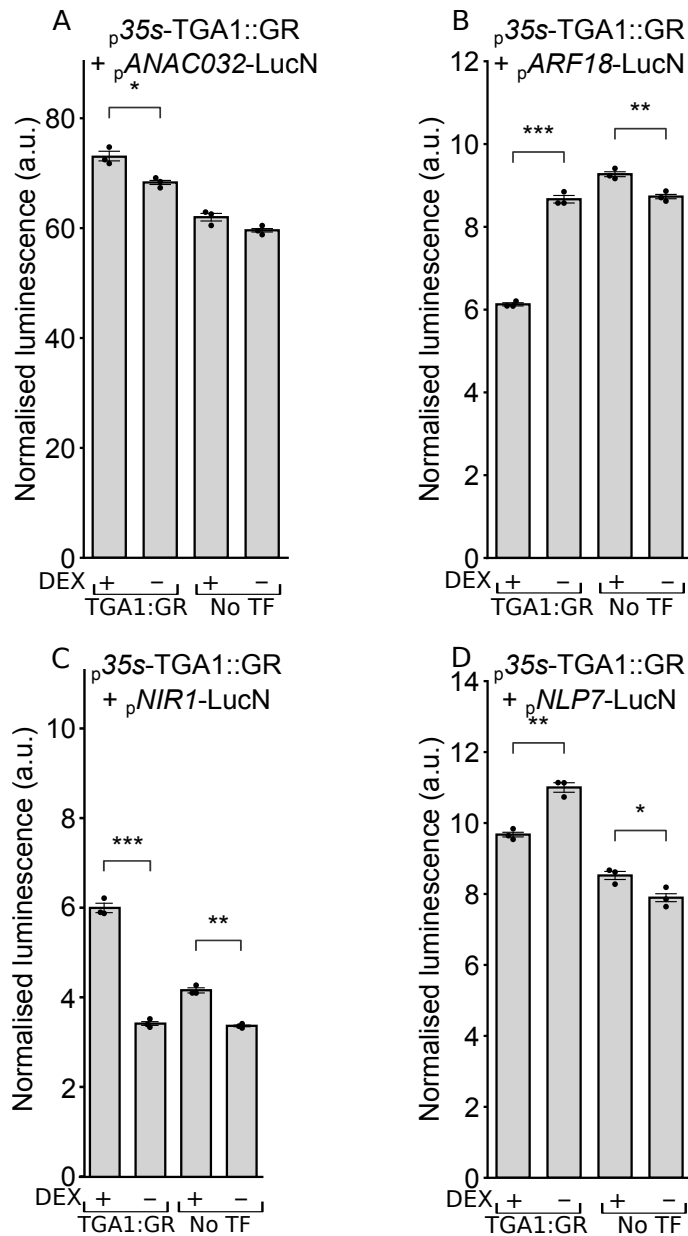


Figure 3.19: **Normalised luminescence of promoter-LucN reporters following co-expression with p_{CaMV35s}-TGA1::GR with and without dexamethasone treatment.** 10 μ M DEX added for three hours after two hours of 20 mM KNO₃ and 20 mM NH₄KNO₃ treatment in the morning after protoplast transfection. No TF control = p_{CaMV35s}-YFP. $N = 3$ (technical replicates). P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

3.5.15 Investigating the role of ARF binding sites in the *ANAC032* upstream promoter and untranslated region.

The co-expression assay can be adapted to examine the role of particular binding sites on the luminescence of the test promoter/5' UTR. Three different ARF TFBSs were mutated in the *ANAC032* promoter and 5' UTR to disrupt all incoming edges from ARF TFs

(fig. 3.20A). In contrast with the native $pANAC032$ -LucN where co-expression with ARFs repressed *ANAC032*, co-expression of the mutated $pANAC032$ -LucN with ARFs did not significantly alter luminescence ($P > 0.05$) compared to the no TF control, regardless of auxin concentration (fig. 3.20B). The luminescence of the mutated $pANAC032$ -LucN was significantly lower than the luminescence of the wild type (WT) $pANAC032$ -LucN in all conditions (WT: No NAA, no ARF18/IAA33: 206.8 ± 19.5 ; $P < 0.01$; No NAA with ARF18/IAA33: 134.2 ± 14.6 ; $P < 0.01$; $1 \mu\text{M}$ NAA, no ARF18/IAA33: 230.0 ± 90.0 ; $P < 0.01$; $1 \mu\text{M}$ NAA with ARF18/IAA33: 154.8 ± 15.3 ; $P < 0.01$; Mutated: No NAA, no ARF18/IAA33: 26.6 ± 0.1 ; No NAA with ARF18/IAA33: 34.5 ± 0.4 ; $1 \mu\text{M}$ NAA, no ARF18/IAA33: 34.3 ± 0.7 ; $1 \mu\text{M}$ NAA with ARF18/IAA33: 38.8 ± 0.6) (fig. 3.20B).

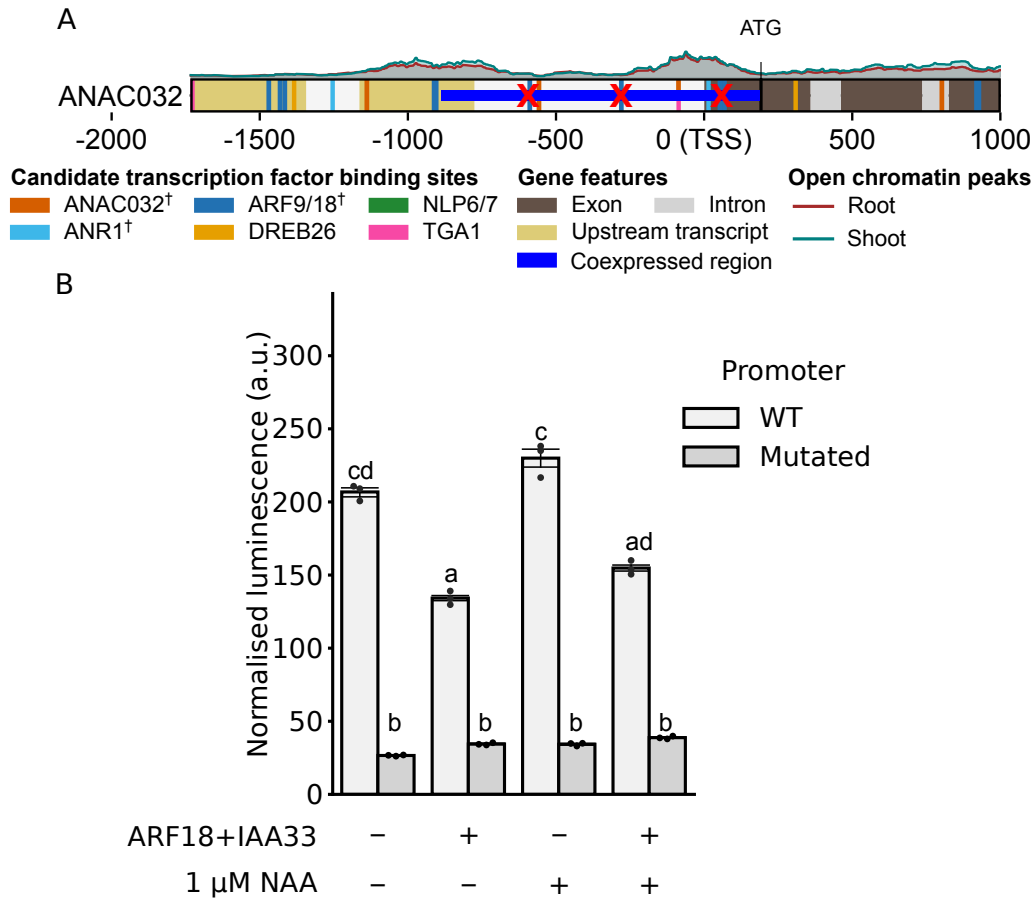


Figure 3.20: **Normalised luminescence of $pANAC032$ -LucN with mutated ARF TFBSs following co-expression with ARF TFs.** A: Position of ablated TFBSs in the *ANAC032* promoter (red crosses) and 5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B: Normalised luminescence of wild type $pANAC032$ -LucN or version containing ablated ARF TFBSs when co-expressed with different $pCaMV35s$ -ARFs with and without 1 μ M 1-Naphthaleneacetic acid (NAA). No TF control = $pCaMV35s$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3$ (technical replicates). Significance calculated using ANOVA and Tukey HSD post hoc tests³⁴⁴. Different letters indicate significant differences between treatments.

3.5.16 Investigating the role of TGA1 binding sites in the *ANAC032* promoter

An *ANAC032* promoter with a mutated TGA1 TFBS was also tested, to see if this prevented the direct regulation of *ANAC032* by TGA1 (fig. 3.21A). The *ANAC032* promoter with a mutated TGA1 TFBS was co-expressed with TGA1::GR, which increased the luminescence of the mutated $pANAC032$ -LucN in the presence of DEX (16.26 ± 0.11 ; $P < 0.05$) compared to no DEX (13.22 ± 0.25) (fig. 3.21B), similar to the response of the native *ANAC032* promoter (fig. 3.19A, fig. 3.21B). However, the luminescence of the mutated

$p_{ANAC032}$ -LucN was significantly lower than the WT promoter with and without DEX ($P < 0.01$; DEX: WT: 73.0 ± 1.6 ; Mutated: 16.3 ± 0.0 ; No DEX: WT: 68.3 ± 0.5 ; Mutated: 13.2 ± 0.0) (fig. 3.21B).

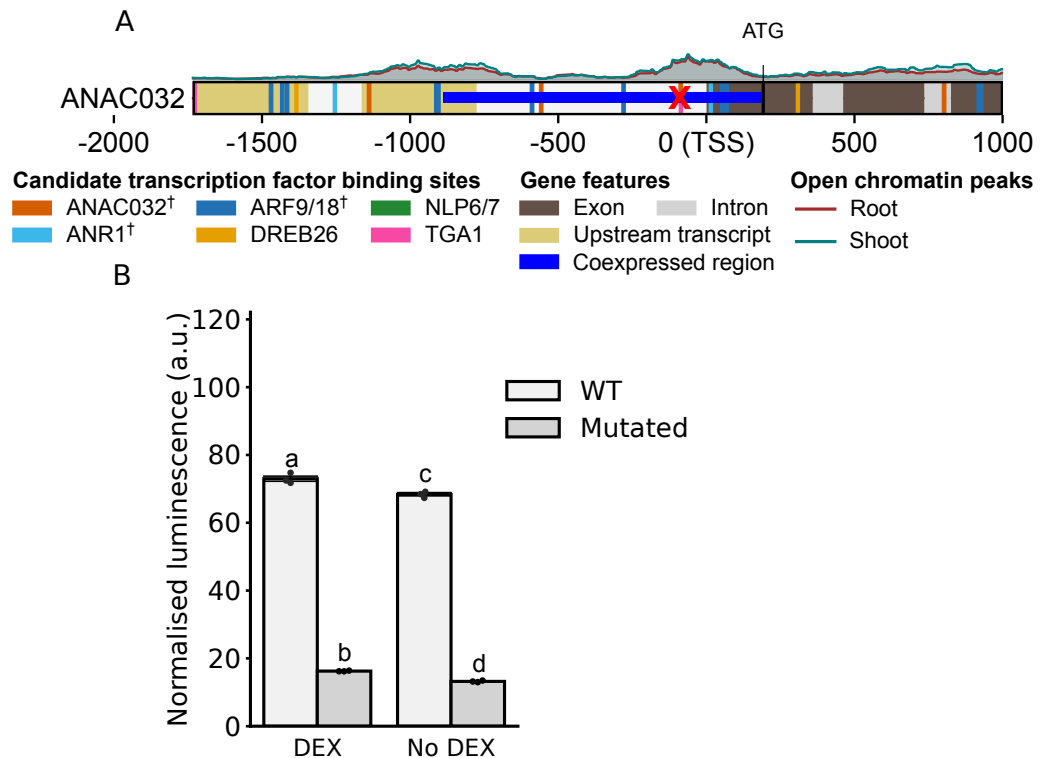


Figure 3.21: **Normalised luminescence of $p_{ANAC032}$ -LucN with a mutated TGA1 TFBS following co-expression with TGA1 TFs.** A: Position of ablated TGA1 motif (red cross) in the *ANAC032* promoter, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B: Normalised luminescence of wild type $p_{ANAC032}$ -LucN or version containing the ablated TGA1 TFBS when co-expressed with $p_{CaMV35s}$ -TGA1::GR, with and without 10 μ M DEX treatment. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3$ (technical replicates). Significance calculated using ANOVA and Tukey HSD post hoc tests³⁴⁴. Different letters indicate significant differences between treatments.

3.5.17 Investigating the role of NLP binding sites in the *NIR1* upstream promoter and untranslated region.

The NRE motif in the *NIR1* 5' UTR is a well-known 43 bp nitrate response element which binds NLP TFs, which are known to activate *NIR1*³⁹⁶. The NRE is a pseudo-palindromic sequence with two half-sites separated by a 10 bp spacer, with both half sites necessary for full nitrate induction of *NIR1*³⁹⁶. The *NIR1* NRE half-sites were mutated as in Konishi

and Yanagisawa³⁹⁶, to test if *NIR1* activation by NLPs would be disrupted (fig. 3.22A).

Luminescence of the mutated p_{NIR1} -LucN with ablated NRE was significantly lower than that of the WT promoter when co-expressed with NLP6 or NLP7 (WT: NLP6: $P < 0.01$, 3.1 ± 0.0 ; NLP7: $P < 0.01$, 4.6 ± 0.0 ; Mutated: NLP6: 0.02 ± 0.0 ; NLP7: 0.03 ± 0.0) (fig. 3.22B and C). There was no significant difference in luminescence of the mutated p_{NIR1} -LucN with the no TF control compared to the WT promoter with the no TF control ($P > 0.05$) (fig. 3.22B and C). There was also no significant difference in luminescence of the mutated promoter with and without co-expression of NLP protein ($P > 0.05$).

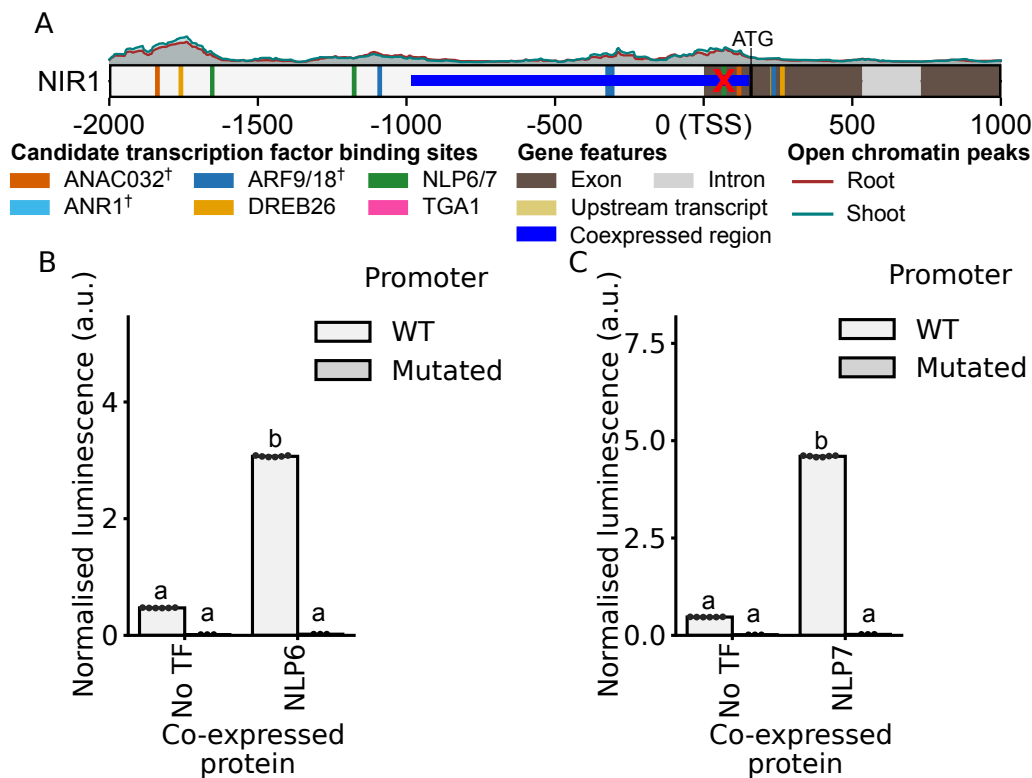


Figure 3.22: **Normalised luminescence of *NIR1*-LucN with a mutated NLP TFBS following co-expression with NLP TFs.** A: Position of ablated NLP (NRE) motif (red cross) in the *NIR1* 5' UTR, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B–C: Normalised luminescence of wild type p_{NIR1} -LucN and version containing the ablated NRE motif when co-expressed with $p_{CaMV35s}$ -NLPs. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3-6$ (technical replicates). Significance calculated using ANOVA and Tukey HSD post hoc tests³⁴⁴. Different letters indicate significant differences between treatments.

3.5.18 Investigating the role of ANAC032 binding sites in the *NLP7* promoter

As mentioned previously, the luminescence of p_{NLP7} -LucN was significantly reduced by ANAC032 (fig. 3.17B). To test if this response could be disrupted, candidate ANAC032 TFBSs were mutated in the *NLP7* promoter (fig. 3.23A). When co-expressed with ANAC032, the mutated p_{NLP7} -LucN luminescence was still significantly reduced (0.50 ± 0.01) compared to the no TF control (1.0 ± 0.01 ; $P < 0.01$; fig. 3.23B). However, raw luminescence of the mutated p_{NLP7} -LucN construct (1.0 ± 0.01) was significantly higher than that of the WT promoter (0.7 ± 0.0) when there was no co-expressed TF ($P < 0.01$; fig. 3.23B).

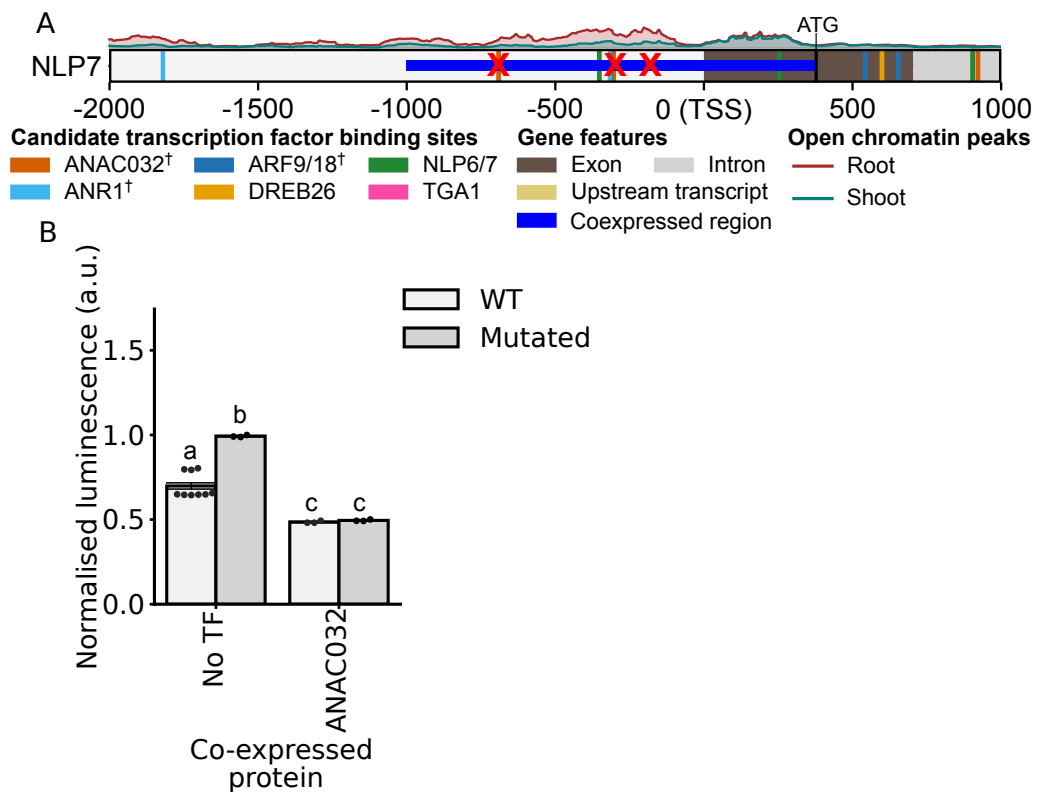


Figure 3.23: Normalised luminescence of p_{NLP7} -LucN with mutated ANAC032 TFBSs following co-expression with ANAC032. A: Position of ablated ANAC032 TFBSs in the *NLP7* promoter, with open chromatin peaks above (ATAC-seq³⁵⁰), and the co-expressed region in protoplast assays in blue. B: Normalised luminescence of WT and mutated p_{NLP7} -LucN when co-expressed with $p_{CaMV35s}$ -ANAC032. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. †, TFBS inferred from closely related TF, see section 3.5.1. $N = 3-9$ (technical replicates). Significance calculated using ANOVA and Tukey HSD post hoc tests³⁴⁴. Different letters indicate significant differences between treatments.

3.5.19 Reporter co-expression assays provide insights into indirect regulation within the subnetwork

Indirect interactions are more difficult to interpret from the co-expression assays since incoming edges from TFs outside of the subnetwork likely generate noise. However, from the knowledge of direct interactions within the network, it is possible to make some predictions about specific indirect interactions and use the data generated to test these. To test the hypothesis that ANAC032 indirectly represses *ANR1* via repression of NLPs, ANAC032 was co-expressed with p_{ANR1} -LucN. Co-expression data supported a slight indirect repression of *ANR1* by ANAC032, potentially through downregulation of NLP6 and NLP7. However, co-expression assays showed a reduction in luminescence of p_{ANR1} -LucN following co-expression with NLPs, but NLPs are known activators³⁹³. Indeed, overexpression of NLP7 increased *ANR1* expression in Arabidopsis and increased lateral root formation⁴⁰⁵, which ANR1 is known to do⁴⁷. This contradiction could be because important regulatory regions might have been missed in regions upstream and downstream of the region included in co-expression assays.

ANAC032 was co-expressed with p_{ARF18} -LucN to test the hypothesis that co-expression of ANAC032 would indirectly increase p_{ARF18} -LucN luminescence through downregulation of *NLPs*. ANAC032 indirectly increased p_{ARF18} -LucN luminescence, supporting this hypothesis. This was potentially through NLP6 which was upregulated by ANAC032 and which increased p_{ARF18} -LucN luminescence. However, unlike *NLP6*, *NLP7* was repressed by ANAC032 which demonstrates the complexity of the interactions in the subnetwork.

A previous study provides evidence that ANR1 is capable of directly activating its gene targets⁴⁰⁶. ANR1 was co-expressed with p_{ARF18} -LucN to test the hypothesis that co-expression of ANR1 would indirectly increase *ARF18* luminescence through activation of *NLPs*. Although co-expression of ANR1 did increase the luminescence of *NLPs*, it slightly decreased p_{ARF18} -LucN luminescence. Therefore, indirect regulation of *ARF18* by ANR1 is unlikely to be as important as other regulators of *ARF18*.

To test the hypothesis that ANR1 would indirectly activate *DREB26* through NLPs, ANR1 was co-expressed with p_{DREB26} -LucN. *NLP6* and *NLP7* were indeed activated by ANR1, and *NLP6* and *NLP7* also increased p_{DREB26} -LucN luminescence. Co-expression of ANR1

also increased p_{DREB26} -LucN luminescence, supporting this hypothesis. There was also evidence that ANR1 indirectly activated *NIR1* through NLPs, since co-expression of both ANR1 and NLPs increased p_{NIR1} -LucN luminescence.

ARF18 significantly increased the luminescence of p_{NLP7} -LucN with no auxin treatment, but reduced the luminescence of p_{NLP7} -LucN with 1 mM auxin. Since ARF18 is a known transcriptional repressor^{400,407}, the co-expression data suggest an indirect edge from ARF18 to *NLP7*. IAA33 slightly enhanced the activation of *NLP7* with no auxin treatment and enhanced the repression of *NLP7* in the presence of auxin. This is supported by the fact that IAA33 stabilises the activity of ARF repressors, preventing their degradation in the presence of auxin⁴⁰¹.

TGA1 was found in TARGET assays to directly repress *ANAC032*, *NLP6* and *NIR1*³¹⁸. Since no TGA1 TFBS were identified in the *NIR1* upstream region used in co-expression assays, it was possible to test the indirect regulatory effects on *NIR1* using the co-expression assay. To test the hypothesis that TGA1 would indirectly repress *NIR1* through repression of *ANAC032* (which activated *NIR1*), TGA1::GR was co-expressed with p_{NIR1} -LucN with DEX treatment. In the presence of DEX, TGA1::GR increased p_{NIR1} -LucN luminescence, and also increased $p_{ANAC032}$ -LucN luminescence. This contradicted the TARGET assay data³¹⁸ and did not support the hypothesis that TGA1 would indirectly repress p_{NIR1} -LucN luminescence following TGA1::GR co-expression. A possible reason for this contradiction is that some of the TGA1 data may be unreliable since the presence of the GR-tag may affect the activity of TGA1 (see section 3.5.14).

Summary

In summary, 22 direct edges were supported by TRAM *in vitro* interactions and co-expression assays (figs. 3.24 and 3.25). *ANAC032* formed a coherent type 2 feedforward loop (see fig. 1.1) to *NIR1* through *NLP7*. Mutating TFBSs in promoters significantly altered the expression of those promoters, with mutations causing a decrease in expression of *ANAC032* (figs. 3.20 and 3.21), *NIR1* (fig. 3.22) and causing an increase in expression of *NLP7* (fig. 3.23).

In general, ARF TFs repressed their targets, while NLP6 and NLP7 activated their targets apart from *ANR1*, and *ANAC032* generally activated target genes although it repressed

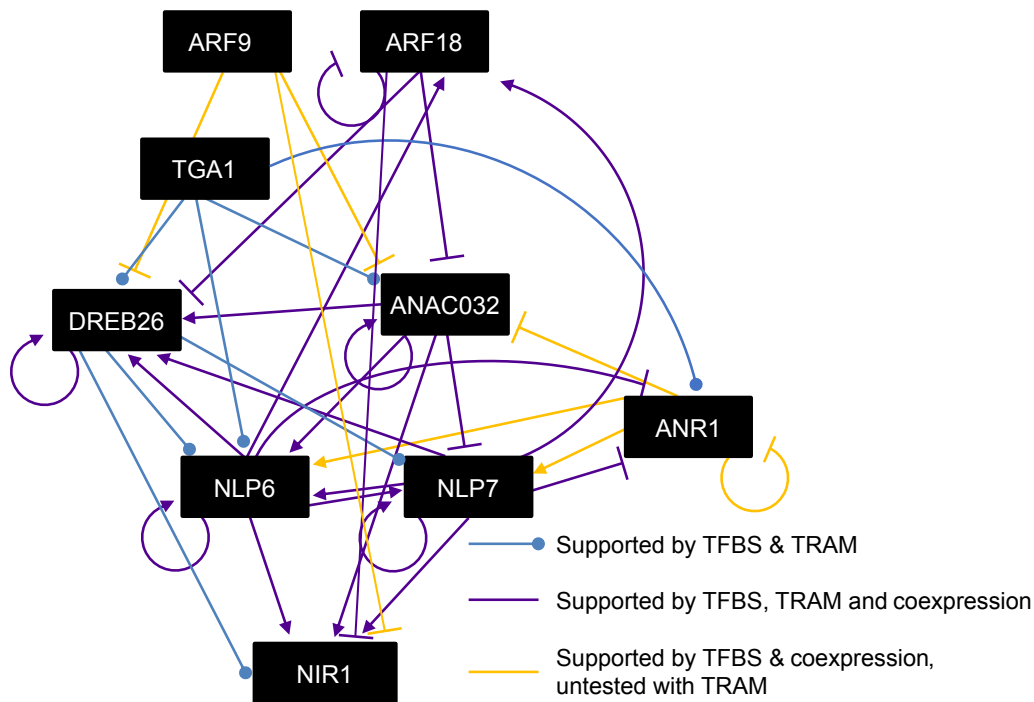


Figure 3.24: **Summary of significant regulatory interactions determined between TFs and their targets in the N-response regulatory subnetwork.** Significant edges ($P < 0.05$) determined from protoplast co-expression data which were supported by TRAM *in vitro* binding data and/or the presence of a candidate TFBS. Arrows represent upregulation of target, and perpendicular lines represent downregulation of target.

NLP7 (fig. 3.25). The addition of NAA when co-expressing ARF TFs with test promoters did not have much effect on the luminescence from the co-expressed promoters in most cases, other than slight increases in overall expression (fig. 3.11D–F, fig. 3.12F–H, fig. 3.13F–G and fig. 3.15F–H). The overall regulatory effect of ARF TFs was largely unaltered by the addition of NAA. However, in some cases the co-expressed promoter responded to NAA. The addition of NAA increased the luminescence from *DREB26*-LucN and *NLP6*-LucN with and without the co-expression of ARF TFs (fig. 3.14G–I). The luminescence from *NLP7*-LucN was increased by the addition of NAA on its own or when co-expressed with ARF9 (fig. 3.17F). However, the addition of NAA and co-expression with ARF18 decreased the luminescence of *NLP7*-LucN compared to the control, whereas without NAA *NLP7*-LucN luminescence was increased when co-expressed with ARF18 (fig. 3.17G).

The co-expression of ANAC032, NLP6 and NLP7 significantly increased the expression of the *CaMV35s* luciferase calibrator (fig. 3.9), which could create noise and obscure the interactions of interest in the various co-expression experiments. This means that only strong effects of these TFs on the expression of their targets could be reliably detected, such

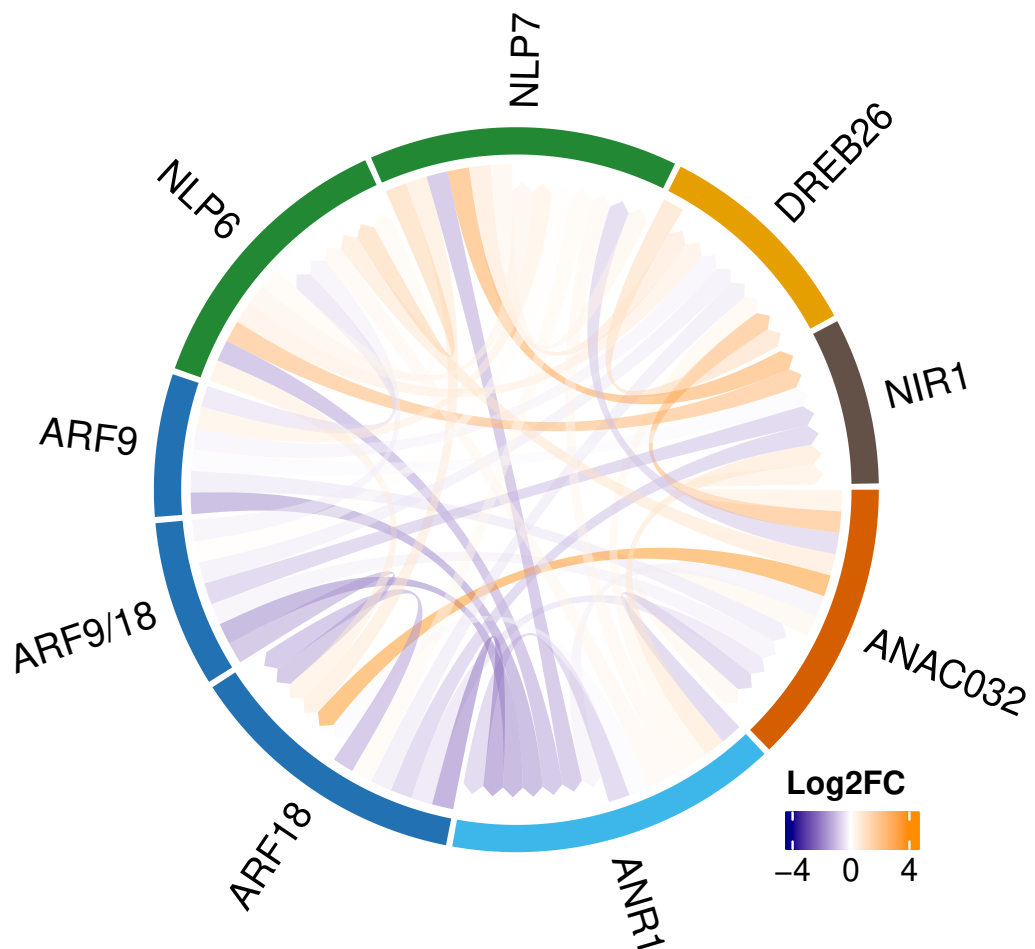


Figure 3.25: **Summary of significant edges determined between TFs and their targets in the N-response regulatory subnetwork using co-expression luciferase assays** Significant edges ($P < 0.05$) determined from protoplast co-expression data. Width of outside segments represent the sum of the number of incoming and outgoing edges. Only ARF TFs with no coexpressed IAA33 and no auxin are shown. Edges are coloured in orange when target was activated and in blue when target was repressed. Shades of colours are determined by log₂ fold change of target CRM luminescence when coexpressed with the TF of interest compared to the no TF control.

as the strong reduction in luminescence of *ANR1*-LucN when co-expressed with NLP6 and NLP7 (fig. 3.12D and E). The effects of the other TFs on the luminescence of the *AtuNOS* luciferase calibrator, although not as strong as the effect of ANAC032 and NLP6/7 on the *CaMV35s* luciferase calibrator, were significant (fig. 3.10) so only strong effects on luminescence from co-expression assays should be considered.

Interestingly, there was a correlation between the distance of the TFBS and the strength of the regulatory effect of the TF on the promoter, with effects being stronger the closer the TFBS was to the TSS. For example, the ANR1 TFBS was located in the *ANAC032* 5' UTR, 7 bp from the TSS (fig. 3.11A). The co-expression of ANR1 with *ANAC032*-LucN

resulted in a significant decrease in luminescence compared to the control (fig. 3.11C). Two ANR1 TFBSs in the *NLP6* promoter were located over 800 bp upstream from the TSS (fig. 3.16A), and the closest ANR1 TFBS to the TSS in the *NLP7* promoter was 307 bp upstream. In both of these cases, ANR1 increased the luminescence of the promoter-LucN (fig. 3.16C and fig. 3.17C), which was the opposite effect as for the ANR1 TFBS close to the TSS in *ANAC032*.

3.6 Discussion

3.6.1 Reporter co-expression assays support direct edges within the sub-network

The aim of this chapter was to gain an understanding of the edges in the subnetwork of TFs identified by Gaudinier et al.²⁵. 22 edges were supported by co-expression and TRAM assays, and additionally 7 edges were supported by co-expression whose targets contained a candidate TFBS that was not tested with TRAM.

In general, ARF9 and ARF18 repressed their targets in protoplast co-expression assays, consistent with the known role of these TFs as repressors^{400,407}. ARF9 and ARF18 reduced p_{DREB26} -LucN luminescence when there was no auxin present, as did ARF9 and 18 when co-expressed together regardless of auxin concentration. ARF9 and 18 also significantly reduced $p_{ANAC032}$ -LucN luminescence when co-expressed.

Protoplast co-expression assays largely agreed with RNA-seq data. For example, in a *NLP7* T-DNA insertion mutant *ANR1* expression was increased²⁵, which supports the co-expression assay result of *NLP7* repressing *ANR1*. When ARF18 was overexpressed, the expression of *NIR1*, *NLP7* and *NLP6* was reduced (unpublished, see Gene Expression Omnibus GSE121222), which supports the co-expression assay results of ARF18 repressing these genes. When *ANAC032* was overexpressed, RNA-seq data showed that *DREB26* was upregulated compared to wild type plants⁴⁰⁸, which supports the co-expression assay result of *ANAC032* activating *DREB26*.

NLPs are known to directly activate *NIR1*^{26,396}, so the activation of *NIR1* by *NLP6* and *NLP7* in the co-expression assay was expected. *ANAC032* reduced p_{NLP7} -LucN luminescence when co-expressed. The predicted feedforward loop from ARF18 to *NLP7*

via DREB26 and ANAC032 was only partially supported by protoplast co-expression data. ARF18 to *NLP7* via ANAC032, and ARF18 to *DREB26* were supported, but DREB26 to *NLP7* could not be tested in co-expression assays since the predicted DREB26 TFBS was located in *NLP7* exon 1 which was not included in the p_{NLP7} -LucN plasmid. DREB26 TFBSs were often not present in the tested regions in co-expression assays as they were often located downstream of the ATG. Larger regions which include these DREB26 TFBSs need to be tested in co-expression assays to find outgoing DREB26 edges.

ARF18 was activated by NLP6 and NLP7 in the co-expression assay. This suggests presence of a positive feedback loop from ARFs to *NLP7* when no auxin is present via repression of *ANAC032*. There was evidence of autoregulation of *ARF18*, *ANAC032*, *DREB26*, *NLP6*, *NLP7* and *ANR1*, which in the case of positive autoregulation (NLPs, DREB26) could potentially slow down network responses and amplify noise⁴⁰⁹ and, in the case of negative autoregulation (ARF18, ANR1), could lead to a faster response time and increased robustness⁴¹⁰.

Co-expression of ANAC032 or NLPs increased luminescence of p_{DREB26} -LucN. ARF9/18 TFs reduced p_{DREB26} -LucN luminescence in the co-expression assay when there was no auxin, but there were mixed effects in the presence of auxin with an activation by ARF9, and a repression by ARF9/18/IAA33 suggesting complex regulatory interactions and feedback which would need investigating further to understand. ARF TFs are split into three classes, with activators in class A, and repressors in classes B (which includes ARFs 9 and 18) and C⁴⁰⁰. The addition of NAA did not have a large effect on the co-expression assay results with ARF TFs, potentially because the saturation with ARF TFs obscured any effects of auxin on the regulatory activity of ARF9 and ARF18. It has been proposed that ARF activators compete for DNA-binding with ARF repressors⁴¹¹. *DREB26* is strongly activated by auxin⁴¹², so class A ARF activators might activate *DREB26* in the presence of auxin, partly mitigating the repressive effect of the repressors ARF9/18. However, if this was the case then, upon overexpression of ARF9 or ARF18, it might be expected that *DREB26* would be repressed because class A activators should be diminished due to saturation with ARF9 or ARF18. However, *DREB26* was activated indirectly by ARF9 and non-significantly by ARF18 in the presence of auxin. It is possible that class A activator ARFs bind with higher affinity to the TFBSs in *DREB26* than

class B ARFs 9 and 18. ARFs bind to 6 bp auxin response elements with the sequence TGTCNN. Class A ARFs can bind to pairs of auxin response elements in inverted repeat, everted repeat, and direct repeat orientations, while class B ARFs mainly bind to inverted repeats^{181,398}. Additionally, class A ARFs bind with highest affinity to inverted repeats with 7–8 bp and 17–18 bp spacing between TGTCNN motifs, while class B ARFs only bind with high affinity to inverted repeats with 7–8 bp spacing³⁹⁸. Class A ARF dimers can bind to inverted repeat motifs with a wider range of spacing between motifs than class B ARFs. Binding affinity of a class A ARF dimer (ARF5) only reduced slightly when spacing between motifs was changed to 5, 6 or 9 bp, while binding affinity of a class B ARF dimer (ARF1) was greatly reduced compared to spacing of 7 or 8 bp⁴¹³. There is a pair of direct repeat auxin response elements 39 bp upstream of the TSS in the *DREB26* promoter with 11 bp spacing between them. This suggests that class A ARFs might bind to *DREB26* with much higher affinity than class B ARFs, which could explain why *DREB26* expression is induced in the presence of auxin even when co-expressed with ARF9 or ARF18, since the auxin mediated derepression of class A ARFs means that class A ARFs outcompete ARF9 and 18 for binding with the *DREB26* promoter. In the absence of auxin, class A ARFs are repressed by aux/IAA proteins so ARF9 and 18 can bind with low affinity and repress *DREB26* when they are overexpressed. Interestingly, when both ARF9 and 18 are co-expressed at the same time, *DREB26* is repressed in the presence of auxin. It has been suggested that ARFs could form alternative dimer complexes allowing for interaction with different DNA motifs with orientation and spacing not bound by ARF homodimers⁴¹⁴. ARF9 and 18 are known to form heterodimers with each other⁴¹⁵, which could alter their DNA-binding domain structure compared to ARF9 or 18 homodimers, potentially making the ARF9/18 heterodimer more promiscuous and able to bind to the direct repeat motifs in the *DREB26* promoter with higher affinity. Further experiments testing the affinity of ARF9 and 18 heterodimers compared to homodimers to different motifs using protein-DNA binding affinity assays such as TRAM are needed to confirm this hypothesis.

The use of a $p_{CaMV35s}$ -YFP control plasmid in place of co-transfection with $p_{CaMV35s}$ -TF has not been demonstrated in previous ratiometric protoplast co-expression assays³⁹³, but the data obtained suggest that this method should be adopted to ensure balanced transcriptional load. Ratiometric assays are dependent on steady expression of an experimental and batch calibrator. In an initial test, *CaMV35s* and *AtuNOS* promoters were

found to change expression in response to the presence of some TFs. Ideally, a wider range of endogenous or synthetic constitutive promoters should be tested to identify a single promoter that could be used for all TFs.

Protoplast co-expression assays in this chapter contained technical replicates but not biological replicates. During protocol development, similar experiments showed the same overall trends in luminescence (data not shown), but these cannot be directly compared to the experiments in this chapter as the experimental conditions were not identical, and the final experimental calibrators and controls had not been selected. It will be important in the future to repeat the protoplast co-expression experiments with the same experimental conditions to confirm the findings.

3.6.2 TF-DNA interaction data supports the architecture of the subnetwork

The identification of candidate binding sites in the upstream promoters and UTRs of *ANAC032*, *ANR1*, *ARF18*, *DREB26*, *NLP6*, *NLP7* and *NIR1* provided some support for the interactions between TFs and target promoters in the subnetwork identified using Y1H²⁵. Y1H edges from ARF9/18 to *ANAC032*, *ANR1*, *DREB26* and *NLP6* were supported by the identified candidate TFBSs. Additionally, edges from ARF9 to *NIR1*, *NLP6/7* to *NIR1*, TGA1 to *NLP6*, *ANAC032* to *NLP7*, *DREB26* to *ANAC032* and *DREB26* to *DREB26* and *NLP7* were supported by the identified candidate TFBSs. However, the Y1H edge from *DREB26* to *ANR1* was not supported since no candidate *DREB26* TFBSs were identified in the *ANR1* promoter/5' UTR. Additional candidate TFBSs for edges not identified in Y1H data were found, such as a *DREB26* TFBS in the *NIR1* exon 1 and a TGA1 TFBS in the *ANAC032* promoter. This is most likely due to the test promoters in the Y1H assay not capturing all of the TFBSs, and because false negatives can occur in Y1H assays due to improper folding or localisation, or binding of TFs to TFBSs might require post-translational modifications that only occur *in planta*²⁵. It is also possible that not all candidate TFBSs were identified. For example, transient binding of bZIPs²²⁴ and *NLP7*²⁶ can lead to TFBSs being missed in ChIP-seq and DAP-seq data. Further, since the TFBSs of *ANAC032*, *ANR1* and *ARF9/18* were not in the DAP-seq database¹⁸¹, their TFBSs were inferred from closely related TFs. Although subsequent data from TRAM found that the TFs could bind to these sites, it is possible that some TFBSs may not

have been identified in promoters/5' UTRs of genes in the subnetwork. Finally, some TFs have been shown to bind to alternate motifs when in a heterodimer with another TF⁴¹⁶. At present, the incomplete datasets for plant TFBSs hinder the ability to identify all candidate sites.

Candidate binding sites were experimentally tested using TRAM. Only 8/20 tested TFBSs were found to bind ANAC032, suggesting that the inferred TFBS from ATAF1/ANAC002 may not be the optimal DNA sequence for binding of ANAC032. Alternatively, ANAC032 may require an additional protein to bind with strong affinity. The TRAM assay could not be performed for ANR1 or ARF9 since these proteins were not successfully expressed. 10/29 DNA probes were significantly bound by ARF18. The low proportion of bound probes could be because class B ARFs such as ARF18 mainly bind to inverted TGTCNN repeats^{181,398} with 7–8 bp spacing³⁹⁸, and most of the identified and tested TFBSs were not inverted TGTCNN repeats. 6/11 DREB26 TFBSs were found to bind the DREB26 protein using TRAM. Since this TFBS is a repetitive sequence, there was high potential for false positive candidate TFBSs when promoter regions were scanned using FIMO³⁶⁴. NLP6 and NLP7 were found to bind 12/17 and 13/17 TFBSs respectively, indicating a potential difference in the binding preferences of these two TFs. 7/8 DNA probes containing TGA1 candidate TFBSs were bound by TGA1 in TRAM assays, indicating that the position weight matrix used for the TGA1 TFBS was good at predicting true binding events. More data is needed to find the true TFBS core and flanking region for optimal binding of TFs in the subnetwork, and to identify if there might be alternative binding sites that are recognised by heterodimers with other TFs from the wider N-response network.

These data, together with the existing Y1H data²⁵ provided experimental evidence of physical interactions between TFs and target genes. However, they do not provide information about any regulatory consequences of binding, which is why the reporter co-expression assay was employed.

3.6.3 Reporter co-expression assays are a useful tool for understanding TF function

A strength of reporter co-expression assays is that they can be used to test potential direct edges once candidate TFBSs have been identified and confirmed *in vitro* to bind

TFs in a variety of contexts. They can complement other methods such as TARSET and TARGET, providing another line of evidence of direct edges. Additionally, they can test incoming edges to genes which respond to cycloheximide and cannot be reliably tested using TARSET and TARGET. For example, incoming edges to *DREB26*, which has been shown to respond to cycloheximide³¹⁸, were successfully tested using co-expression assays in this chapter. However, a limitation of co-expression assays is that the tested promoter region is limited in size and may not capture the true regulatory effects since TFBSs falling outside of the tested region may be important. This was an issue especially for testing regulatory targets of DREB26, since several DREB26 TFBSs were located outside of the tested region in protoplast assays. Additionally, the tested region in co-expression assays is unlikely to include the native chromatin state, which could lead to incorrect conclusions of the regulatory effects. However, disrupted chromatin state is also likely a limitation of TARGET and TARSET assays. Isolation of plant protoplasts leads to genome-wide increases in chromatin accessibility, leading to differentially expressed genes in protoplasts compared to intact plants⁴¹⁷. Overexpression of test TF in whole plants through *Agrobacterium* leaf infiltration or delivery using nanoparticles^{418–420} followed by transcriptomics analysis could be used instead in the future to overcome this limitation, although these methods are lower throughput and less scalable.

The following direct edges were supported by both co-expression and TARSET data (unpublished, Tufan Oz): ARF18 to *ANAC032*, ANAC032 to *NLP7*, NLP6 to *NIR1* and NLP7 to *NIR1*. Additionally, TGA1 to *ARF18* was supported by TARGET data³¹⁸. In some cases, co-expression data represented the indirect regulatory effects found in TARSET assays. For example, ANR1 was found to indirectly activate *NLP7* in both co-expression and TARSET assays. This could be explained by repression of *ANAC032* by ANR1, which potentially led to derepression of *NLP7*, since ANAC032 decreased luminescence of p_{NLP7} -LucN following co-expression.

Co-expression assays also allow the investigation of the role of specific TFBSs by introducing mutations. The *ANAC032* promoter/5' UTR with mutated ARF TFBSs showed increased luminescence following co-expression of ARF18/IAA33 compared to no TF. This was the opposite trend to the wild type $p_{ANAC032}$ -LucN luminescence which was decreased following co-expression with ARF18/IAA33. Additionally, the luminescence of mutated

$p_{ANAC032}$ -LucN was $>4\times$ lower compared to the WT $p_{ANAC032}$ -LucN. This provides a further line of evidence that *ANAC032* is a direct target of ARF TFs.

When the NRE motif in *NIR1* was mutated, NLPs still significantly activated NIR1 although luminescence was much lower overall compared to that from the WT p_{NIR1} -LucN, showing the importance of the NRE for correct NIR1 expression. These results supported those in Konishi and Yanagisawa³⁹⁶, where mutations in the NRE significantly reduced *NIR1* expression.

Overall, mutating TFBSs in promoters significantly reduced the expression of those promoters, showing that the TFBSs are important for correct expression of the genes, and that mutations in TFBSs can be used to change expression levels and potentially rewire GRNs.

Reporter co-expression assays provide useful tools for identifying potential regulatory interactions. However, complementary experiments such as the TARSET/TARGET assays can potentially provide better information as target genes are within the native genomic context. In addition, TARSET/TARGET assays can distinguish between direct and indirect interactions. However, comparisons of data from TARSET/TARGET with co-expression data also highlighted potential issues with TARSET/TARGET assays. When TGA1 was co-expressed with test promoters/5' UTRs, it resulted in the activation of all genes. This contradicted TARGET data in Brooks et al.³¹⁸, where many of the same target genes were repressed by TGA1. This could be due to the GR tag used in the TARGET assay, which might have interfered with the binding of TGA1 to DNA or cofactors. TGA1::GR reduced p_{ARF18} -LucN luminescence upon application of DEX in co-expression assays which included the GR tag. A similar reduction in luminescence was observed after DEX treatment when TGA1::GR was co-expressed with p_{NLP7} -LucN, although overall NLP7 expression was higher with TGA1::GR than with the no TF control. However, when TGA1::GR was co-expressed with $p_{ANAC032}$ -LucN and p_{NIR1} -LucN, it still resulted in an increase in luminescence upon application of DEX, suggesting that the GR tag did not interfere with TGA1 function in the TARGET assay. Additionally, when a candidate TGA1 TFBS was mutated in the *ANAC032* promoter, co-expression of TGA1::GR still increased the mutated $p_{ANAC032}$ -LucN luminescence, suggesting that the native genomic context of *ANAC032* is required for its correct regulation by TGA1. The luminescence of the mutated

$p_{ANAC032}$ -LucN was significantly lower than the WT $p_{ANAC032}$ -LucN regardless of DEX concentration, showing the importance of the TGA1 TFBS. However, the disruption of the TGA1 TFBS also disrupted an ANAC032 TFBS, which could have multiplied the effect of the mutation and makes the interpretation ambiguous. Since protoplasts were harvested 3 hours after DEX treatment during co-expression with TGA1::GR, as opposed to after overnight incubation with TGA1 in the original co-expression experiment, the differences in regulation of test promoters might be due to the time of harvest. This suggests that TGA1::GR represses genes such as *ARF18* in the short term after 3 hours, but after 14 hours there is an overall activation effect, potentially due to feedback loops. It would be useful in the future to conduct time course experiments to compare the regulatory consequences of test TFs on their target genes over time. However, a compelling reason for the discrepancy between the TARGET and luciferase assays is that in targets repressed by TGA1, TFBSs are mainly located in the CDS, a region which was not included in the CRM in protoplast co-expression assays³¹⁸. In induced TGA1 targets, TGA1 TFBSs were mainly located in the promoter region 0.5 kb upstream of the TSS³¹⁸. For genes with enriched TFBSs in regions downstream of the promoter and 5' UTR, alternative assays to protoplast co-expression such as TARGET would be more appropriate for testing repression effects.

ANAC032 still significantly reduced luminescence of p_{NLP7} -LucN when candidate ANAC032 TFBSs were ablated in the *NLP7* CRM. The luminescence activity of the mutated p_{NLP7} -LucN was higher than the WT promoter, but was not significantly different from the WT promoter following co-expression with ANAC032, suggesting an additional mechanism of repression of *NLP7* by ANAC032, potentially via undiscovered TFBSs, through an interacting cofactor which binds different TFBSs to ANAC032, or indirectly through ANAC032 regulating another TF which then regulates *NLP7*.

A potential limitation of overexpressing TFs in protoplasts is that any additional factors that are either required for or contribute to the normal regulatory activity of specific TFs via protein-protein interactions may not be present at sufficiently high quantities. For example, TCP20 has been shown to interact with NLP7 with the resulting heterodimer having a role in root meristem growth under N-starvation⁶⁴. Overexpression of NLP7 is likely to result in more NLP7 homodimers than NLP7-TCP20 heterodimers. Similarly, ARF9/18 have

been shown to dimerise with each other⁴¹⁵ and to interact with aux/IAA proteins, which repress the activity of ARFs when the auxin concentration is low³⁹⁹. As there may have been insufficient aux/IAA proteins in the protoplasts for any such interactions to occur, protoplasts were supplemented with NAA and IAA33. This did not have a large effect on the activity of ARF9/18, however, and it is not certain that IAA33 is the correct partner for ARF9/18. Although these proteins have been shown to interact *in vitro*⁴⁰¹, further data are required to determine if this happens in plant roots. It will be important in the future to study more TF-DNA interactions and regulatory effects with dual luciferase co-expression assays along with complementary TARGET assays to study both short- and long-term responses of plants to nitrate to expand the validated GRN. Large scale experiments are needed to fully understand the GRN, but targeted experiments such as mutation of TFBSs in the promoters of important N-response TFs to disrupt specific edges, for example using CRISPR, could be used to change plant nitrate responses. Testing the effect of mutations in *cis*-regulatory regions, such as the mutated ARF TFBSs in the *ANAC032* CRM, in the native context would be useful so that transcriptomics and phenotyping assays could be conducted to assess the effect on N-use efficiency.

Some genes such as *DREB26* have been shown to respond to cycloheximide³¹⁸, therefore, alternative methods might be required. Although co-expression assays can provide supporting evidence for direct regulation when paired with evidence of TF-DNA binding, they cannot confirm direct regulation since native TFs are still active. One potential solution could be to sample several timepoints following induction of DEX to identify suites of genes that are more likely to be directly or indirectly regulated by the test TF, rather than relying on cycloheximide to distinguish direct interactions³⁹².

3.6.4 Conclusions

In summary, TRAM TF-DNA interaction data supported 29 edges in the subnetwork. Protoplast co-expression data was successfully used to characterise edges in the N-response subnetwork and test whether they activate or repress targets. 29 TF-promoter edges (excluding TGA1 outgoing edges) were confirmed to significantly alter the expression of target genes, with support for 22 direct edges when combined with confirmation of binding with TRAM. Protoplast co-expression assays can only putatively support direct regulatory effects, and it is beneficial to complement them with TARGET or TARSET assays. The

promoters/5' UTRs used in protoplast co-expression assays miss potentially important TFBSs in upstream and downstream regions, and do not contain native chromatin states. Despite these limitations, the data are useful for understanding the overall gene regulatory subnetwork and predicting the effects of perturbation or manipulation. Additionally, protoplast co-expression data are useful for evaluating and identifying network motifs. For example, the predicted ARF18 to *NLP7* feedforward loop through ANAC032 and DREB26 was supported. In combination with TF-DNA binding affinity data such as TRAM, and complementary TF overexpression assays such as TARGET, the protoplast co-expression data will allow targeted manipulation or engineering of promoter sequences, altering the expression of genes in the N-response subnetwork and causing changes to network dynamics. This will potentially allow for the improvement of beneficial traits, such as nitrogen-use efficiency. So far, data from this chapter have been used to identify TFBSs in the regions of promoters that are mutated using CRISPR approaches in chapter 4, and to inform the design and engineering of a genetic feedback controller in the subnetwork (chapter 6).

Chapter 4

Development of a CRISPR library for engineering variation into *cis*-regulatory sequences

4.1 Preface

My contribution to this chapter included all plasmid design and assembly, CRISPR library construction, plant transformation, sequencing library preparation, most of the sequencing analysis, all variant calling and mutation analysis, and phenotyping of mutant lines. Genomics Pipelines (Earlham Institute) performed PacBio sequencing, and Dr. David Swarbreck (Earlham Institute) performed the CCS and demultiplexing of samples. Primers for qRT-PCR were obtained from Dr. Tufan Oz (Earlham Institute).

4.2 Introduction

There is a need to accelerate crop improvement to meet future food demands of the human population⁴²¹. Application of nitrogen fertilisers increases yield, but overuse can cause eutrophication²⁹ and release of the greenhouse gas nitrous oxide³⁰. To reduce the need for nitrogen fertilisers, the development of crops with more efficient nitrogen-use efficiency is desirable. The regulation and integration of plant responses to stimuli and nutrients such as nitrate are coordinated by complex GRNs made up of TFs and their target genes⁴.

Manipulation of GRNs has potential to improve traits such as yield, growth, and nitrogen (N)-use efficiency. Once a GRN has been characterised, it can be analysed to identify hubs of highly connected TFs or nodes, which can indicate network robustness and provide targets for engineering⁴.

However, loss-of-function mutations in TF hubs can disrupt important biological processes and lead to undesirable traits⁴²². In contrast, disrupting specific edges between TFs and their targets by introducing mutations that disrupt TFBSs may avoid pleiotropic effects and enable the subtle tuning of network responses.

Improving crops has traditionally relied on selective and mutation breeding, but this is time-consuming and labour-intensive. Precision genome editing, where precise changes are made at specific loci without incorporating DNA into the genome, has the potential to simplify the plant development process^{423–425}. Precision genome editing using CRISPR/Cas systems has been used to introduce mutations into the genomes of many species by adding double strand breaks and relying on errors in plant repair mechanisms⁴²⁶. In plants, the most common mechanism for repairing double strand breaks is NHEJ; however, in dividing cells HDR is most dominant where it repairs stalled replication forks²⁶⁴. HDR is relatively inefficient in plants, especially due to the difficulties of delivering sufficient quantities of repair template into the cells⁴²⁷. The NHEJ pathway is more error prone than HDR and can occasionally lead to insertions and deletions. Although the exact mutation introduced by NHEJ repair cannot be controlled using wild type Cas9 proteins, more control has been enabled via the use of base editors, in which deaminase domains are fused to inactivated Cas proteins^{428,429}, and prime editors, in which a fused reverse transcriptase domain uses a modified guide RNA as a template for a specific repair^{430,431}.

Many important mutations driving crop domestication, breeding and evolution have been identified in *cis*-regulatory regions⁴³². These *cis*-regulatory mutations often cause subtle phenotypic changes by altering timing, duration, or amplitude of gene expression. Deletions in *cis*-regulatory regions can significantly alter gene expression³⁰⁶. In contrast, mutations in CDSs result in loss-of-function or alterations to protein structure¹¹². However, using CRISPR to make specific edits in *cis*-regulatory regions requires knowledge of which TFBSs or edges to disrupt. Since promoters are complex, it is difficult to predict the effect of a mutation on gene expression, so approaches that do not rely on extensive knowledge of

the *cis*-regulatory connections provide useful alternatives.

Forward genetic screening is a powerful approach for identifying regulatory or gene functions. This approach requires a population with genetic variation, and a method with which to identify individuals with phenotypes that differ from the wild type⁴³³. In plants, mutagenesis techniques have largely involved physico-chemical mutagens such as ethylmethane sulphonate, gene silencing or T-DNA transformant populations⁴³³. The CRISPR-Cas9 system has commonly been used to generate large libraries for screening in animals^{354,434–439}. However, there are relatively few examples of CRISPR-Cas9 libraries in plants^{440–442}. To date, CRISPR libraries have been used to create libraries of knockout lines in tomato⁴⁴³, rice^{440,441}, maize⁴⁴² and soy⁴⁴⁴. CRISPR libraries have also been used for ‘CRISPR directed evolution’ of specific target genes in rice^{445–447}.

Few studies have targeted *cis*-regulatory regions using pooled CRISPR library approaches. In tomato, eight sgRNAs were used to target a 2 kb region in the promoter of *SiCLAVATA* (*CLV*) β , which encodes a signalling peptide known to be involved in controlling meristem size⁴⁴⁸. This generated a range of mutants with varying locule number and fruit size in a mutant collection of 1152 F1 plants³⁰⁶. In wheat, a pooled sgRNA CRISPR library approach was used to target the promoter of the *Q* gene, which is involved in major domestication traits in wheat, but deletions in the promoter region did not change expression levels or phenotype significantly⁴⁴⁹. In rice, a CRISPR library approach was used to target the promoter, 5' UTR and downstream region of *ideal plant architecture 1* (*IPA1*), generating a range of mutants with variations in tiller number and panicle size, including a mutant containing a 54 bp deletion in the promoter which increased both panicle number and size, enhancing grain yield³⁰⁷.

In all the above-mentioned studies, Sanger sequencing of PCR amplicons was used to identify the mutations. An alternative approach is to use pooled barcodes with multiplexed amplicon sequencing. This approach allows for larger scale experiments⁴⁵⁰.

In chapter 3, the elucidation of a regulatory subnetwork is described for nitrate-responses consisting of eight TFs and the *NIR1* gene. Within that subnetwork, NLP7 is an intracellular nitrate sensor, an early responder to nitrate, and is known to be a master regulator in the nitrate response^{26,49}. NLP7 promotes lateral root development by activating *TAR2*⁵². ARF9 and ARF18 have been shown to regulate nitrogen-related genes such as *nitrate*

transporter 2.4 (NRT2.4) and *ANAC032*²⁵. Another TF, DREB26, was shown to affect nitrate assimilation²⁵. In chapter 3, evidence showed that ARF TFs regulate *DREB26*. Y1H data suggest that DREB26 regulates *NLP7*²⁵. Further, evidence from chapter 3 showed that ARF9 and 18 are involved in a feedforward loop to *NLP7* through *ANAC032* and potentially through DREB26. *ANAC032* was found to repress *NLP7* (see fig. 3.1). A hypothesis is that disrupting edges in the feed forward loop will affect the network dynamics and the response to nitrate.

Loss-of-function mutants of these transcription factors is undesirable as a strategy for engineering NUE because they may have roles in other tissues and processes. For example, in aerial tissues, ARF18 has a role in shade avoidance⁴⁰⁷. It is desirable to only disrupt the specific subnetwork. One way to achieve this is to disrupt TFBSs in the *cis*-regulatory regions of target genes. This requires knowledge of all TFs, and other proteins that bind to the target promoters. However, although progress has been made at characterising this subnetwork, the dataset is likely to be incomplete. Therefore, a CRISPR library approach was used to add genetic variation to the promoters of *ARF9*, *ARF18*, *DREB26* and *NLP7*, four TFs we predict to be important in the N-response in Arabidopsis.

In this chapter, the construction of a pooled CRISPR library is described which targets these promoters as well as the use of a pooled barcode amplicon sequencing approach to identify mutations. Five plant lines containing interesting mutations, several of which overlapped candidate TFBSs, were characterised using qRT-PCR and RSA phenotyping.

4.3 Aims

The first aim of this chapter is to generate a library of mutants by adding variation into the promoter and 5' UTR regions of four genes encoding TFs involved in regulating root growth in response to nitrate using a CRISPR library approach. The second aim is to study the effect of disrupting interactions between TFs and their candidate genes in *cis*-regulatory regions on transcriptomics and phenotype.

4.4 Results

4.4.1 Selection of genetic regions

In chapter 5, I show that the region closest to the TSS or start codon is more important for regulatory function than regions further upstream, as architectural differences here explained the different expression patterns of different gene categories. This is supported by observations in Yu et al.⁴⁵¹, where TFBSs in Arabidopsis were enriched in regions closest to the TSS compared to more distal regions, and also in Cai et al.²⁷⁸, where moving TFBSs closer to the TSS led to an increase in gene expression. Additionally, promoters containing more open chromatin were more actively expressed, and those with motifs falling within open chromatin were more important. Therefore, *cis*-regulatory regions 430–832 bp upstream of the TSS falling within open chromatin, and the 5' UTR were selected to increase the chance of disrupting important edges. ATAC-seq data³⁵⁰ (see section 2.7.7 Open chromatin coverage) were used to annotate the upstream regulatory regions falling within open chromatin in *ARF9* (1078 bp), *ARF18* (651 bp), *DREB26* (854 bp) and *NLP7* (1006 bp) (fig. 4.1, turquoise region) (see section 2.6.1).

4.4.2 Identification of targets

SpCas9 NGG PAM sites were identified using CRISPOR³⁵¹. Protospacer sequences with on-target efficiency³⁵⁴ of <40 % were filtered out and 30 protospacers per gene were identified (see section 2.6.1; fig. 4.1, red lines).

To enable the possibility of larger deletions^{452–454} and increase the likelihood that each line would contain at least one mutation, protospacers were sorted into pairs roughly 90–110 bp apart, with each protospacer allocated into at least one pair (see section 2.6.1). In total, there were 96 protospacer pairs, with 23 pairs in *ARF9*, 24 pairs in *ARF18*, 23 pairs in *DREB26* and 26 pairs in *NLP7*.

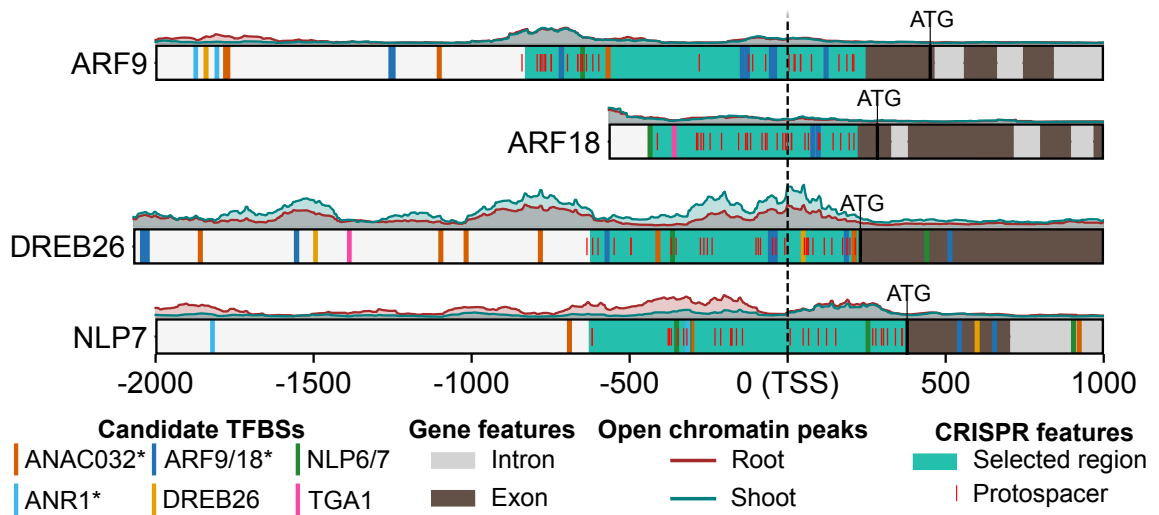


Figure 4.1: **Selected genetic regions for CRISPR targeting.** The *cis*-regulatory regions selected for targeting with CRISPR are shown in turquoise. CRISPR-Cas9 protospacer targets are shown as thin red lines. The vertical black dashed line shows the location of the TSS. *, TFBS inferred from closely related TF, see section 3.5.1.

4.4.3 Construction of a CRISPR library

122 sgRNA scaffolds were constructed by PCR amplification of a cloned sgRNA scaffold sequence³³¹ using primers to introduce each spacer sequence (see section 2.6.1). Guide scaffolds were co-assembled with the *AtU6-26* promoter producing 122 sgRNA expression cassettes. 96 multigene Cas9 expression constructs each containing a pair of sgRNAs were assembled resulting in ~24 constructs for each gene. Figure 4.2 shows a schematic of these plasmids.

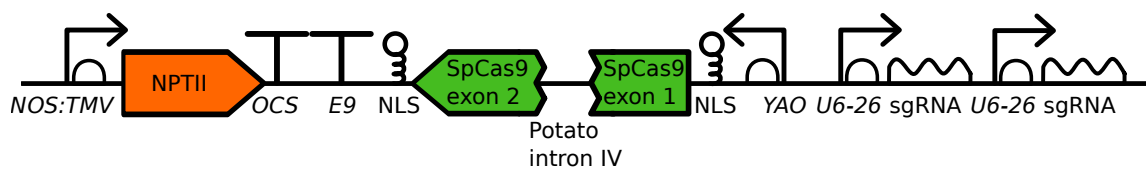


Figure 4.2: **Schematic of SpCas9 expression cassettes.** sgRNAs scaffolds were amplified from a sgRNA template first described by Chen et al.³³¹. These amplicons were used to create expression cassettes which were inserted into the pCsA backbone, creating multigene constructs according to the Loop syntax³²⁰. NLS, nuclear localisation signal.

96 strains of *Agrobacterium*, each containing a Cas9 expression construct with a different pair of sgRNAs, were combined and used to transform 100 Col-0 Arabidopsis plants (see sections 2.2.3 and 2.2.4). Seeds were collected, germinated on kanamycin selective media and 200 T1 lines were transplanted to soil. 50 T1 plants were chosen at random and

T2 seeds were collected. Ten T2 seeds from each parental line were sown. The four *cis*-regulatory regions in T2 plants were amplified using two PCR steps which added two unique barcodes to each amplicon (see section 2.6.4). Samples were multiplexed to allow for simultaneous analysis of the large number of amplicons⁴⁵⁰ and sequenced using PacBio Sequel II SMRT sequencing. Figure 4.3 outlines the strategy used to generate the library of plants with novel mutations in the promoter regions of the four TFs.

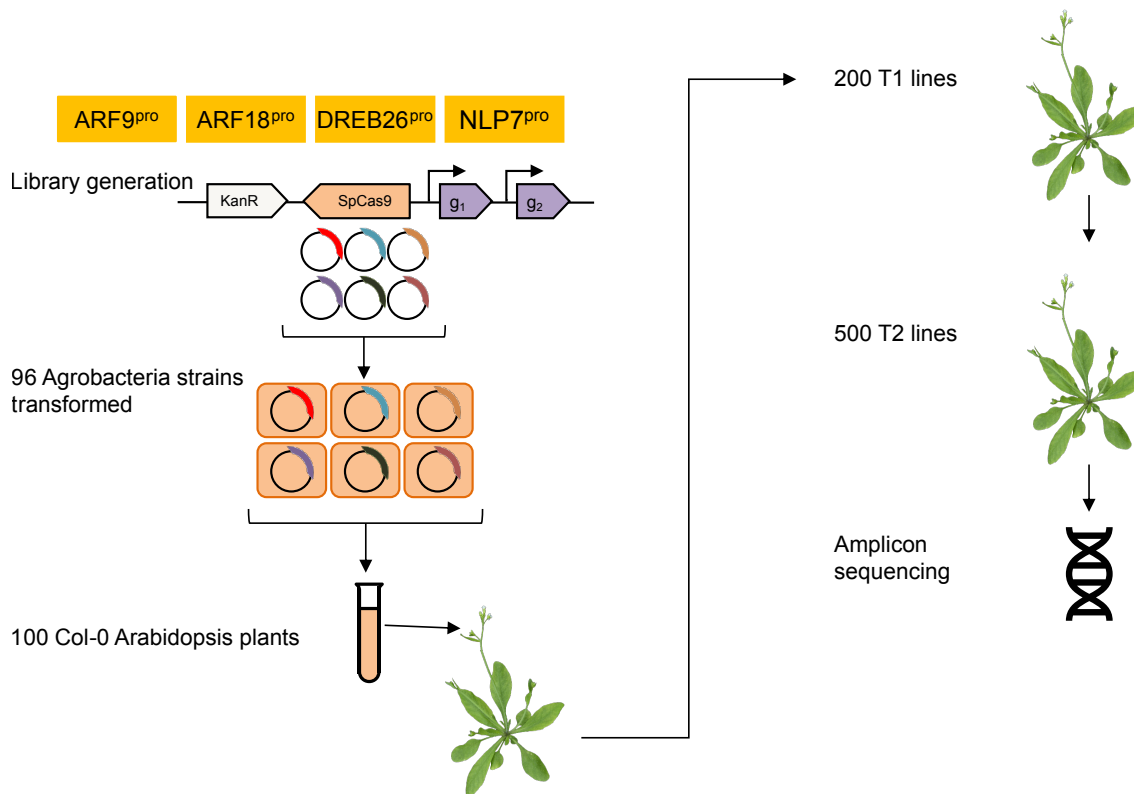


Figure 4.3: **Generation of a *cis*-regulatory mutation library in Arabidopsis.** A library of 96 CRISPR plasmids was generated with each plasmid containing two guides targeting the promoter and 5' UTR of four TFs in the nitrogen subnetwork. 96 *Agrobacterium* strains were transformed with a single plasmid each, and then mixed equally. Arabidopsis plants were transformed, and the four amplicons in T2 plants were sequenced.

4.4.4 A wide range of promoter mutations were generated

Amplicon sequencing generated 4049544 raw reads of mean length 39721 bp of the libraries described above. CCS was performed, generating 2313978 high fidelity (HiFi) reads with yield of 2907046832 bp with median read quality of Q44, indicating good quality reads. Following CCS, reads were demultiplexed by barcodes and reference genes, and variant calling and mutation analyses were performed.

The median number of mapped reads for *ARF9* was 314.5, for *ARF18* was 680.0, for *DREB26* was 342.5 and for *NLP7* was 660.0 (fig. 4.4).

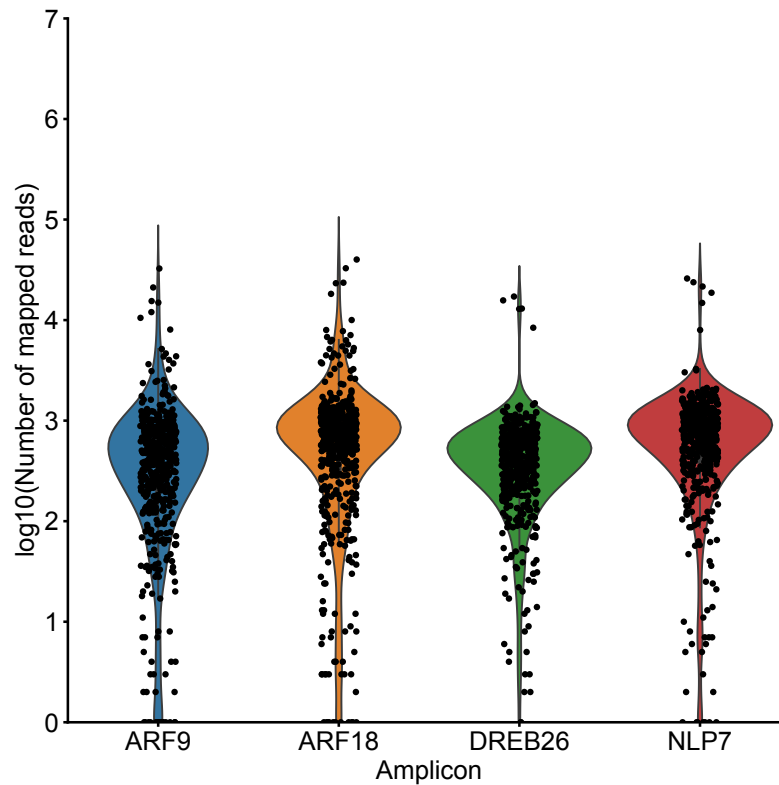


Figure 4.4: **Distribution of \log_{10} read counts per amplicon.** Read counts represent the number of reads mapping to that amplicon in a plant line.

Lines containing more than two alleles were defined as being chimeric. Chimeric lines can occur when Cas9 is expressed in the somatic cells of unedited plant lines, therefore introducing a range of different mutations in individual cells or sectors of tissue⁴⁵⁵. In total, 327/499 T2 lines that were sequenced contained a mutation in at least one *cis*-regulatory region. Of these, 315 lines were non-chimeric (96.3%). 49.6% of non-chimeric mutated lines were homozygous, 23.7% contained at least one heterozygous mutation and 20.6% contained at least one biallelic mutation. 40 lines contained mutations in two genes, and one line contained mutations in three genes.

The greatest number of mutations were found within *ARF9*, with 147/468 lines (31.4%) containing a mutation. This was followed by *NLP7* and *ARF18* which had 112/478 (23.4%) and 67/373 (18.0%) lines containing mutations, respectively. *DREB26* had the lowest number of mutations with 43/366 lines (11.7%) containing mutations (table 4.1). The GC

content of the sequenced *cis*-regulatory region of *ARF9* was 32%, of *ARF18* was 35%, of *DREB26* was 29% and of *NLP7* was 27%. The GC content did not correlate with the number of mutations. *NLP7* and *ARF18* had the highest median reads at each mutation site (646 and 623 respectively), while *DREB26* and *ARF9* had the lowest (344 and 361 respectively) (table 4.1 and fig. 4.5). Of the 315 non-chimeric lines, 64.8% of *ARF9* lines, 40.9% of *ARF18* lines, 37.5% of *DREB26* lines and 49.5% of *NLP7* lines contained homozygous mutations. Most insertions were single nucleotide insertions, with 98.4% of *ARF9*, and 100% of *ARF18*, *DREB26* and *NLP7* insertions being single nucleotide polymorphisms (SNPs).

Overall, 19.7% of all deletions were over 10 bp in length. *ARF18* and *NLP7* had the highest percentage of deletions over 10 bp in length, at 27.3% and 25.6% respectively, although *ARF18* only had 11 non-chimeric deletions in total (table 4.1 and fig. 4.6). The largest deletion in *ARF18* was 34 bp. *ARF9* had 50 non-chimeric deletions, with the largest being 31 bp in length. *DREB26* had 22 non-chimeric deletions, with the largest being 21 bp in length, and *NLP7* had 39 non-chimeric deletions, the longest being 34 bp.

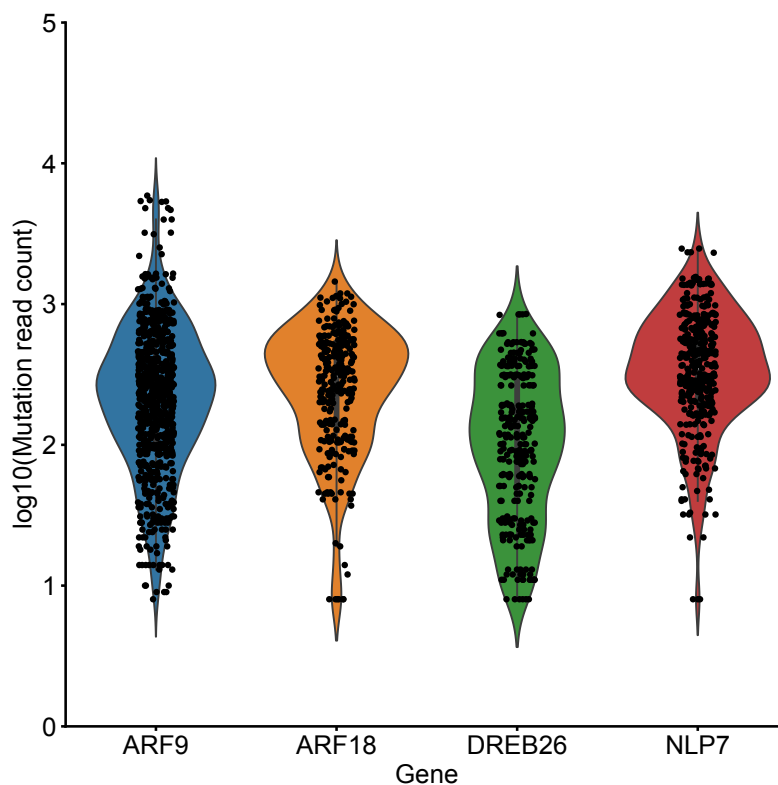


Figure 4.5: **Distribution of \log_{10} read counts at each mutation site for each gene.** Read counts include reads from wild type alleles.

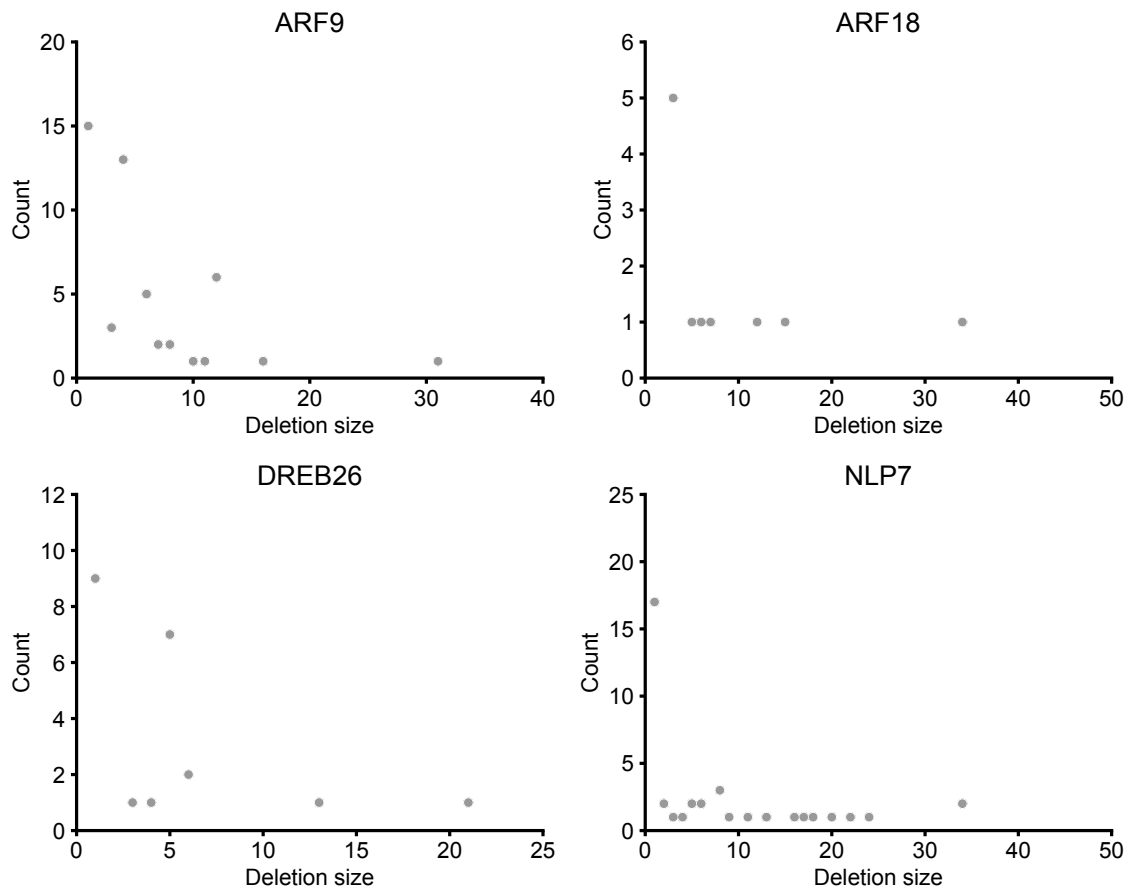


Figure 4.6: **Deletion size counts in the promoter/5' UTRs of four genes targeted using CRISPR.** The number of plant lines containing deletions of varying sizes in each gene (*ARF9*, *ARF18*, *DREB26* and *NLP7*).

The distribution of mutations relative to the TSS in each gene is shown in fig. 4.7. *ARF9*, *ARF18* and *DREB26* had many mutations within the desirable region surrounding the TSS. *ARF9* also had many mutations at around 740 bp upstream of the TSS. *NLP7* had two peaks of mutations, ~250 bp upstream and ~250 bp downstream of the TSS.

In summary, the CRISPR library was a success, with the generation of a range of mutations in the *cis*-regulatory regions of *DREB26*, *ARF9*, *ARF18* and *NLP7*, and the vast majority of these were non-chimeric SNPs. Around half of the mutations were homozygous, and the rest were heterozygous or biallelic.

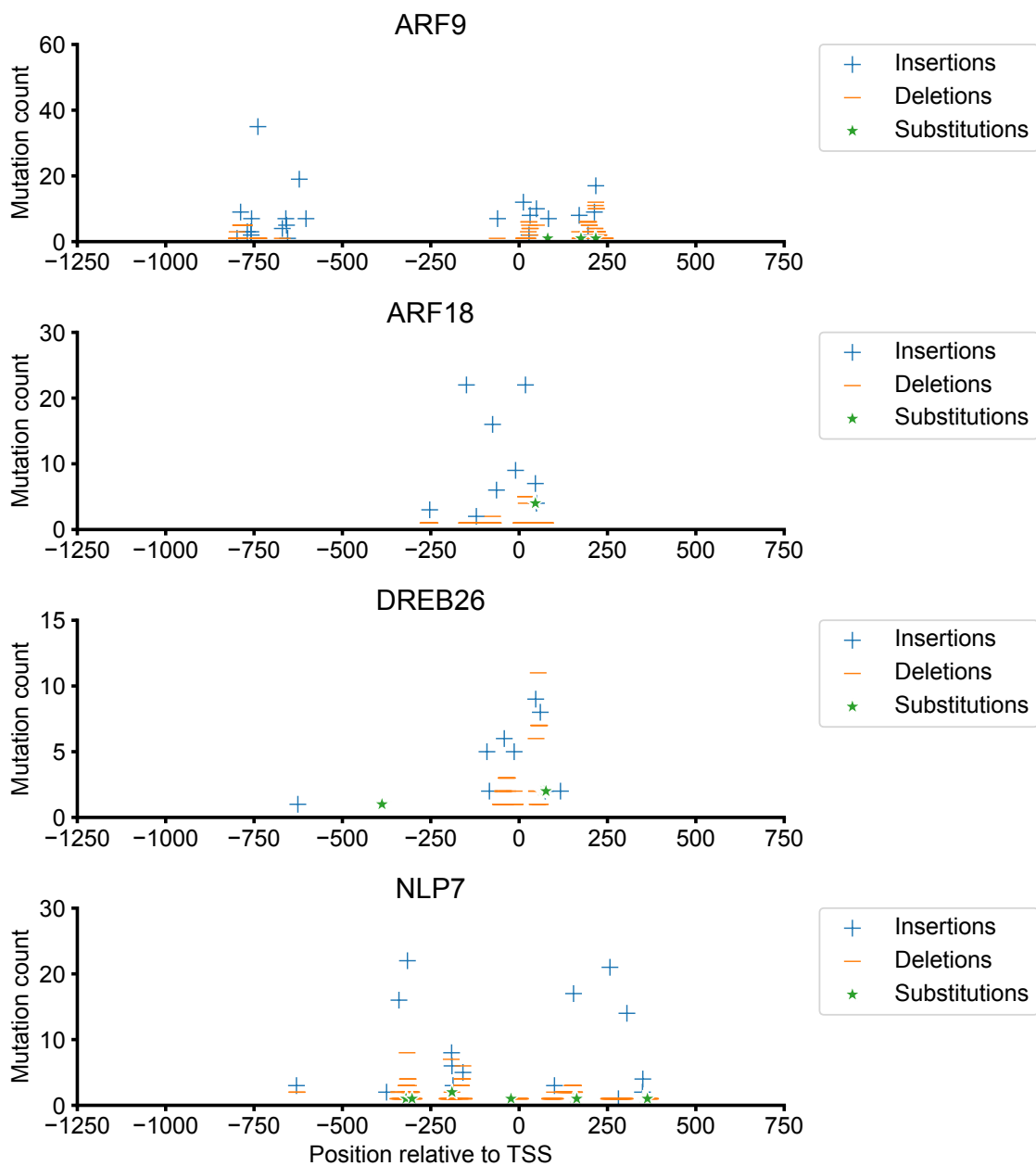


Figure 4.7: **Distribution of mutations relative to the TSS.** The distribution of insertions, deletions, and substitutions in non-chimeric mutated lines relative to the TSS (obtained from Araport 11³⁶³) in each gene.

Table 4.1: **Details of the genotypes, mutation types and overlapping TFBSs of the four CRISPR library gene targets.** WT, wild type.

	<i>ARF9</i>	<i>ARF18</i>	<i>DREB26</i>	<i>NLP7</i>
Genotypes				
Total plant lines with mapped amplicon reads	468	373	366	478
Total mutated lines	147	67	43	112
% non-chimeric mutated lines	96.6	98.5	93	95.5
% homozygous mutated lines (non-chimeric)	64.8	40.9	37.5	49.5
% heterozygous mutated lines (non-chimeric)	8.5	37.9	47.5	29.9
% biallelic mutated lines (non-chimeric)	26.8	21.2	15	20.6
Mutation types				
Total mutations	236	106	73	180
% insertions (of total mutations)	77.1	82.1	54.8	70.6
% deletions (of total mutations)	21.2	10.4	30.1	21.7
% of deletions over 10 bp length	18	27.3	9.1	25.6
% substitutions (of total mutations)	1.7	7.5	15.1	7.8
Median mutation site reads (incl. WT alleles)	361	623	344	646
Overlapping TFBSs				
% of insertions overlapping TFBSs	18.7	60.9	45	32.3
% of deletions overlapping TFBSs	30	63.6	72.7	17.9
% of substitutions overlapping TFBSs	0	0	18.2	7.1

4.4.5 Selection of plant lines for further study

To determine if mutations to the *cis*-regulatory regions had any impact on gene expression or plant growth, several lines were selected for further analysis. These were selected using data collected on the zygosity and locations of mutations. First, lines containing 3 or more mutations at a single locus (chimeric lines) were excluded. Lines with biallelic mutations were also deprioritised as, in *cis*-regulatory regions, two different mutations may have different consequences. Lines in which mutations disrupted a binding site of a TF were prioritised, as were lines with larger (> 4 bp) deletions. Using these criteria, 10 lines were selected (table 4.2). Due to time constraints only the first 5 were analysed. These were selected based on observations made during the production of T3 seeds, which suggested changes to growth and development.

Table 4.2: **Types of mutations in plant lines selected for further analysis.**

Plant ID	Gene	Genotype	Mutation
plntEPSWT20069-9	<i>NLP7</i>	homozygous	Homozygous 11 bp deletion 258 bp downstream of TSS disrupts <i>NLP7</i> TFBS.
plntEPSWT20125-4	<i>NLP7</i> , <i>ARF9</i>	biallelic	<i>NLP7</i> : biallelic 13 bp deletion/1 bp insertion 316 bp upstream of TSS, disrupts ANR1, <i>NLP7</i> and ANAC032 TFBSs. <i>ARF9</i> : Biallelic 12 bp deletion/1 bp insertion 207-218 bp downstream of TSS.
plntEPSWT20130-4	<i>ARF18</i>	biallelic	Biallelic 15 bp deletion/1 bp deletion 6 bp downstream of TSS. Homozygous 1 bp insertion 76 bp upstream of TSS.
plntEPSWT20134-3	<i>DREB26</i>	homozygous	Homozygous 5 bp deletion 55 bp downstream of TSS.
plntEPSWT20139-9	<i>ARF9</i>	homozygous	Homozygous 31 bp deletion 215 bp downstream of TSS.
plntEPSWT20127-10	<i>DREB26</i>	homozygous	Homozygous 4 bp deletion 11 bp upstream of TSS.
plntEPSWT20144-5	<i>DREB26</i>	biallelic	Biallelic 1 bp or 13 bp deletions 48 bp downstream of TSS disrupting <i>DREB26</i> TFBS.
plntEPSWT20154-4	<i>DREB26</i> , <i>ARF9</i>	homozygous	Homozygous 1 bp insertion 85 bp upstream of TSS. Homozygous 10 bp insertion/substitution 63 bp downstream of TSS. Also has homozygous insertion in <i>ARF9</i> 770 bp upstream of TSS.
plntEPSWT20142-4	<i>NLP7</i>	homozygous	Homozygous 1 bp insertion overlapping <i>NLP7</i> TFBS 257 bp downstream of TSS. Homozygous 23 bp substitution 351 bp downstream of TSS.
plntEPSWT20142-8	<i>NLP7</i>	homozygous	Homozygous 1 bp insertion 349 bp downstream of TSS, homozygous 5 bp deletion 253 bp downstream of TSS overlapping <i>NLP7</i> TFBS.

The T3 progeny seeds of selected T2 lines were grown in 1 and 10 mM KNO₃ conditions on vertical plates to assess the effect of the mutations on the nitrate-response (see section 2.2.2 for details). Seeds for the Col-0 controls were collected from equivalent generations of plants grown alongside the mutant lines. Root system architecture (RSA) was analysed 9 days after germination using the SMARTROOT plugin (v4.21)³³⁸ for (Fiji Is Just) ImageJ (v2.3.0)³³⁹, to look for changes in primary root length, number and density of lateral roots, average lateral root length, total root length, and the ratio of lateral root length to total root length (see section 2.2.8). Type II two-way ANOVAs were used to test for significant effects of genotype on the RSA parameters, and Tukey's post hoc tests were used when significant to test significance of individual comparisons.

Root tissue was collected for RNA extraction 12-days following germination, and the relative expression of genes containing mutations in their *cis*-regulatory regions was analysed by qRT-PCR (see section 2.5.5 for details). The expression of *NIR1* was assessed in all lines as a proxy for genes regulated by the subnetwork (fig. 3.24).

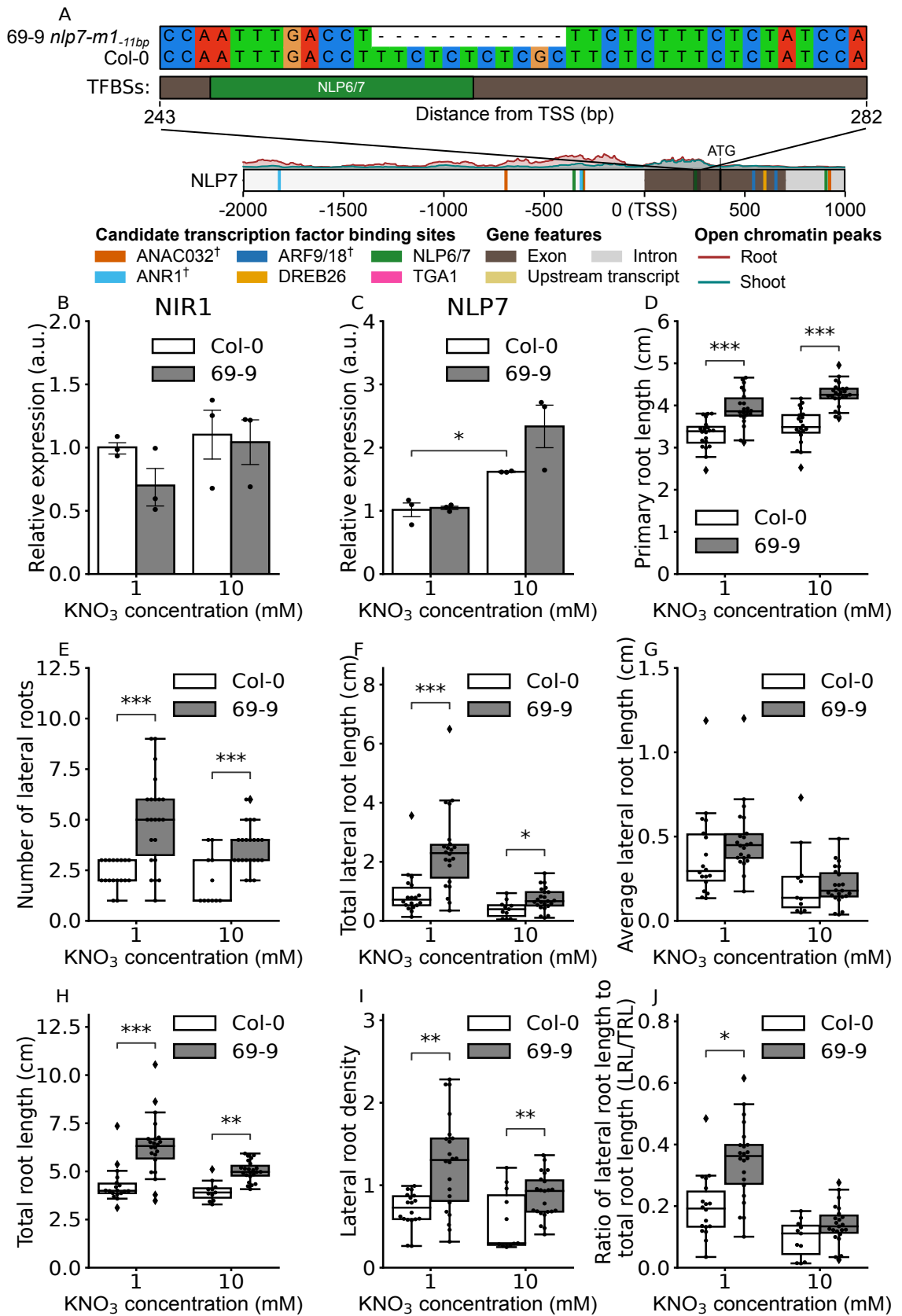
4.4.6 Characterisation of plant lines

Line plntEPSWT30069-9 containing *nlp7-m1_11bp* showed an increased lateral root length:total root length ratio

Line plntEPSWT30069-9 was selected for analysis because it contained an 11 bp deletion overlapping a candidate NLP7 TFBS in the 5' UTR of *NLP7* (fig. 4.8A). Several other candidate TFBSs were predicted to overlap the deletion (TFBSs of far1-related sequence 9 (FRS9), basic pentacysteine 1 (BPC1), basic pentacysteine 5 (BPC5) and transcription factor IIIA (TFIIIA)) (Supplementary fig. S1). Although this caused no significant change in relative expression of *NLP7* (fig. 4.8C) or its target *NIR1* (fig. 4.8B) compared to Col-0 on either nitrate concentration (Welch's t-tests, $P > 0.05$), there were significant root system architectural changes (Type II two-way ANOVAs). There was a significantly higher mean ratio of lateral root length to total root length (LRL/TRL) at 1 mM KNO₃ in line plntEPSWT30069-9 (Tukey's post hoc tests; 0.35 ± 0.13 cm; $P < 0.05$) than Col-0 (fig. 4.8J; 0.2 ± 0.1 cm). At both 1 mM and 10 mM KNO₃, line plntEPSWT30069-9 had significantly longer mean primary roots (fig. 4.8D,K; plntEPSWT30069-9: 1 mM 3.94 ± 0.41 cm, $P < 0.001$; 10 mM 4.26 ± 0.29 cm, $P < 0.001$; Col-0: 1 mM 3.32 ± 0.35 cm, 10 mM

3.51 ± 0.42 cm), more lateral roots (fig. 4.8E; plntEPSWT30069-9: 1 mM 4.95 ± 2.21, $P < 0.001$; 10 mM 3.7 ± 1.15, $P < 0.001$; Col-0: 1 mM 2.33 ± 0.69, 10 mM 2.0 ± 1.26), higher total lateral root length (fig. 4.8F; plntEPSWT30069-9: 1 mM 2.36 ± 1.38 cm, $P < 0.001$; 10 mM 0.73 ± 0.38 cm, $P < 0.05$; Col-0: 1 mM 0.92 ± 0.77 cm, 10 mM 0.39 ± 0.29 cm), higher mean total root length (fig. 4.8H; plntEPSWT30069-9: 1 mM 6.29 ± 1.55 cm, $P < 0.001$; 10 mM 5.02 ± 0.53 cm, $P < 0.01$; Col-0: 1 mM 4.29 ± 0.93 cm, 10 mM 3.97 ± 0.52 cm) and higher lateral root density (fig. 4.8I; plntEPSWT30069-9: 1 mM 1.26 ± 0.58, $P < 0.01$; 10 mM 0.87 ± 0.27, $P < 0.01$; Col-0: 1 mM 0.7 ± 0.21, 10 mM 0.57 ± 0.37) than Col-0.

Although line plntEPSWT30069-9 had an increased number and density of lateral roots compared to Col-0, there was no significant difference in average lateral root length between genotypes (fig. 4.8G; $P > 0.05$).



K

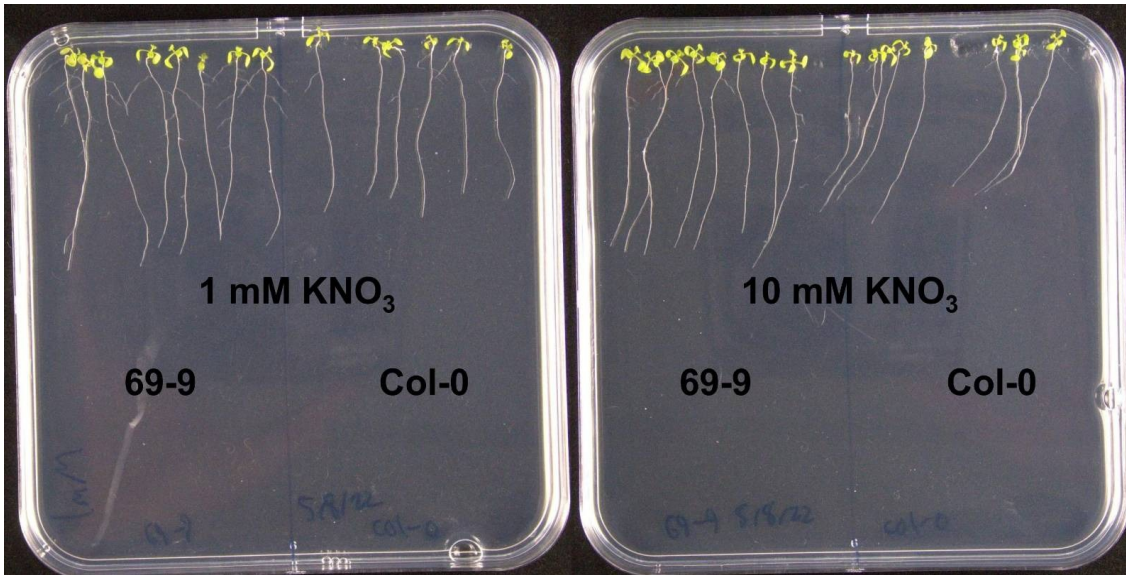


Figure 4.8: **Characterisation of line plntEPSWT30069-9 containing *nlp7-m1-11bp*.** A: Location of mutation relative to transcription start site. Relative expression of B: *NIR1* and C: *NLP7* in line plntEPSWT30069-9 and Col-0 plants on 1 mM and 10 mM KNO_3 . Significance was calculated using either independent t-tests (when variances were equal) or Welch's t-tests (when variances were not equal)³⁴⁹. qRT-PCRs performed using RNA extracted from roots 12 days after germination. $N = 3$ biological replicates, each consisting of a pool of roots from ~50 seedlings. Each qRT-PCR was performed in duplicate. D to J: RSA of line plntEPSWT30069-9 and Col-0 9 days after germination on 1 mM and 10 mM KNO_3 . Significance calculated using ANOVA and Tukey HSD post hoc tests when necessary³⁴⁴. $N = 8$ biological replicates (plants) and three technical replicates (plates). K: Photos of representative Col-0 and plntEPSWT30069-9 plants 9 days after germination grown on 1 and 10 mM KNO_3 . **, $P < 0.01$. *, $P < 0.05$. †, TFBS inferred from closely related TF, see section 3.5.1.

Line plntEPSWT30125-4 containing *arf9-m1-12bp*, *nlp7-m2-13bp/nlp7-m3+1bp* and *nlp7-m4-1bp/nlp7-m5+1bp* showed an increased lateral root length:total root length ratio

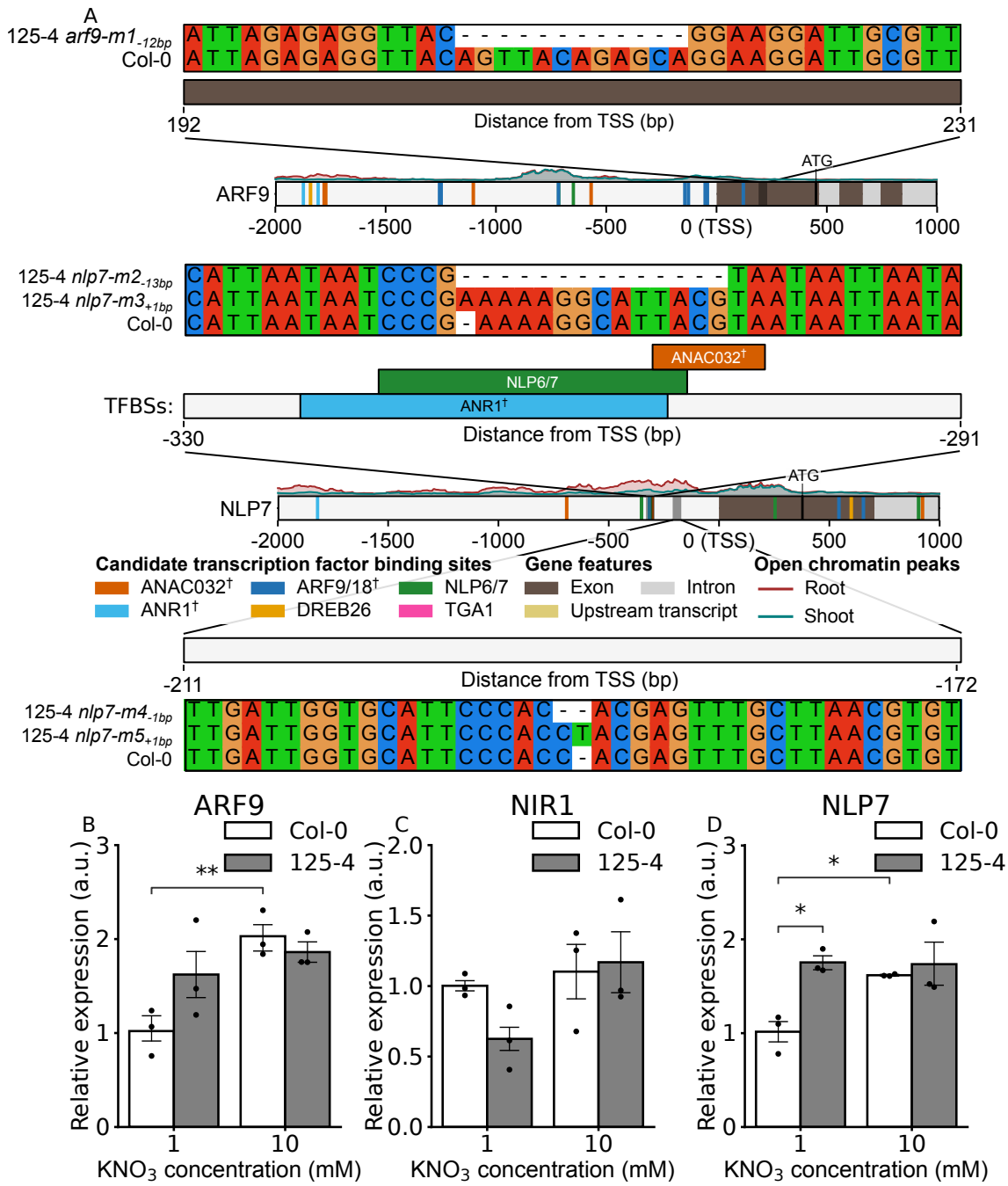
Line plntEPSWT30125-4 had a very similar RSA phenotype to line plntEPSWT30069-9, with significantly longer primary roots (fig. 4.9E,L; plntEPSWT30125-4: 1 mM 3.95 ± 0.36 cm, $P < 0.001$; 10 mM 3.89 ± 0.32 cm, $P < 0.001$; Col-0: 1 mM 3.46 ± 0.31 cm, 10 mM 3.3 ± 0.34 cm), more lateral roots (fig. 4.9F; plntEPSWT30125-4: 1 mM 4.16 ± 1.99 , $P < 0.001$; 10 mM 3.47 ± 1.47 , $P < 0.01$; Col-0: 1 mM 2.11 ± 1.18 , 10 mM 1.92 ± 1.24), longer total lateral root length (fig. 4.9G; plntEPSWT30125-4: 1 mM 1.86 ± 0.91 cm, $P < 0.001$; 10 mM 0.88 ± 0.49 cm, $P < 0.01$; Col-0: 1 mM 0.69 ± 0.81 cm, 10 mM 0.39 ± 0.41 cm),

longer total root length (fig. 4.9I; plntEPSWT30125-4: 1 mM 5.83 ± 1.07 cm, $P < 0.001$; 10 mM 4.76 ± 0.61 cm, $P < 0.001$; Col-0: 1 mM 4.18 ± 0.84 cm, 10 mM 3.65 ± 0.65 cm), and a higher ratio of lateral root length to total root length (fig. 4.9K; plntEPSWT30125-4: 1 mM 0.30 ± 0.11 , $P < 0.001$; 10 mM 0.18 ± 0.09 , $P < 0.05$; Col-0: 1 mM 0.15 ± 0.12 , 10 mM 0.10 ± 0.08) than Col-0.

At 1 mM KNO_3 , line plntEPSWT30125-4 had a significantly higher average lateral root length (fig. 4.9H; $M = 0.46$ cm, $P < 0.05$) and lateral root density (fig. 4.9J; $M = 1.04$, $P < 0.01$) than Col-0 (PRL: $M = 0.29$ cm; LRD: $M = 0.61$).

Line plntEPSWT30125-4 contained a biallelic 13 bp deletion or 1 bp insertion in the *NLP7* promoter of line plntEPSWT30125-4 which overlapped the TFBSs of ANR1, NLP6/7 and also ANAC032 in the case of the deletion (fig. 4.9A (middle)). To test if this influenced gene expression of *NLP7*, qRT-PCRs were performed on extracted RNA, revealing that *NLP7* expression was significantly increased in line plntEPSWT30125-4 ($M=1.75$; Welch's t-tests, $t = -5.29$, $P < 0.05$) compared to Col-0 ($M=1.02$) at 1 mM KNO_3 (fig. 4.9D). Although there was an increase in *NLP7* expression, *NIR1* expression was not significantly changed, although it was slightly decreased at 1 mM KNO_3 ($P > 0.05$) (fig. 4.9C).

Line plntEPSWT30125-4 also contained a homozygous 12 bp deletion in the *ARF9* 5' UTR (fig. 4.9A (top)), although *ARF9* expression was not changed significantly in the mutant compared to Col-0 (fig. 4.9B) ($P > 0.05$).



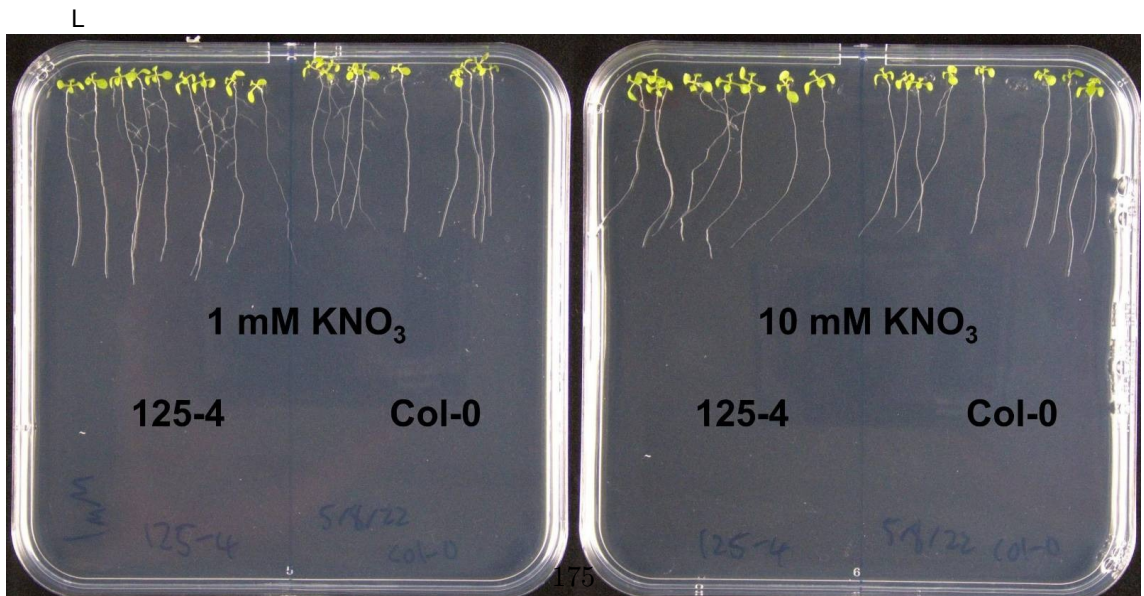
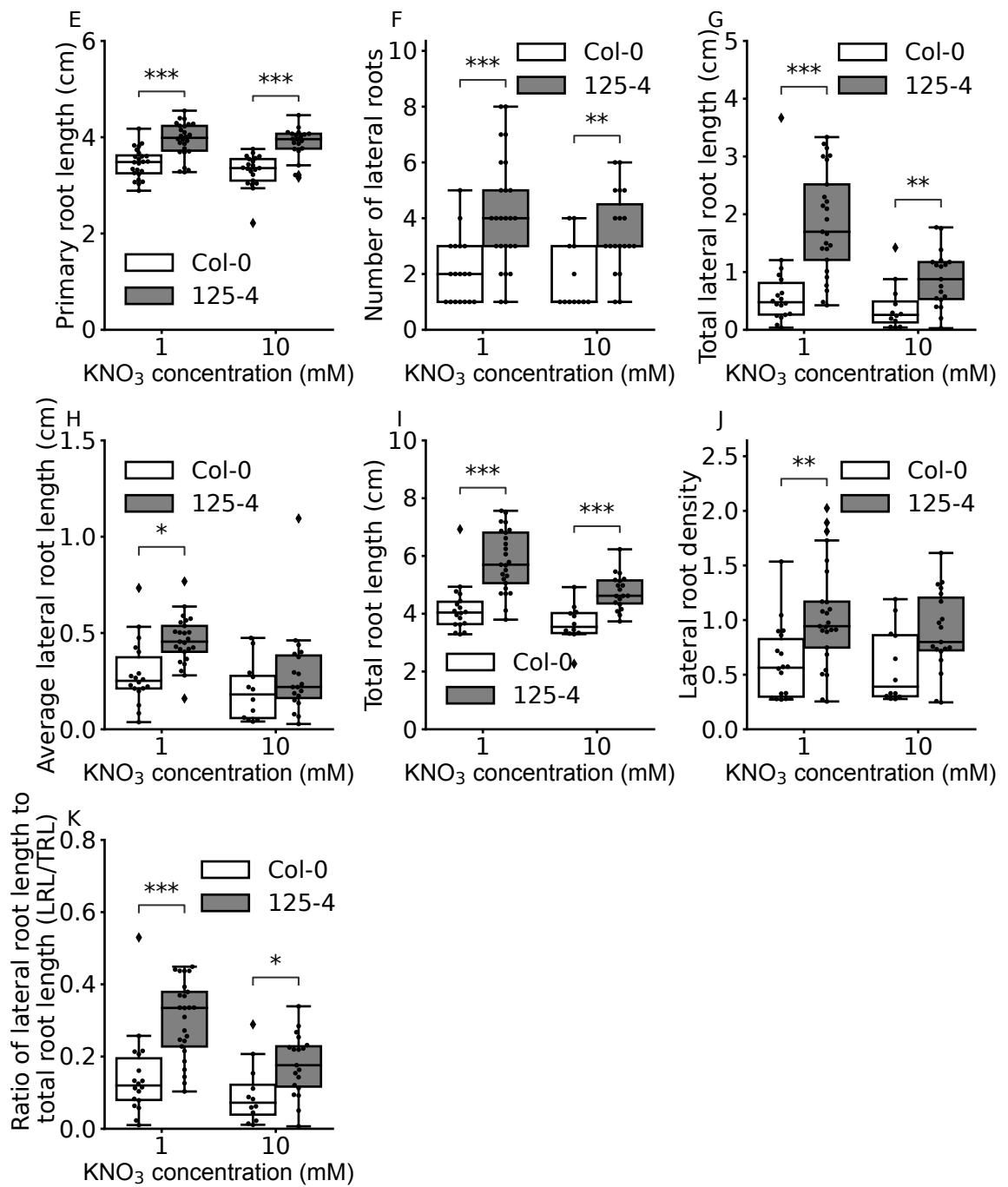
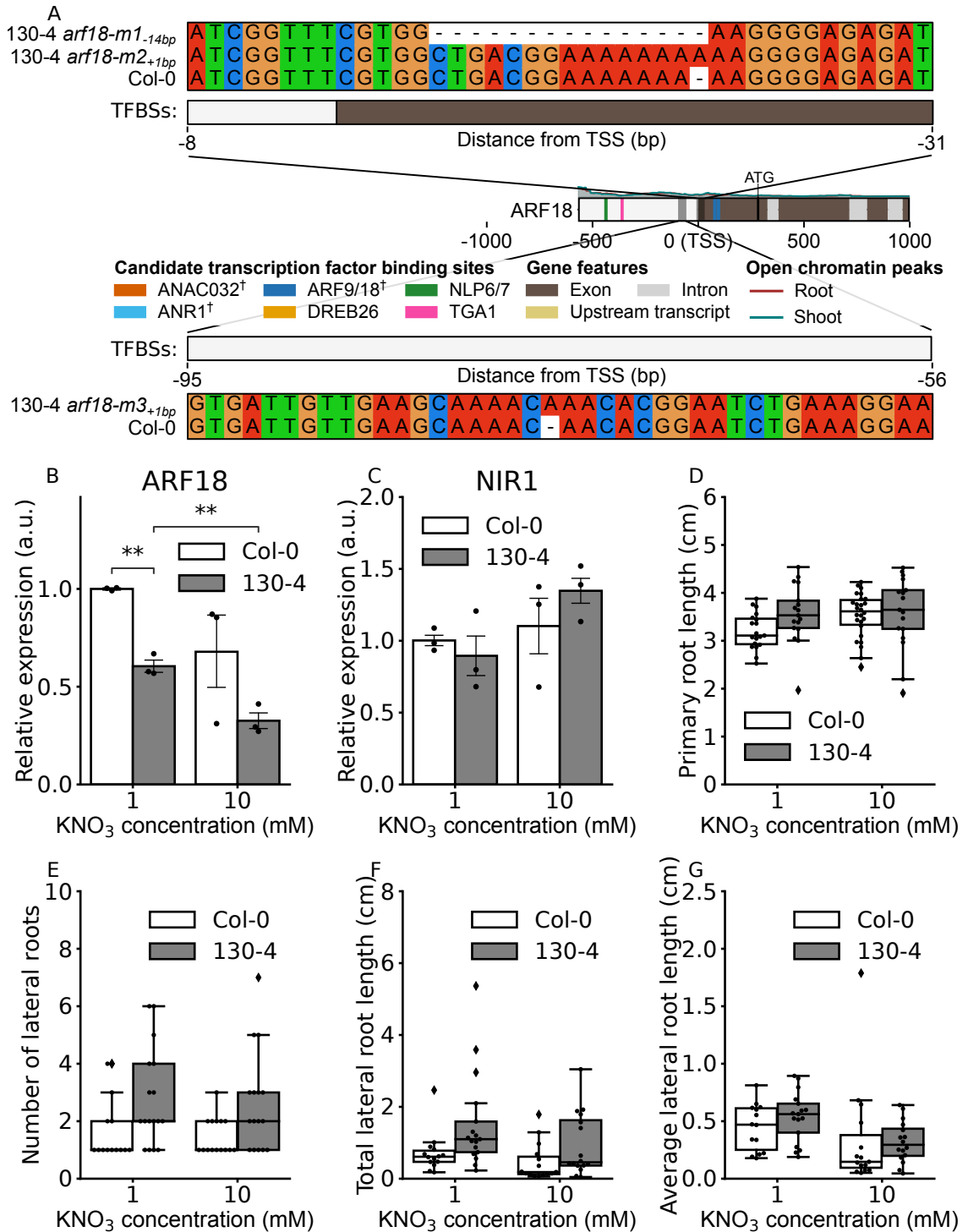


Figure 4.9: **Characterisation of line plntEPSWT30125-4 containing *arf9-m1-12bp*, *nlp7-m2-13bp/nlp7-m3+1bp* and *nlp7-m4-1bp/nlp7-m5+1bp*.** A: Mutation locations relative to transcription start site in *ARF9* and *NLP7*. Relative expression of B: *ARF9*, C: *NIR1* and D: *NLP7* in line plntEPSWT30125-4 and Col-0 plants on 1 mM and 10 mM KNO₃. Significance was calculated using either independent t-tests (when variances were equal) or Welch's t-tests (when variances were not equal)³⁴⁹. qRT-PCRs performed using RNA extracted from roots 12 days after germination. $N = 3$ biological replicates, each consisting of a pool of roots from ~50 seedlings. Each qRT-PCR was performed in duplicate. E to K: RSA of line plntEPSWT30125-4 and Col-0 9 days after germination on 1 mM and 10 mM KNO₃. Significance calculated using ANOVA and Tukey HSD post hoc tests when necessary³⁴⁴. $N = 8$ biological replicates (plants) and three technical replicates (plates). L: Photos of representative Col-0 and plntEPSWT30069-9 plants 9 days after germination grown on 1 and 10 mM KNO₃. **, $P < 0.01$. *, $P < 0.05$. †, TFBS inferred from closely related TF, see section 3.5.1.

Line plntEPSWT30130-4 containing *arf18-m1-14bp/arf18-m2+1bp* and *arf18-m3+1bp* showed a decrease in *ARF18* expression

Line plntEPSWT30130-4 was selected for analysis because it contained a biallelic 14 bp deletion or 1 bp insertion in the *ARF18* 5' UTR, along with a 1 bp insertion the promoter, both of which were close to the TSS increasing the chance of a transcriptional change (fig. 4.10A). These mutations overlapped candidate TFBSs of vernalisation 1 (VRN1), related to vernalisation 1 (RTV1), high cambial activity 2 (HCA2) and BPC1 (Supplementary fig. S2 top). Mutation *arf18-m1-14bp* also overlapped a TFBS of BPC5. Some plants from line plntEPSWT30130-4 lacked chlorophyll pigment and were excluded from analyses (fig. 4.10K). This was because the CRISPR construct had likely been inserted into a chlorophyll gene and was most likely still present in those plants. The expression of *ARF18* was significantly lower in line plntEPSWT30130-4 ($M = 0.60, \pm 0.06$) than Col-0 ($M = 1.00, \pm 0.01$) on 1 mM KNO₃ (fig. 4.10B; Welch's t-tests $t = 0.12, P < 0.01$). The *ARF18* expression was also significantly lower at 10 mM (0.33 ± 0.08) than 1 mM (0.60 ± 0.06) KNO₃ in line plntEPSWT30130-4 (Welch's t-tests $t = -5.18, P < 0.01$). In chapter 3, *ARF18* was shown to significantly repress *NIR1* (fig. 3.15G), a well-studied nitrate assimilation gene (fig. 3.24). The relative expression of *NIR1* was tested by qRT-PCR to test whether the reduction in *ARF18* expression affected its expression. However, *NIR1* expression was not significantly different between line plntEPSWT30130-4 and Col-0 on either nitrate concentration, although it was slightly higher in line plntEPSWT30130-4 at 10 mM (fig. 4.10C; $P > 0.05$).

The RSA of line plntEPSWT30130-4 was analysed on different nitrate concentrations to test whether the reduction in *ARF18* expression led to any changes. The RSA of line plntEPSWT30130-4 was very similar to that of Col-0, although at 1 mM KNO_3 the total root length was significantly higher in line plntEPSWT30130-4 (5.08 ± 1.32 cm) than Col-0 (fig. 4.10H; 3.93 ± 0.75 cm; $P < 0.05$).



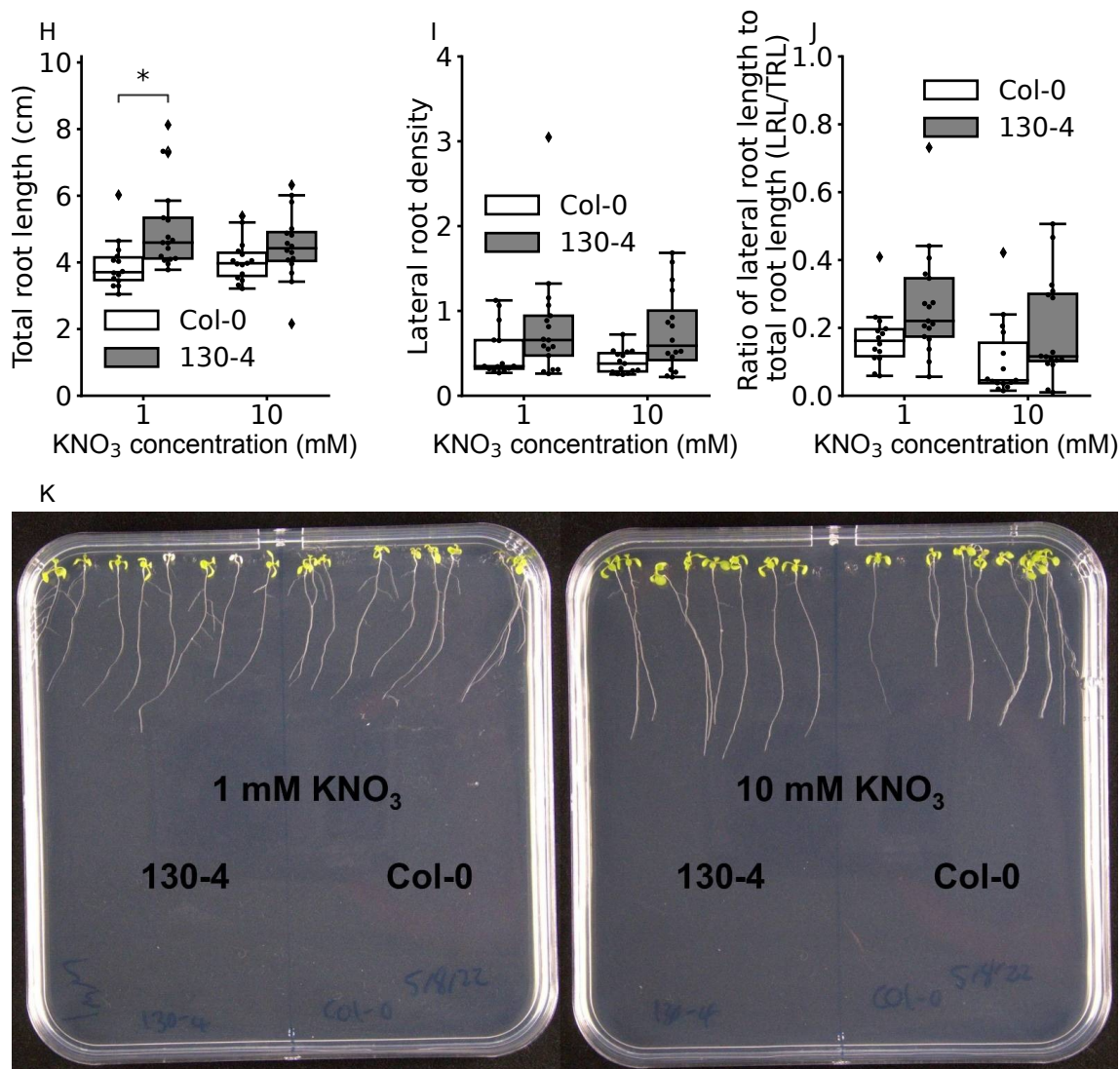
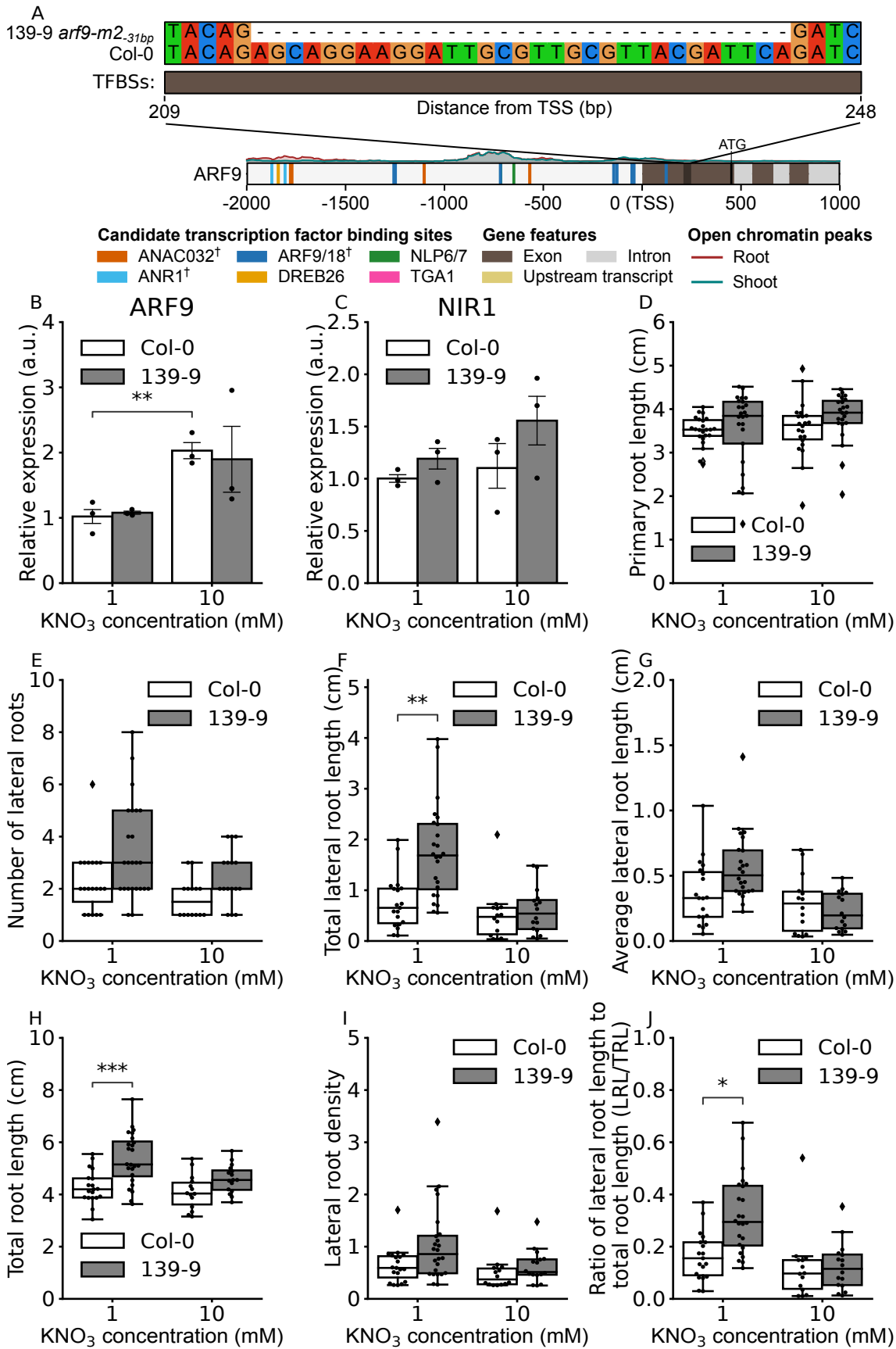


Figure 4.10: **Characterisation of line plntEPSWT30130-4 containing *arf18-m1*_{-14bp}/*arf18-m2*_{+1bp} and *arf18-m3*_{+1bp}.** A: Mutation locations relative to transcription start site. Relative expression of B: *ARF18* and C: *NIR1* in line plntEPSWT30130-4 and Col-0 plants on 1 mM and 10 mM KNO₃. Significance was calculated using either independent t-tests (when variances were equal) or Welch's t-tests (when variances were not equal)³⁴⁹. qRT-PCRs performed using RNA extracted from roots 12 days after germination. *N* = 3 biological replicates, each consisting of a pool of roots from ~50 seedlings. Each qRT-PCR was performed in duplicate. D to J: RSA of line plntEPSWT30130-4 and Col-0 9 days after germination on 1 mM and 10 mM KNO₃. Significance calculated using ANOVA and Tukey HSD post hoc tests when necessary³⁴⁴. *N* = 8 biological replicates (plants) and three technical replicates (plates). K: Photos of representative Col-0 and plntEPSWT30069-9 plants 9 days after germination grown on 1 and 10 mM KNO₃. **, *P* < 0.01. *, *P* < 0.05. †, TFBS inferred from closely related TF, see section 3.5.1.

Line plntEPSWT30139-9 containing *arf9-m2-31bp* showed an increase in the ratio of lateral root length:total root length at 1 mM KNO₃

The function of ARF TFs is known to be partially redundant⁴⁵⁶, so the effect of a 31 bp deletion in the 5' UTR of *ARF9* in line plntEPSWT30139-9 (fig. 4.11A) was tested to see if it would have a similar effect on RSA as mutations in *ARF18*. Interestingly, line plntEPSWT30139-9 had significantly a higher total lateral root length (fig. 4.11F,K; plntEPSWT30139-9: 1.76 ± 0.93 cm, $P < 0.01$; Col-0: 0.75 ± 0.52 cm), total root length (fig. 4.11H; plntEPSWT30139-9: 5.36 ± 1.0 cm, $P < 0.001$; Col-0: 4.29 ± 0.64 cm) and ratio of lateral root length to total root length (fig. 4.11J; plntEPSWT30139-9: 0.32 ± 0.15 , $P < 0.05$; Col-0: 0.16 ± 0.09) than Col-0 at 1 mM nitrate. The increase in total root length compared to Col-0 in line plntEPSWT30139-9 was similar to the increase in root length in line plntEPSWT30130-4 containing the *ARF18* mutation (fig. 4.10H) compared to Col-0.

To test if these changes in RSA were due to a change in ARF9 expression, qRT-PCR was performed on RNA extracted from roots of line plntEPSWT30139-9 and Col-0 grown on 1 mM and 10 mM KNO₃. Surprisingly, there was no significant change in expression of *ARF9* or *NIR1* on either nitrate concentration compared to Col-0 ($P > 0.05$) (fig. 4.11 B and C).



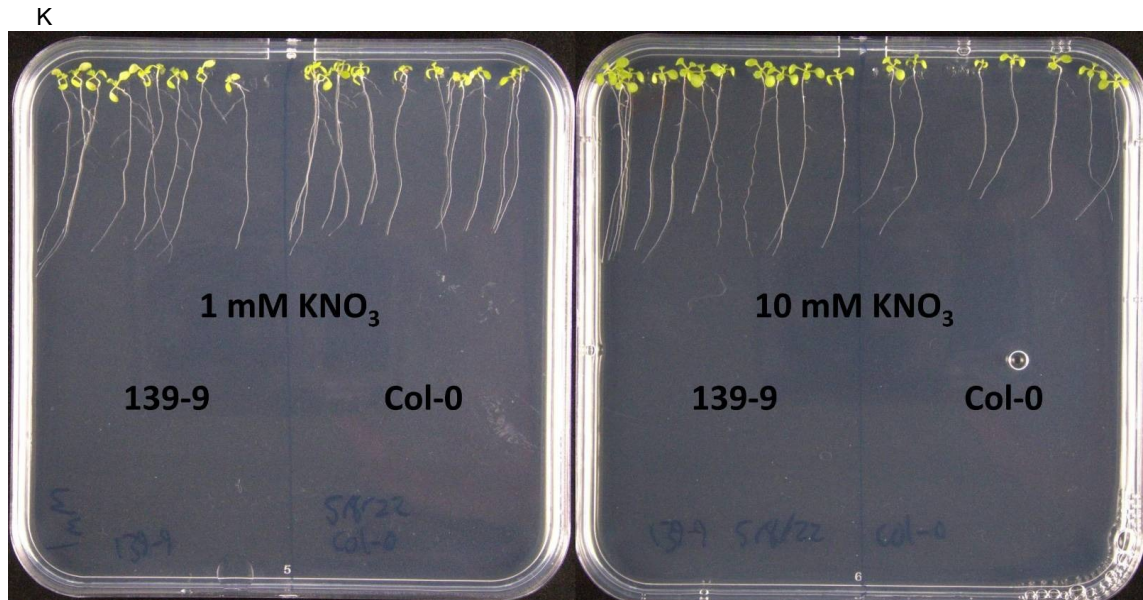
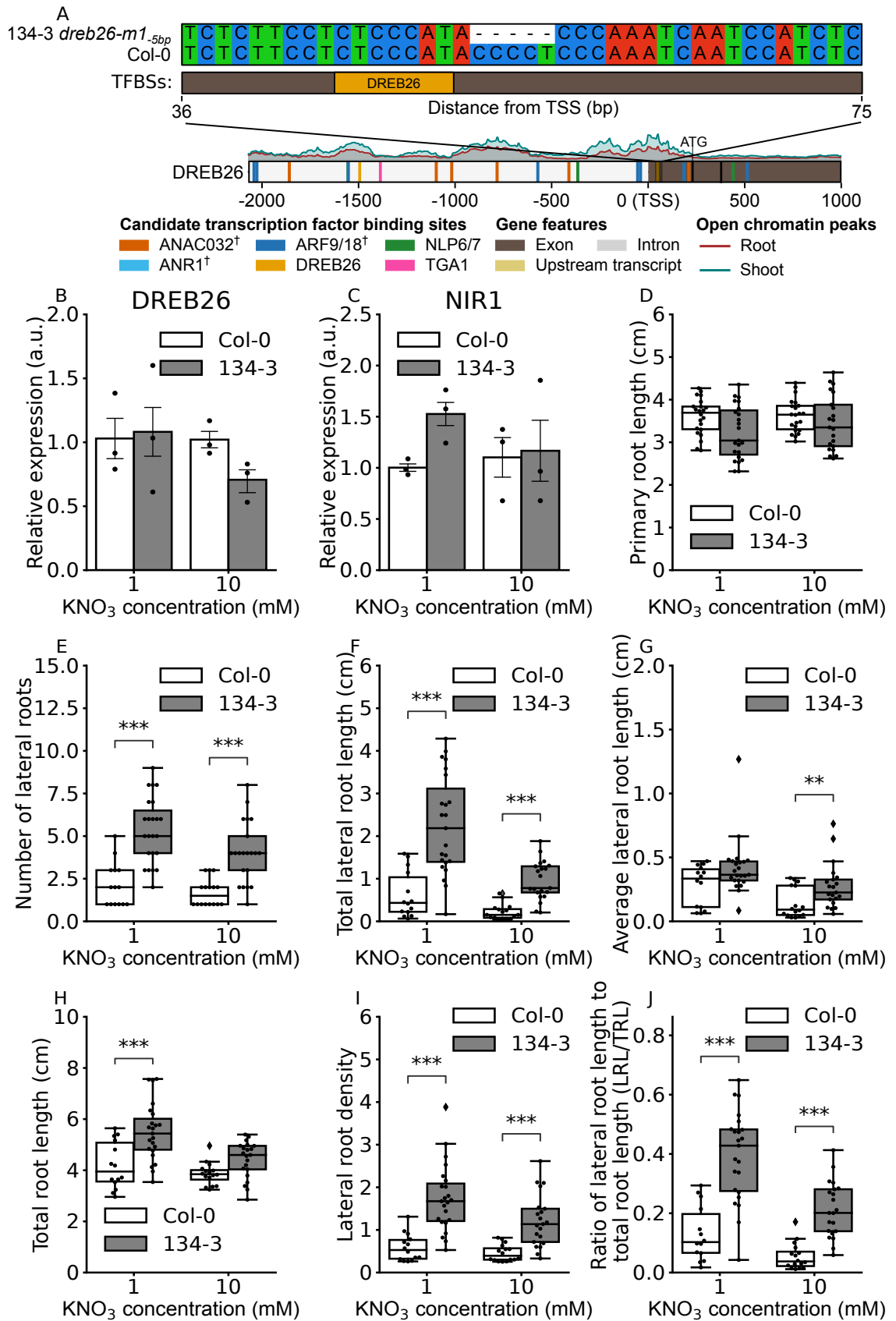


Figure 4.11: **Characterisation of line plntEPSWT30139-9 containing *arf9-m2.31bp*.** A: Location of mutation relative to transcription start site. Relative expression of B: *ARF9* and C: *NIR1* in line plntEPSWT30139-9 and Col-0 plants on 1 mM and 10 mM KNO_3 . Significance was calculated using either independent t-tests (when variances were equal) or Welch's t-tests (when variances were not equal)³⁴⁹. qRT-PCRs performed using RNA extracted from roots 12 days after germination. $N = 3$ biological replicates, each consisting of a pool of roots from ~50 seedlings. Each qRT-PCR was performed in duplicate. D to J: RSA of line plntEPSWT30139-9 and Col-0 9 days after germination on 1 mM and 10 mM KNO_3 . Significance calculated using ANOVA and Tukey HSD post hoc tests when necessary³⁴⁴. $N = 8$ biological replicates (plants) and three technical replicates (plates). K: Photos of representative Col-0 and plntEPSWT30069-9 plants 9 days after germination grown on 1 and 10 mM KNO_3 . **, $P < 0.01$. *, $P < 0.05$. †, TFBS inferred from closely related TF, see section 3.5.1.

Line plntEPSWT30134-3 containing *dreb26-m1.5bp* showed an increase in the ratio of lateral root length:total root length

DREB26 is a putative target of NLP7 (fig. 3.14F), ARF9 and 18 (fig. 3.14G–I). Therefore, line plntEPSWT30134-3 containing a homozygous 5 bp deletion in the *DREB26* 5' UTR was investigated (fig. 4.12), to see if it would cause transcriptional changes or changes in RSA. The deletion fell within 2 bp of a *DREB26* candidate TFBS, and overlapped candidate TFBSs of BPC1, TFIIIA, basic pentacysteine 6 (BPC6), FRS9 and BPC5 (Supplementary fig. S3). Interestingly, the ratio of lateral root length to total root length was significantly higher at both nitrate concentrations in line plntEPSWT30134-3 compared to Col-0 (fig. 4.12J; plntEPSWT30134-3: 1 mM 0.39 ± 0.15 , $P < 0.001$; 10 mM 0.21 ± 0.09 , $P < 0.001$; Col-0: 1 mM 0.13 ± 0.09 , 10 mM 0.05 ± 0.04). However, there was no significant

difference in expression between genotype or nitrate concentration for *DREB26* or its putative target *NIR1*, although *NIR1* expression was higher in line plntEPSWT30134-3 than Col-0 at 1 mM KNO₃, albeit non-significantly (fig. 4.12B and C; Welch's t-tests, $P > 0.05$). Compared to Col-0 at both nitrate concentrations, line plntEPSWT30134-3 had significantly more lateral roots (fig. 4.12E,K; plntEPSWT30134-3: 1 mM 5.35 ± 1.92 , $P < 0.001$; 10 mM 4.05 ± 1.80 , $P < 0.001$; Col-0: 1 mM 2.14 ± 1.29 , 10 mM 1.69 ± 0.79), longer total lateral root length (fig. 4.12F; plntEPSWT30134-3: 1 mM 2.26 ± 1.16 cm, $P < 0.001$; 10 mM 0.95 ± 0.45 cm, $P < 0.001$; Col-0: 1 mM 0.62 ± 0.54 cm, 10 mM 0.22 ± 0.18 cm) and a higher lateral root density (fig. 4.12I; plntEPSWT30134-3: 1 mM 1.74 ± 0.79 , $P < 0.001$; 10 mM 1.21 ± 0.61 , $P < 0.001$; Col-0: 1 mM 0.59 ± 0.32 , 10 mM 0.46 ± 0.20). At 10 mM KNO₃, line plntEPSWT30134-3 had significantly higher average lateral root length (0.27 ± 0.18 cm; $P < 0.01$) than Col-0 (fig. 4.12G; 0.14 ± 0.12 cm). Additionally, at 1 mM KNO₃, line plntEPSWT30134-3 had significantly higher mean total root length (5.49 ± 1.12 cm; $P < 0.001$) than Col-0 (fig. 4.12H; 4.19 ± 0.92 cm).



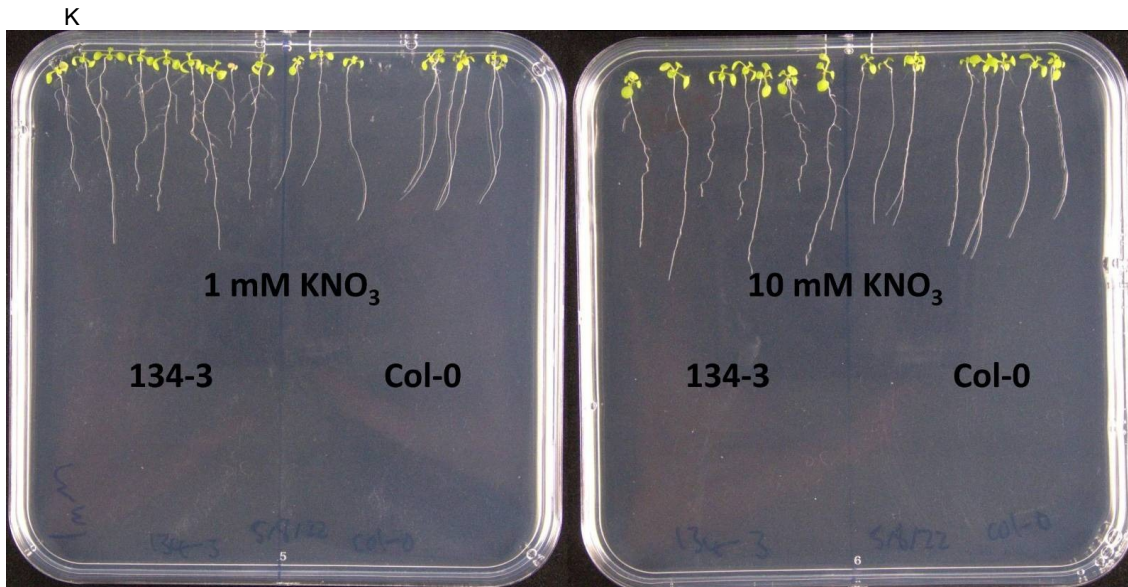


Figure 4.12: **Characterisation of line plntEPSWT30134-3 containing *dreb26-m1-5bp*.** A: Location of mutation relative to transcription start site. Relative expression of B: *DREB26* and C: *NIR1* in line plntEPSWT30134-3 and Col-0 plants on 1 mM and 10 mM KNO_3 . Significance was calculated using either independent t-tests (when variances were equal) or Welch's t-tests (when variances were not equal)³⁴⁹. qRT-PCRs performed using RNA extracted from roots 12 days after germination. $N = 3$ biological replicates, each consisting of a pool of roots from ~50 seedlings. Each qRT-PCR was performed in duplicate. D to J: RSA of line plntEPSWT30134-3 and Col-0 9 days after germination on 1 mM and 10 mM KNO_3 . Significance calculated using ANOVA and Tukey HSD post hoc tests when necessary³⁴⁴. $N = 8$ biological replicates (plants) and three technical replicates (plates). K: Photos of representative Col-0 and plntEPSWT30134-3 plants 9 days after germination grown on 1 and 10 mM KNO_3 . **, $P < 0.01$. *, $P < 0.05$. †, TFBS inferred from closely related TF, see section 3.5.1.

4.5 Discussion

This library of 327 lines provided a large resource for the identification of mutations of interest in the promoter/5' UTR regions of four genes. However, as sequence data was not obtained for all four genes in all 500 lines, there are likely to be more mutations and mutated plant lines than reported. Furthermore, only 200 T1 lines were selected and many more additional T1 seeds were collected that could be analysed in the future to identify further mutations. Additionally, only T2 progeny from 50 T1 parental lines were analysed, so further T2 progeny from the remaining 150 T1 parental lines could be analysed in the future. 41 lines contained mutations in more than one gene. This is most likely due to co-infection by multiple *Agrobacterium* strains. The number of sequencing reads per plant

line were sufficient for identification of mutations and zygosity, however, fewer reads were obtained for *DREB26* than for other genes. This is most likely because the *DREB26* target region was difficult to amplify, possibly due to several repetitive AT sequences in the *DREB26* amplicon. *DREB26* had the lowest percentage of mutated plants at 11.7%. This could be due to the lower number of successfully sequenced *DREB26* amplicons from different plant lines, the fact that sgRNAs are less efficient at targeting T-rich motifs, and also could be because fewer lines containing *DREB26* mutations were analysed by chance. In the future, the use of Cas12a (Cpf1) could improve targeting of T-rich motifs⁴⁵⁷.

3.7% of mutant lines were chimeric with presence of at least three different alleles. This suggests that some Cas9 activity occurred within somatic cells rather than within the egg cell or zygote⁴⁵². Due to time constraints, the presence of T-DNA had not yet been determined at the time of writing, therefore, some plant lines might still encode a functional Cas9. This could potentially result in the introduction of additional off-target mutations.

Overall, there were relatively few substitutions (6.2% of non-chimeric mutations), with most mutations being insertions (73.3% of non-chimeric mutations) followed by deletions (20.5% of non-chimeric mutations). This is consistent with previous studies using CRISPR in plants^{452,458-460}, which have similar ranges of mutations arising from NHEJ and, possibly, by microhomology-mediated end joining (MMEJ), typically observed when using Cas9 for mutagenesis in plants^{452,461,462}. The use of CRISPOR³⁵¹ to design sgRNAs likely improved the editing efficiency of the CRISPR constructs due to selection of better quality sgRNAs.

Lines plntEPSWT30069-9 and plntEPSWT30125-4 both showed an increased lateral root length:total root length ratio. Both plant lines contained deletions overlapping *NLP7* TFBSs in the *NLP7 cis*-regulatory region. These binding sites are suggestive of positive autoregulation of *NLP7*. At 1 mM KNO₃, the *NIR1* expression in both plant lines was lower than in Col-0, although the difference was not significant. This could be due to a lower expression of *NLP7* due to decreased autoregulation during the early response to nitrate.

The deletion in line plntEPSWT30069-9 also overlaps candidate binding sites for FRS9, BPC1, BPC5 and TFIIIA. It would be useful to experimentally determine whether these TFs bind to the candidate TFBSs using TF-DNA binding affinity assays, since candidate sites identified using PWMs are not always functional⁴⁶³ and might bind with higher

affinity to other related TFs from the same family³⁸⁹. FRS9, BPC1 and BPC5 are involved in the response to ethylene and stress responses^{464,465}. TFIID regulates the 5S rRNA gene⁴⁶⁶. This highlights that, even when targeting *cis*-regulatory regions, it is difficult to delete edges from TFs within one pathway without disrupting other edges related to other potential pathways. The biallelic *nlp7-m2-13bp/nlp7-m3+1bp* mutations in line plntEPSWT30125-4 also overlap candidate binding sites of ANR1, and in the case of the deletion, ANAC032. Since all three TFBSs bind TFs implicated in the nitrate response, the network dynamics in response to nitrate could be disrupted even more than if only a single TFBS was mutated. Interestingly, *NLP7* expression was higher at 1 mM KNO₃ in line plntEPSWT30125-4. This could be due to the deletion in the *ARF9* promoter, or it could be due to a feedback loop involving other genes due to the lower initial expression of *NLP7* causing an increase in *NLP7* expression longer term.

Lines plntEPSWT30069-9 and plntEPSWT30125-4 had similar RSAs, with longer total lateral root length, more lateral roots, higher total root length, higher lateral root density and a higher ratio of lateral root length to total root length on one or both nitrate concentrations compared to Col-0. This is most likely due to possible disruption of the timing and level of expression of the master regulator and nitrate sensor⁴⁹ *NLP7*. *NLP7* has a role in lateral root development through regulation of the IAA biosynthesis gene *TAR2*⁵². *NLP7* loss-of-function mutants have a lower auxin signal in the lateral root primordia⁵² and have a larger primary root and more lateral roots at 1 and 10 mM nitrate compared to wild type^{25,405}. The root phenotypes of line plntEPSWT30069-9 and line plntEPSWT30125-4 are consistent with *NLP7* loss-of-function phenotypes.

Previous studies have shown that overexpression of *NLP7* also increased primary root length and lateral root density in both nitrate rich and nitrate poor conditions, but to a greater extent than *NLP7* loss-of-function mutants⁴⁰⁵. Loss-of-function of *NLP7* confers a constitutive N-starved phenotype on high and low nitrate conditions. In contrast to *NLP7* loss-of-function mutants, *NLP7* overexpression led to higher root and shoot biomass at high and low nitrate conditions and resulted in up-regulation of genes involved in nitrate transport, assimilation, and signalling. *NLP7* loss-of-function mutants had lower biomass on high and low nitrate conditions. Since both *NLP7* overexpression and loss-of-function mutants both show similar RSA phenotypes⁴⁰⁵, it cannot be said for certain whether lines

plntEPSWT30069-9 and plntEPSWT30125-4 had increased or reduced expression of *NLP7*. In *NLP7* overexpression lines, *NIR1* expression is higher compared to Col-0, while in *NLP7* loss-of-function mutants the expression of *NIR1* is lower than in Col-0⁴⁰⁵. However, *NIR1* expression in both lines plntEPSWT30069-9 and plntEPSWT30125-4 was lower at 1 mM nitrate compared to Col-0, suggesting that the *NLP7* expression was lower in both lines. However, since *NLP7* expression was significantly higher in line plntEPSWT30125-4 at 1 mM nitrate compared to Col-0, and *NLP7* expression was slightly higher at 10 mM nitrate in line plntEPSWT3069-9 compared to Col-0, we cannot rule out that *NLP7* expression was increased by the mutations in its *cis*-regulatory region. Additionally, since the *nlp7-m1-11bp* deletion in line plntEPSWT3069-9 falls within the *NLP7* 5' UTR, disruption of translation efficiency due to disrupted ribosome binding sites in the mRNA cannot be discounted.

Line plntEPSWT30130-4 contained mutations *arf18-m1-14bp/arf18-m2+1bp* and *arf18-m3+1bp*. The mutations *arf18-m1-14bp/arf18-m2+1bp* overlapped binding sites for additional TFs (VRN1, RTV1, HCA2 and BPC1), and also BPC5 in the case of the deletion. The *arf18-m3+1bp* mutation also disrupted the binding site of the stress response gene BPC5⁴⁶⁷. VRN1 and RTV1 are known repressors involved with vernalisation for the flower development, although they have non-specific binding^{468,469}. Therefore, this mutation might lead to a change in flowering time. HCA2 is known to promote radial growth and elongation of roots⁴⁷⁰. Disruption of the HCA2 binding site might have implications for root growth through *ARF18*.

NIR1 expression was not significantly affected in line plntEPSWT30130-4. Since *ARF18* is known to repress *NIR1* (see chapter 3), it was expected that *NIR1* expression would be increased by a reduction in expression of *ARF18*. However, RNA-seq data in Gaudinier et al.²⁵ showed that *NIR1* expression was not significantly affected by *ARF18* loss-of-function. The only significant change in RSA in line plntEPSWT30130-4 was a higher total root length on 1 mM nitrate. This is generally consistent with the RSA of *ARF18* loss-of-function mutants, which did not have much change in RSA²⁵. The *ARF18-3* loss-of-function mutant had a significantly lower ratio of lateral root length to total root length compared to Col-0 at 1 mM KNO₃²⁵, which was the opposite trend shown in line plntEPSWT30130-4. This demonstrates the subtleties of promoter edge disruption compared to the loss-of-function

of the downstream gene.

Line plntEPSWT30139-9 contains a large deletion, *arf9-m2-31bp*, in the *ARF9* 5' UTR. Since the expression of *ARF9* in line plntEPSWT30139-9 did not change significantly compared to Col-0, it is likely that this mutation disrupted translation rather than transcription. Similar to line plntEPSWT30130-4 containing mutations in the *ARF18* promoter, the expression of *NIR1* in line plntEPSWT30139-9 was slightly higher than Col-0 at 10 mM KNO₃. The RSA phenotype of line plntEPSWT30139-9 was similar to that of line plntEPSWT30130-4, with a higher total root length than Col-0 on 1 mM KNO₃, although line plntEPSWT30139-9 also had an increased total lateral root length and ratio of lateral root length to total root length on 1 mM KNO₃. This contrasts with Gaudinier et al.²⁵, who found that the *ARF9-2* loss-of-function mutant had no significant root system architectural difference to Col-0.

Line plntEPSWT30134-3 contained the mutation *dreb26-m1-5bp* which disrupted the candidate TFBSs of the stress response genes BPC5, BPC1 and FRS9, along with TFIIIA and BPC6. BPC6 is involved with the ethylene response⁴⁶⁵, which is a known stress-responsive hormone besides its role in plant growth and development⁴⁶⁷. The consensus TFBS sequences of all of these TFs are highly repetitive TC sequences, suggesting that they are non-specific TFBSs. The fact that a DREB26 motif is present overlapping these stress-response related binding sites makes sense because DREB26 is involved with the drought stress and ethylene response⁴⁷¹. The expression of *DREB26* was slightly lower in line plntEPSWT30134-3 than Col-0 at 10 mM KNO₃, although the effect was not significant. There was a significant increase in the number of lateral roots, total lateral root length, lateral root density and ratio of lateral root length to total root length at both nitrate concentrations. This suggests that the disruption of the *DREB26* promoter had a positive effect on lateral root development. Evidence from validation of the regulatory subnetwork suggests that DREB26 is likely a regulator of *NLP7*, a known positive regulator of nitrate responses^{40,41,472}. It might, therefore, be hypothesised that DREB26 is a negative regulator of nitrate responses. Since the *dreb26-m1-5bp* deletion occurs in the 5' UTR, we cannot determine whether the RSA phenotypic changes are due to a change in transcription or translation efficiency.

It would be interesting to characterise more T1 lines to increase the size of the T2 mutation

library. This would allow for the identification of a greater range of mutations across each gene. It would also be useful to identify and grow the homozygous progeny of heterozygous and biallelic lines containing interesting mutations. Where only one allele disrupted a functional element, the analysis of homozygous lines might result in a clearer phenotype. A limitation of looking at gene expression of 12-day old plants is that the gene expression only represents the long-term nitrate response. It would be informative to look at the transcriptomics immediately and soon after changing the concentration of nitrate in the media. These results would allow for a better understanding of the early response to nitrate before later transcriptional cascades and feedback loops change the expression of genes of interest. Further work is needed to determine whether the *NLP7* expression is increased or decreased in the short and long-term in lines plntEPSWT30069-9 and plntEPSWT30125-4. The CRISPR library was a useful tool for studying the effect of disruptions between TFs and their target genes. The use of a pooled CRISPR library allowed for the generation of a large number of lines with a diverse range of mutations, and the phenotypic effects of some of these mutations were characterised. The use of a multiplexed sequencing approach was successful in identifying the mutations in the CRISPR lines. A potential downside of using a CRISPR library approach is that since mutations were generated by chance, the coverage of mutations across the promoter missed some predicted and tested TFBSs. The use of a complementary approach disrupting individual TFBSs of interest would be beneficial to guarantee that the TFBS of interest is disrupted. Alternatively, the number of screened lines could be increased to improve the likelihood of finding a mutation in the TFBS of interest. It will be important in the near future to determine whether the Cas9 construct was present in the lines that were characterised and whether the Cas9 construct had any effect on the phenotype by removing the Cas9 through breeding. This is needed before any solid conclusions can be made about the effect of the mutations on the phenotype. The presence of any off-target mutations should also be determined, and these removed by breeding, to ensure that the phenotypic effects were due to the known mutations in the CRISPR library. Additionally, multiple lines should be characterised with various mutations in the same promoters to determine whether the phenotypic effects are consistent, or specific for individual mutations. The fact that root phenotypes were altered in characterised lines suggests that NUE might also have been altered, as the surface area

in contact with the growth medium will have been changed. It would be interesting to characterise the NUE of these lines to determine whether the mutations in the CRISPR library have an effect on NUE. This would involve collecting additional data such as plant biomass, plant tissue nitrogen content and nitrogen uptake. It would also be useful to generate additional CRISPR libraries targeting more genes of interest, to increase the potential for identifying interesting mutations which improve NUE. Finally, it would be beneficial to use similar CRISPR library approaches in related crop species to determine how consistent the results are with those found in *A. thaliana*.

4.5.1 Conclusions

Overall, the mutations in the *NLP7* promoter disrupted potential edges from ANAC032, ANR1 and *NLP7*, causing longer and more dense lateral roots. In the previous chapter, ANAC032 was shown to repress *NLP7* expression, disrupting part of the feedforward loop from ARF18 to *NLP7*. This edge disruption could potentially have disrupted network dynamics and reduced the robustness of the nitrogen-response subnetwork, causing the altered lateral root growth. The disrupted ANR1 and *NLP7* TFBSs in the *NLP7* promoter might also have disrupted network dynamics, causing a change in either the time of or intensity of expression of *NLP7*.

The deletion in the *ARF18 cis*-regulatory region caused a reduction in *ARF18* expression, which slightly increased total root length. The deletion in the *ARF9 cis*-regulatory region also caused an increased total root length, suggesting that ARF9 and ARF18 have similar functions. This could have reduced the overall repression effect of ARF9/18 repressors on genes in the nitrate-response subnetwork in response to auxin, leading to slightly higher root growth. The activity of ARF TFs is known to be functionally redundant⁴⁵⁶, so disrupting the *cis*-regulatory regions of multiple repressor ARFs might have a greater effect on the network dynamics of the nitrate-response subnetwork and would potentially cause more obvious phenotypes.

Although the 5 bp deletion in the *DREB26* 5' UTR did not cause a significant change in *DREB26* expression, it might have changed ribosome binding affinity causing a lower concentration of DREB26 protein, leading to a derepression of DREB26 targets. The change in network dynamics caused by this deletion increased the number and density

of lateral roots. The putative reduction in DREB26 protein potentially disrupted the feedforward loop from ARF18 to *NLP7* through DREB26, which might have increased *NLP7* expression and caused the increased lateral root growth. This phenotype was similar to the effect of the mutation disrupting the ANAC032 TFBS in the *NLP7 cis*-regulatory region (line plntEPSWT30125-4), which potentially disrupted the other half of the feedforward loop from ARF18 through *ANAC032* to *NLP7*. It would be interesting to cross lines plntEPSWT30125-4 and plntEPSWT30134-3 to see if the disruption of both ANAC032 and DREB26 feedforward loop edges is additive to the phenotype.

In conclusion, the CRISPR library contained a range of interesting mutations, some of which were shown to disrupt the regulation of four TFs involved in the early response to nitrate. This demonstrates that CRISPR libraries may be useful engineering strategies for engineering crop traits. The CRISPR library generated in this chapter will be an important resource for further investigations of the N-response gene regulatory network, including understanding network dynamics and identifying genotypes that lead to improvements in NUE.

Chapter 5

Architectural differences in the promoters of constitutive and variably expressed genes

5.1 Preface

All analyses in this chapter were performed by myself.

5.2 Introduction

Constitutively expressed genes are defined as genes that have consistent expression levels under normal (absence of stress) conditions across all cell and tissue types and developmental stages^{473,474}. In contrast, variable genes include developmental stage, cell-specific genes where expression is mainly confined to one or a few tissues^{474,475}, as well as responsive genes, which respond to changes in the environment including predictable changes such as light and dark but also biotic and abiotic stress. It is important to note that the degrees of expression variability and tissue-specificity are on a spectrum, and even constitutively expressed genes show significant cell to cell variability in Arabidopsis⁴⁷⁶. This chapter aims to investigate differences in upstream regulatory regions of Arabidopsis genes to identify if there are architectural differences in the upstream regulatory regions of constitutive and variable genes that might aid the design of minimal synthetic promoters. Previous

analyses in plants and other eukaryotes have already identified differences between genes within these categories. Arabidopsis constitutive genes are enriched for methylation in their gene body regions^{477–479} which may help them maintain transcription levels. Tissue-specific genes are enriched for methylation in their promoters⁴⁷⁷, which may help prevent gene expression in tissues or cells where they are not required. Additionally, while few genes contain TATA boxes¹⁵⁷, TATA-containing promoters have a larger GC-skew (higher proportion of G and C compared to the background) peaking at the TSS⁴⁸⁰. In addition, Arabidopsis constitutive promoters are enriched for GA repeat regions while variable promoters are depleted in GA repeat regions⁴⁸¹. GA repeats have been hypothesised to have a similar function to CpG (CG) islands in mammalian promoters, which are also enriched in constitutive promoters as compared to variable promoters^{481–483}. While CpG dinucleotides are often methylated, those within promoter CpG islands are usually unmethylated⁴⁸⁴. Perhaps related to the presence of CpG islands, mammalian constitutive promoters have a higher GC content than responsive promoters^{485,486}. GC-rich regions show higher flexibility in^{487,488} and lower nucleosome formation potential^{489,490} in animals and plants. In Arabidopsis, GC-rich sequence has a higher DNase I sensitivity with more open chromatin⁴⁹⁰. Together, these findings suggest that the promoter regions of constitutive genes may maintain accessibility to their regulating TFs via a high GC content in confirmations that do not induce methylation. It can therefore be hypothesised that Arabidopsis constitutive promoters will have a higher GC content than variable promoters. In plants, the CDSs of constitutively expressed, housekeeping genes have been shown to be under stronger selective constraint than tissue-specific genes⁴⁹¹. In contrast, promoter sequences are not as conserved as CDSs. For example, strength of expression but not primary sequence is conserved between Arabidopsis and *Oryza sativa* (rice)⁴⁹². Highly expressed plant constitutive genes were more conserved than highly expressed tissue-specific genes and also more conserved than constitutive genes with low expression, however, highly expressed tissue-specific genes were less conserved than tissue-specific genes with low expression⁴⁹¹.

Another interesting difference between constitutive and variable genes are differences in TSSs. Morton et al.¹⁵⁷ used paired-end analysis of TSSs (PEAT) to create a dataset of TSSs from *A. thaliana* root samples. This dataset was used to create a machine learning

model called the Plant PEAT Peaks (3PEAT) model, which predicts the TSS probability at any given nucleotide within a TSR. Variable genes were associated with narrow TSRs, while weaker expression was associated with constitutive expression¹⁵⁷. This was also true in humans, with variably expressed genes having a narrower TSRs than constitutive genes⁴⁸³.

In plants, though few genes have TATA boxes, they are enriched in genes that respond to multiple stimuli as compared to tissue-specific genes, and early responders to stress were found to be more likely to contain a TATA box than genes that responded later⁴⁹³. Furthermore, TATA boxes were found to be enriched in variable genes compared to constitutive genes in *Arabidopsis*⁴⁹⁴.

Differences in sequence composition, open chromatin, and the breadth of the TSR might also suggest differences in the numbers and diversity of TFs that bind to promoter regions of constitutive and variable genes. Studies in humans show conflicting evidence for differences in TFBS coverage between promoter types. Constitutive genes were found to have higher DNA entropy (less order) than tissue-specific genes in humans, suggesting that tissue-specific genes are more complex and have a higher density of CREs⁴⁹⁵. Constitutive promoters were also found to have more nucleotides covered by TFBSs than tissue-specific promoters in humans²⁷⁴, suggesting that the promoters of constitutive genes are more complex. Human constitutive promoters were also found to have more TFBSs that attract multiple TFs²⁷⁴, and in pigs, more types of motifs were found in variable promoters than in constitutive⁴⁸⁶. In *Arabidopsis*, genes responding to many stimuli were found to be targeted by more TFs than those that respond to few stimuli, and the first 500 bp upstream of the TSS were found to contain a stronger correlation between breadth of stimuli response and motif density⁴⁹³. Stochastically expressed genes were found to be shorter and targeted by more TFs than constitutive genes in *Arabidopsis*⁴⁹⁶. Together with the observations above, this suggests that the functional regions of variable promoters may be short, with the TFs regulating them binding close to the TATA box, potentially also contributing to a narrow TSR. However, as previous observations have been based on comparisons of genes that respond to different numbers of stimuli, further analyses are required to determine differences between constitutive and variably expressed genes.

Some genes in GRNs have more connections than others, and these are called hubs. These

genes can regulate the activity of many other genes in the network. Regulating nodes or hubs within GRNs are less likely to be tissue-specific than their target genes. In humans, nearly all TFs are associated with at least one tissue-specific edge in all tissues, suggesting that even constitutively expressed TFs play a role in regulating tissue-specific expression⁴⁹⁷. Network hub genes which interact with lots of other genes were more likely to be non-tissue specific. Tissue-specific genes were less likely to be regulated by canonical TF-DNA interactions in their promoter and were more likely to be regulated by non-canonical means through TF-complexes, alternative TFBSs or interactions outside of the promoter⁴⁹⁷.

5.3 Aims

The first aim is to elucidate promoter architectural differences between different gene expression categories in Arabidopsis. Specifically, whether TATA boxes, TFBS density, GC content, TFBS diversity and chromatin percentage coverage of CRMs differ between constitutive, variable, non-specific and tissue-specific genes. The second aim is to learn design features needed to construct synthetic minimal N-responsive promoters.

5.4 Results

5.4.1 Gene selection

Four categories of genes with different expression patterns in *A. thaliana* were selected to study promoter architecture. Before gene selection, 3299 genes were flagged and removed from the analysis as their coding sequences were overlapping other coding sequences. 484 genes were also filtered from the analysis as the promoter/5' UTR did not generate significant matches to known TFBS motifs identified using DAP-seq¹⁸¹. Sets of constitutive and variable genes were selected from a published dataset (Czechowski et al.³⁷²) in which Arabidopsis genes were ranked based on their coefficient of variation (CV) to identify stably expressed housekeeping genes (see section 2.7.3). The methodology described by Czechowski et al.³⁷² was used to select the top 100 and top 300 constitutive and variable genes, gene categories which were not described in Czechowski et al.³⁷². However, only genes that were found to be expressed in 80% or more of developmental stages and tissues were included in that analysis, meaning that even the most variable genes excluded

tissue/condition-specific genes. Tissue/condition-specific genes ranked using Tau tissue-specificity were therefore also compared with genes with the lowest Tau ranking, which were categorised as non-specific genes. Tau tissue-specificity was calculated using the gene expression data from Czechowski et al.³⁷², and the top 100 and 300 non-specific and tissue-specific genes were selected. Using a custom Python script, 1959 genes were flagged and removed as they had potentially overlapping or bidirectional promoters where the upstream gene was positioned in the opposite direction and was less than 2000 bp away from the TSS. Of these, 17 constitutive, 6 variable, 15 non-specific, 14 tissue-specific and 11 control genes were flagged and removed. The mean \log_2 expression of the genes in each category are shown in fig. 5.1. To select 100 or 300 control genes that were present in both the CV and Tau datasets, the Tau ranking was used to select genes from the central distribution of the expression CV ranked genes. The central distribution was divided into 10 bins and 10 or 30 genes were selected randomly from each bin. For analysis of promoters/5' UTRs, hereafter called *cis*-regulatory modules (CRMs), 1000 bp upstream of the annotated Araport 11³⁶³ TSS or until the nearest gene were extracted. 5' UTRs were extended downstream of the TSS to the closest coding region. As can be seen in fig. 5.1, the tissue-specific genes were only expressed in a small number of tissues, so any conclusions from further analyses are only relevant for these tissues and should not be generalised to all tissues.

The expression CV distribution of the top 100 constitutive genes was narrower than the top 100 variable genes (fig. 5.2). Additionally, the Tau tissue-specificity distribution of the top 100 tissue-specific genes was narrower than the top 100 non-specific genes (fig. 5.3).

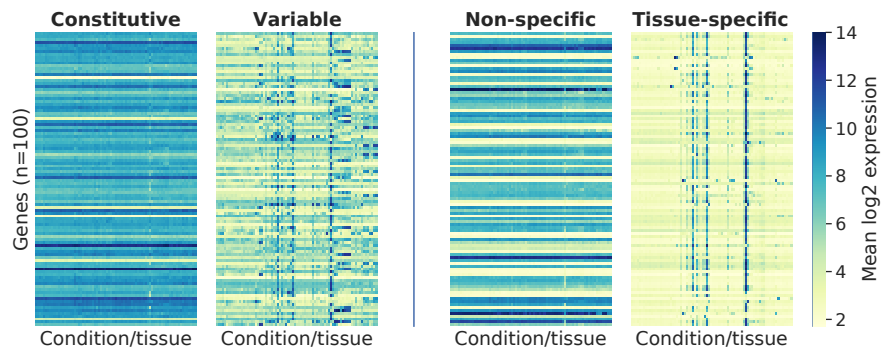


Figure 5.1: Mean \log_2 expression of constitutive, variable, non-specific and tissue-specific gene categories in *Arabidopsis thaliana* across 79 different tissues and developmental stages³⁷². Constitutive and variable categories were selected using coefficient of variation ranking as in Czechowski et al.³⁷². Tissue-specific and non-specific categories were selected using Tau tissue specificity ranking. $N = 100$. The top 100 genes (Y-axis) in each gene expression category are coloured by mean \log_2 expression in 79 different tissues and developmental stages (X-axis)³⁷².

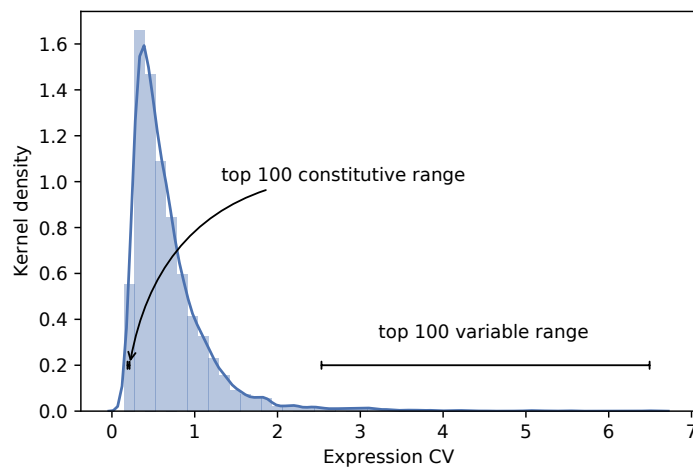


Figure 5.2: Expression CV distribution of the top 100 constitutive and variable genes compared to all CV ranked genes. Constitutive and variable categories were selected using coefficient of variation ranking as in Czechowski et al.³⁷².

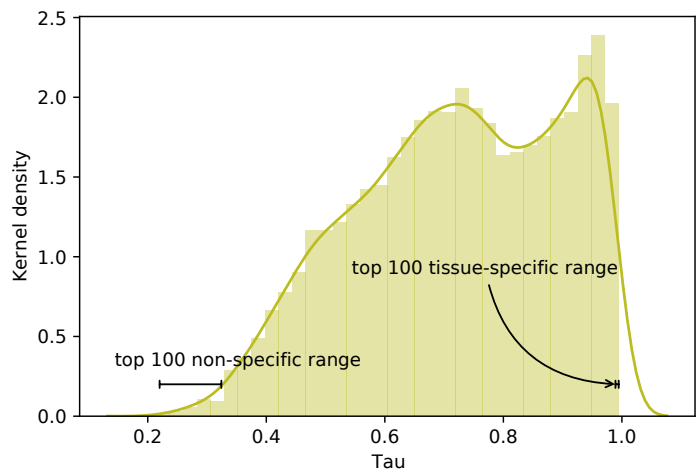


Figure 5.3: **Tau tissue-specificity distribution of the top 100 non-specific and tissue-specific genes compared to all Tau ranked genes.** Tissue-specific and non-specific categories were selected using Tau tissue specificity ranking.

5.4.2 Gene ontology analysis

A gene ontology (GO) enrichment analysis using Fisher's exact test³⁷⁹ and Benjamini/Hochberg FDR correction³⁷¹ found 2 significantly enriched GO terms (biological process GO:0008150) for the top 300 constitutive genes, with 17 genes associated with intracellular protein transport and 9 genes with endocytosis (see section 2.7.10). There were 4 significantly enriched GO terms (cellular component GO:0005575) for the top 300 variably expressed genes, with 73 associated with the nucleus, 55 with being extracellular, 32 with the cytosol and 3 with the plastid. For the top 300 non-specific genes there were 3 significantly enriched cellular component GO terms, with 67 genes associated with the mitochondrion, 60 genes with the cytosol and 28 genes with the Golgi apparatus. For the top 300 tissue-specific genes there were 4 biological process, 7 cellular component and 6 molecular function significantly enriched GO terms. 151 tissue-specific genes were associated with the pollen tube (pollen tube growth GO:0009860, regulation of pollen tube growth GO:0080092, pollen tube guidance GO:0010183, pollen exine formation GO:0010584, pollen tube GO:0090406, pollen tube tip GO:0090404), 67 with the extracellular region, 31 with the chloroplast, 26 with protein binding, 15 with antiporter activity, 14 with the cytosol, 8 with the endomembrane system, 6 with apical plasma membrane, 8 with pectinesterase inhibitor activity and 8 with pectinesterase activity. A KEGG gene set enrichment analysis with Benjamini/Hochberg FDR correction found no enriched terms in constitutive or variable groups ($P > 0.05$).

5.4.3 Analysis of chromatin availability

Publicly available ATAC-seq data (control treated plants³⁵⁰) were superimposed onto the promoter/5' UTRs (CRMs) from the four gene sets (see section 2.7.7). This allowed the chromatin accessibility to be described. The percentage of nucleotides within open chromatin was significantly different between constitutive, variable and control CRMs (Kruskal-Wallis $H = 37.5$, $P < 0.0001$). Dunn's post hoc tests with Bonferroni correction showed that the percentage of nucleotides within open chromatin was significantly lower in variable CRMs ($35.0\% \pm 35.1$) than constitutive ($54.1\% \pm 33.3$; $P < 0.001$) or control ($65.3\% \pm 30.4$; $P < 0.0001$) CRMs (fig. 5.4A). Open chromatin coverage was also significantly different between non-specific, tissue-specific and control gene types (Kruskal-

Wallis $H = 75.3$, $P < 0.0001$). Dunn's post hoc tests with Bonferroni correction revealed that the percentage of nucleotides within open chromatin in tissue-specific CRMs ($20.4\% \pm 30.2$) was significantly lower than in non-specific ($46.9\% \pm 36.1$; $P < 0.0001$) or control ($65.3\% \pm 30.4$; $P < 0.01$) CRMs (fig. 5.4B)

A sliding window analysis revealed that from ~ 350 bp to ~ 650 bp upstream of the start codon the median open chromatin in constitutive CRMs decreased from 100% to zero while variable CRMs had a median percentage open chromatin of 0 across the entire region (fig. 5.5A) (see section 2.7.4). From ~ 250 bp to ~ 450 bp upstream of the start codon the median open chromatin in non-specific CRMs decreased from 100% to zero while tissue-specific CRMs had a median percentage open chromatin of 0 across the entire region (fig. 5.5B).

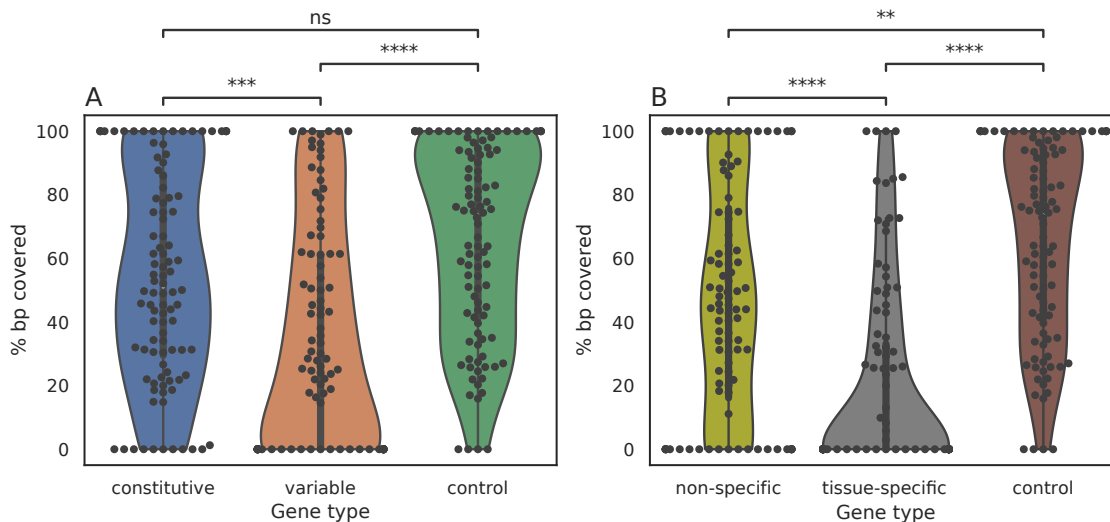


Figure 5.4: **Mean percentage open chromatin in Arabidopsis CRM expression categories.** A: Percentage base pairs (bp) within open chromatin in 100 constitutive, variable and control genes. $N=100$. B: Percentage bp within open chromatin in 100 non-specific, tissue-specific and control genes. $N=100$. Significance was calculated using Kruskal-Wallis and Dunn's post hoc tests with Bonferroni correction. **, $P < 0.01$. ***, $P < 0.001$. ****, $P < 0.0001$. Control genes are identical in plots A and B.

To test whether this was a significant difference or due to different 5' UTR lengths between constitutive and variable genes, the percentage of root-shoot intersect open chromatin of windows was plotted centred around the Araport11 TSS. This analysis separated windows located in promoter regions from those located in 5' UTRs regions. The median percentage of open chromatin decreased upstream of the TSS for both constitutive and variable genes,

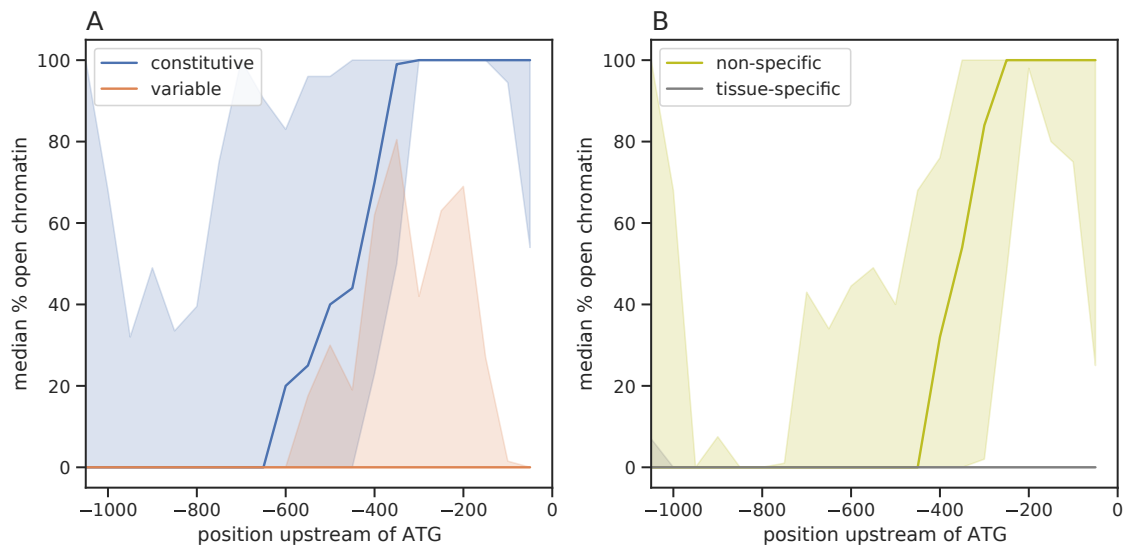


Figure 5.5: **Sliding window analysis of percentage open chromatin in Arabidopsis CRMs.** The median proportion of nucleotides within open chromatin was scored for 100 bp windows across selected CRMs. The windows were overlapping with a 50 bp offset. The median value for each bin is displayed at the central bp of the bin. Shading represents 95% confidence intervals estimated using 10000 bootstraps. A: constitutive and variable GC content sliding windows. $N=100$. B: non-specific and tissue-specific sliding windows. $N=100$.

and constitutive genes had a higher percentage open chromatin -100 to +500 bp around the TSS (fig. 5.6A). The median percentage of open chromatin decreased upstream of the TSS for non-specific genes while tissue-specific CRMs had a median percentage open chromatin of 0 across the whole CRM (fig. 5.6B). When analyses were repeated with the top 300 CRMs in each category the results supported the findings of the top 100 CRMs (figs. S4 to S6).

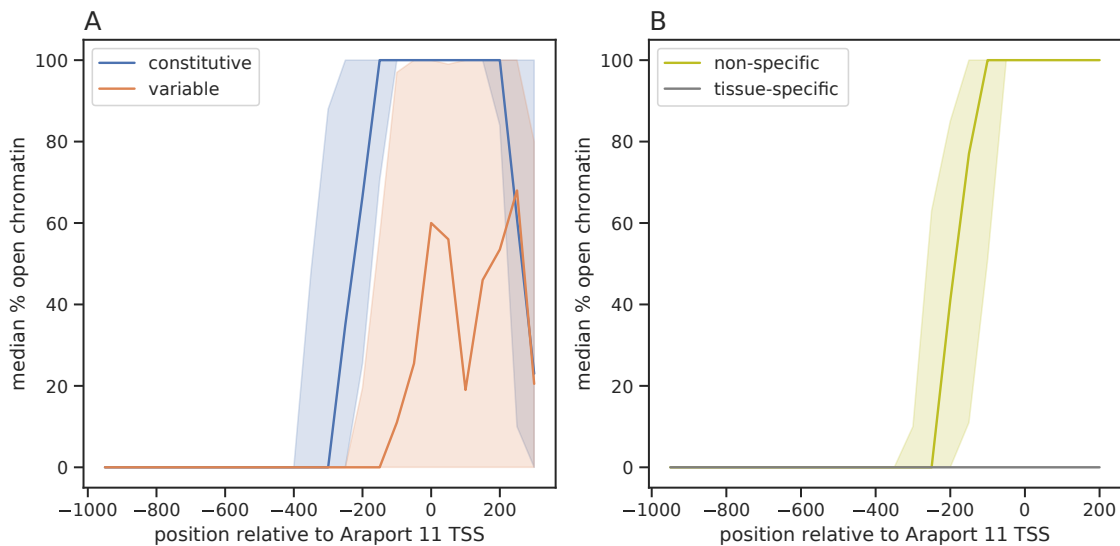


Figure 5.6: **Percentage open chromatin in Arabidopsis CRMs centred around the transcription start site (TSS)**. Windows are offset by 50 bp. Shading represents 95 % confidence intervals estimated using 10000 bootstraps. Median percentage of open chromatin peaks overlapping 100 bp windows. The median value for each bin is displayed at the central bp of the bin. Open chromatin peaks derived from the intersect of root and shoot peaks derived from negative control (treated with NaOH) ATAC-seq data by Potter et al.³⁵⁰. A: constitutive and variable GC content sliding windows. $N=100$. B: non-specific and tissue-specific sliding windows. $N=100$.

5.4.4 Analysis of GC content

GC content was analysed for the top 100 genes of each category (see section 2.7.5). GC content was significantly different between gene types (Kruskal-Wallis $H = 6.5$, $P < 0.05$). Dunn's post hoc tests with Bonferroni correction revealed that mean percentage GC content was significantly higher in constitutive CRMs ($32.8\% \pm 4.3$) than variable CRMs ($31.7\% \pm 3.6$; $P < 0.05$) (fig. 5.7A). There was no significant difference between non-specific, tissue-specific and control categories (Kruskal-Wallis $H = 4.2$, $P > 0.05$; (fig. 5.7B).

A sliding window analysis (see section 2.7.4) revealed that percentage GC content was higher for constitutive genes than variable genes and higher for non-specific genes than tissue-specific genes (fig. 5.8).

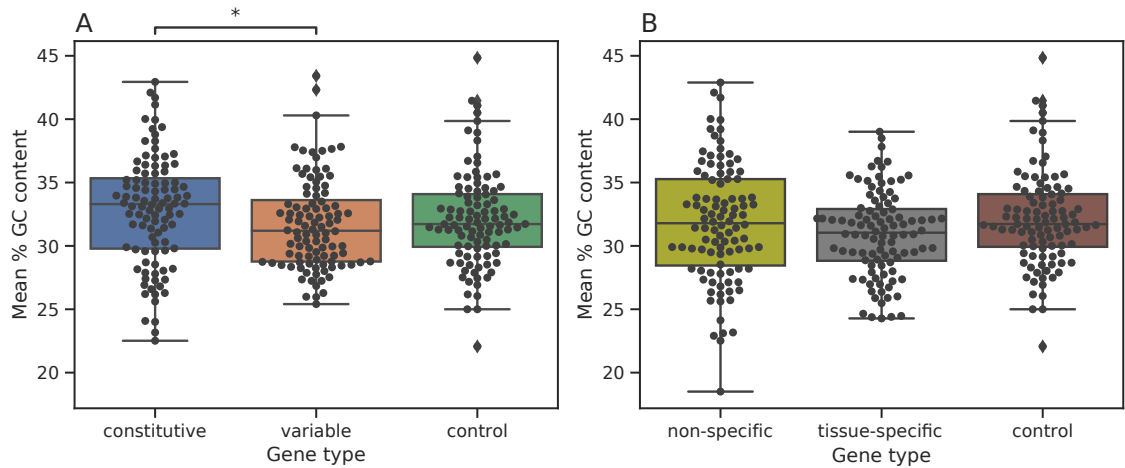


Figure 5.7: **Mean percentage GC content in Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. A: Percentage GC content in 100 constitutive, variable and control genes. *, $P < 0.05$. $N=100$. B: Percentage GC content in 100 non-specific, tissue-specific and control genes. There was no significant difference in GC content between categories. $N=100$. Control genes are identical in plots A and B.

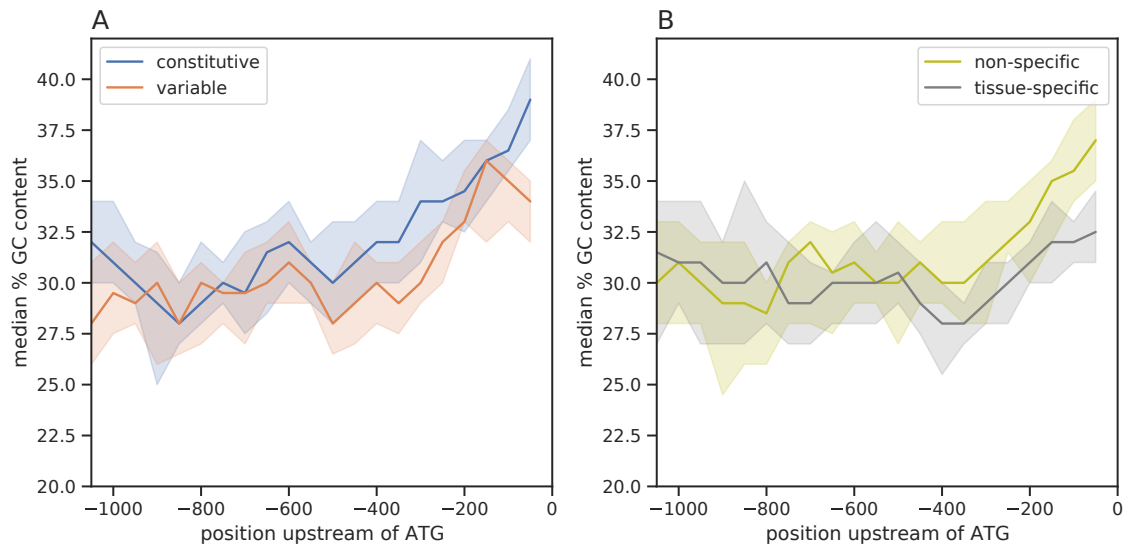


Figure 5.8: **Sliding window analysis of percentage GC content in Arabidopsis CRMs.** The median percentage GC content was scored for 100 bp windows across selected CRMs. The windows were overlapping with a 50 bp offset. The median value for each bin is displayed here at the central bp of the bin. Shading represents 95 % confidence intervals estimated using 10000 bootstraps. A: constitutive and variable GC content sliding windows. $N=100$. B: non-specific and tissue-specific sliding windows. $N=100$.

When the initial analysis was limited to 400 bp upstream of the ATG start codon, there was a significant difference between constitutive, variable and control genes (Kruskal-Wallis $H = 14.8$, $P < 0.001$) (fig. 5.9). Dunn's post hoc tests with Bonferroni correction revealed

that, within this 400 bp region, CRMs of constitutive genes had a significantly higher GC content ($35.0\% \pm 5.0$) than variable CRMs ($33.1\% \pm 4.7$; $P < 0.001$) and control CRMs ($33.8\% \pm 4.2$; $P < 0.05$). In the same 400 bp CRM region there was a significant difference between non-specific, tissue-specific and control CRMs (Kruskal-Wallis $H = 22.8$, $P < 0.0001$). Dunn's post hoc tests with Bonferroni correction revealed that non-specific ($33.6\% \pm 5.6$; $P < 0.001$) and control CRMs ($33.8\% \pm 4.2$; $P < 0.0001$) had a significantly higher GC content than tissue-specific CRMs ($31.2\% \pm 3.9$). When analyses were repeated using the top 300 CRMs in each category, all GC content findings supported those found when using the top 100 CRMs (figs. S7 to S9).

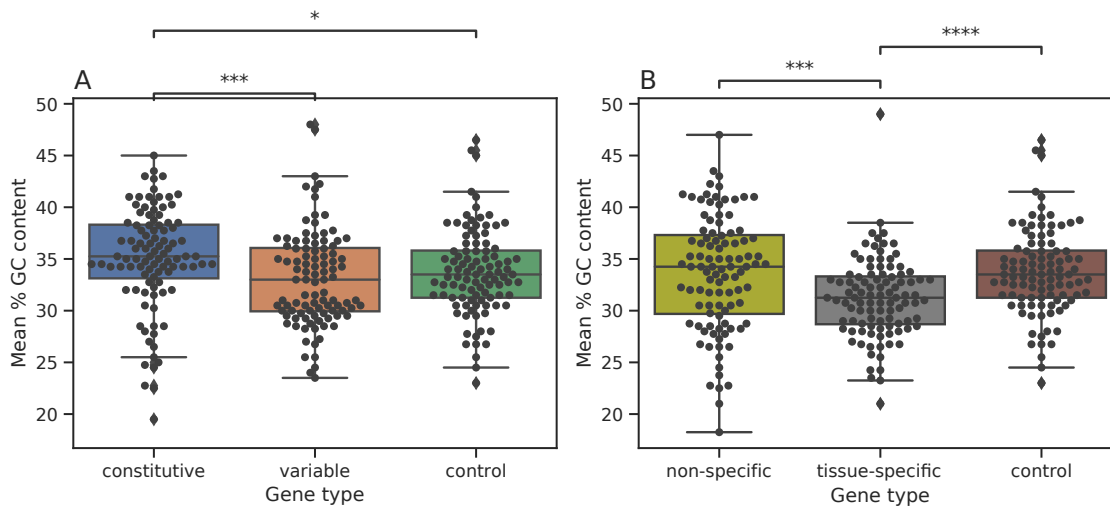


Figure 5.9: **Mean percentage GC content of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Significance was calculated using Kruskal-Wallis followed by Dunn's post hoc tests with Bonferroni correction. A: Percentage GC content in 100 constitutive, variable and control CRMs. $N=100$. B: Percentage GC content in 100 non-specific, tissue-specific and control genes. $N=100$. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. Control genes are identical in plots A and B.

5.4.5 Transcription factor binding site coverage

TFBS coverage within CRMs was compared between the top 100 genes of each category to investigate if variable CRMs have higher coverage, using TFBS motif data from DAP-seq¹⁸¹ (see section 2.7.6). TFBS coverage was not significantly different between CRM types

(Kruskal-Wallis $H = 2.4$, $P > 0.05$; fig. 5.10A). TFBS coverage did not differ significantly between non-specific, tissue-specific or control CRM types at the whole promoter level (Kruskal-Wallis $H = 5.6$, $P > 0.05$; fig. 5.10B).

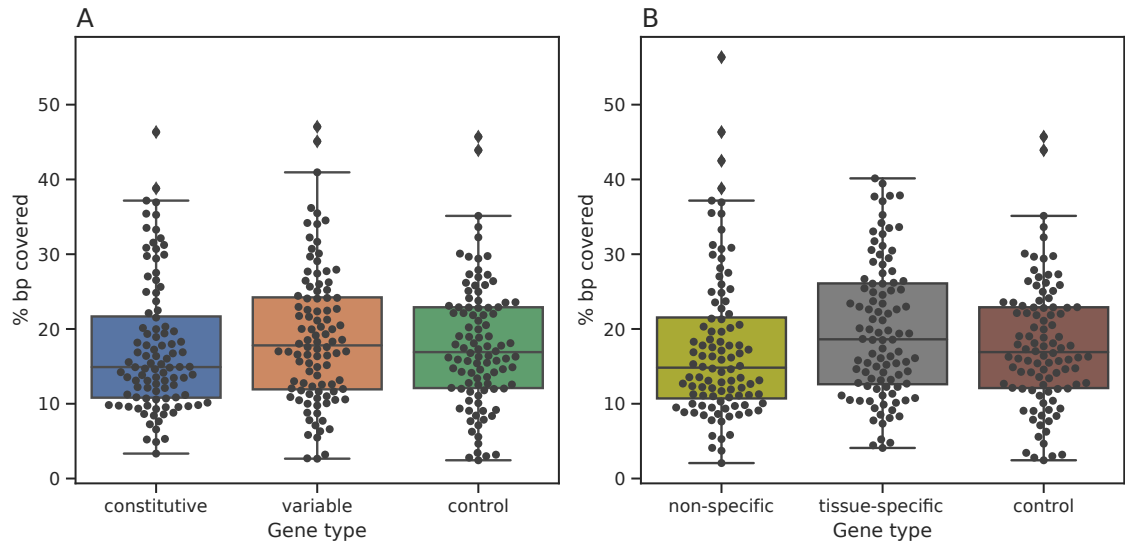


Figure 5.10: **Mean percentage TFBS coverage of Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Significance was calculated using Kruskal-Wallis. A: There was no significant difference in TFBS coverage between constitutive, variable or control CRMs. $N=100$. B: There was no significant difference in TFBS coverage between non-specific, tissue-specific or control CRMs. $N=100$. Control genes are identical in plots A and B. TFBS motif data was obtained from DAP-seq¹⁸¹.

The same analysis was performed restricting analysis to regions of open chromatin. TFBS coverage differed significantly between constitutive, variable and control categories (Kruskal-Wallis $H = 27.1$, $P < 0.0001$; fig. 5.11A). Dunn's post hoc tests with Bonferroni correction revealed that TFBS coverage in open chromatin of variable CRMs ($7.3\% \pm 8.8$) was significantly lower than constitutive CRMs ($11.3\% \pm 8.5$; $P < 0.001$) and control CRMs ($12.9\% \pm 9.1$; $P < 0.0001$). TFBS coverage also differed significantly between non-specific, tissue-specific and control categories (Kruskal-Wallis $H = 72.9$, $P < 0.0001$; fig. 5.11B).

Dunn's post hoc tests with Bonferroni correction revealed that TFBS coverage in open chromatin of tissue-specific CRMs ($3.5\% \pm 6.0$) was significantly lower than non-specific CRMs ($10.0\% \pm 10.1$; $P < 0.0001$) and control CRMs ($12.9\% \pm 9.1$; $P < 0.0001$). Non-specific genes had a significantly lower TFBS coverage in open chromatin than control genes ($P < 0.05$).

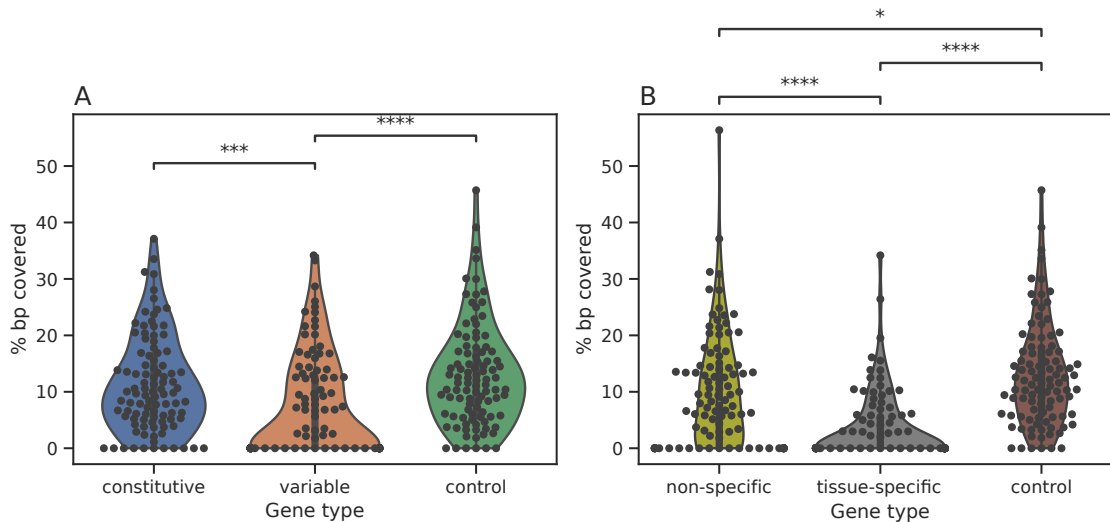


Figure 5.11: **Mean percentage coverage of TFBS within open chromatin regions of Arabidopsis CRMs.** Significance was calculated using Kruskal-Wallis followed by Dunn's post hoc tests with Bonferroni correction. A: Mean percentage coverage of TFBS falling within open chromatin in constitutive, variable and control CRMs. $N=100$. B: Mean percentage coverage of TFBS falling within open chromatin in non-specific, tissue-specific and control CRMs. $N=100$. ****, $P < 0.0001$. ***, $P < 0.001$. *, $P < 0.05$. Control genes are identical in plots A and B. TFBS motif data was obtained from DAP-seq¹⁸¹.

A sliding window analysis revealed that TFBS coverage was higher in variable CRMs than constitutive CRMs (fig. 5.12A) and higher in tissue-specific CRMs than in non-specific CRMs (fig. 5.12B) in the first 400 bp region upstream of the ATG start codon (see section 2.7.4).

Further analysis of this 400 bp region revealed a significant difference between constitutive, variable and control CRM types (Kruskal-Wallis $H = 10.5$, $P < 0.01$; fig. 5.13A). Dunn's post hoc tests with Bonferroni correction showed that variable promoters had a significantly higher percentage bp covered ($26.1\% \pm 14.6$) than constitutive CRMs ($20.0\% \pm 13.7$; $P < 0.01$) and control CRMs ($21.1\% \pm 13.8$; $P < 0.05$).

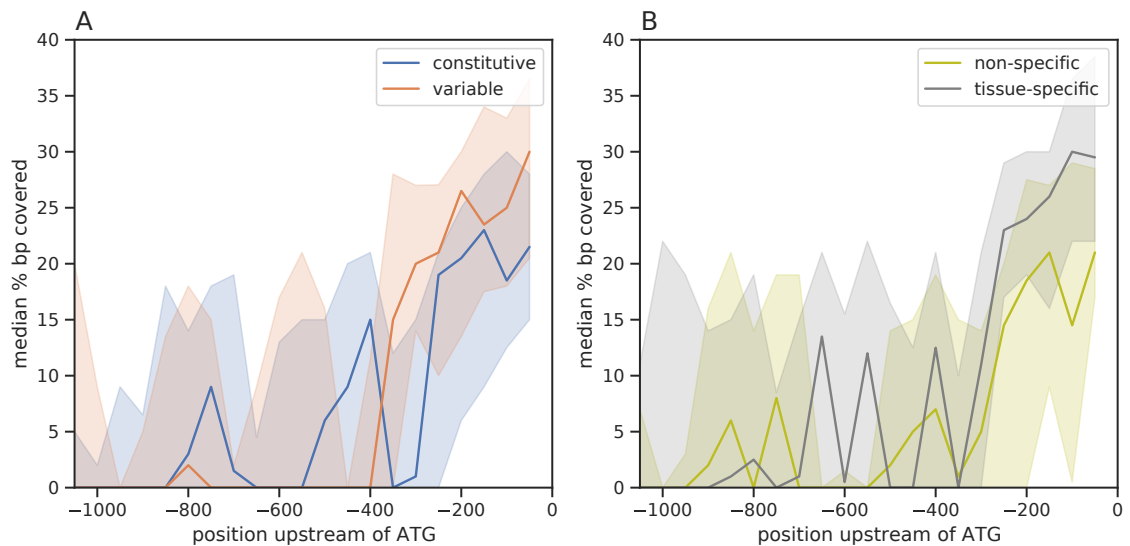


Figure 5.12: **Sliding window analysis of percentage TFBS coverage in Arabidopsis CRMs.** The median percentage TFBS coverage was scored for 100 bp windows across selected CRMs. The windows were overlapping with a 50 bp offset. The median value for each bin is displayed here at the central bp of the bin. Shading represents 95 % confidence intervals estimated using 10000 bootstraps. A: constitutive and variable GC content sliding windows. $N=100$. B: non-specific and tissue-specific sliding windows. $N=100$. TFBS motif data was obtained from DAP-seq¹⁸¹.

In the same 400 bp region there was a significant difference between non-specific, tissue-specific and control CRM types (Kruskal-Wallis $H = 8.5$, $P < 0.05$; fig. 5.13B). Dunn's post hoc tests showed that tissue-specific CRMs had significantly higher TFBS percentage coverage ($25.1\% \pm 14.6$) than non-specific CRMs ($19.4\% \pm 13.3$; $P < 0.05$).

Percentage coverage of TFBSs falling within open chromatin in the 400 bp region was significantly different between constitutive, variable and control CRMs (Kruskal-Wallis $H = 13.2$, $P < 0.01$; fig. 5.14A). Dunn's post hoc tests with Bonferroni correction showed that variable CRMs had a significantly lower TFBS coverage ($10.5\% \pm 14.2$) in open chromatin than constitutive ($15.2\% \pm 13.4$; $P < 0.05$) and control CRMs ($16.1\% \pm 14.0$; $P < 0.01$).

Within the same 400 bp region non-specific, tissue-specific and control CRMs also had significantly different TFBS coverage in open chromatin (Kruskal-Wallis $H = 74.7$, $P < 0.0001$; fig. 5.14B). Dunn's post hoc tests with Bonferroni correction revealed that tissue-specific CRMs had significantly lower TFBS coverage ($2.5\% \pm 7.0$) in open chromatin than non-specific ($13.4\% \pm 14.0$; $P < 0.0001$) and control CRMs ($16.1\% \pm 14.0$; $P < 0.0001$).

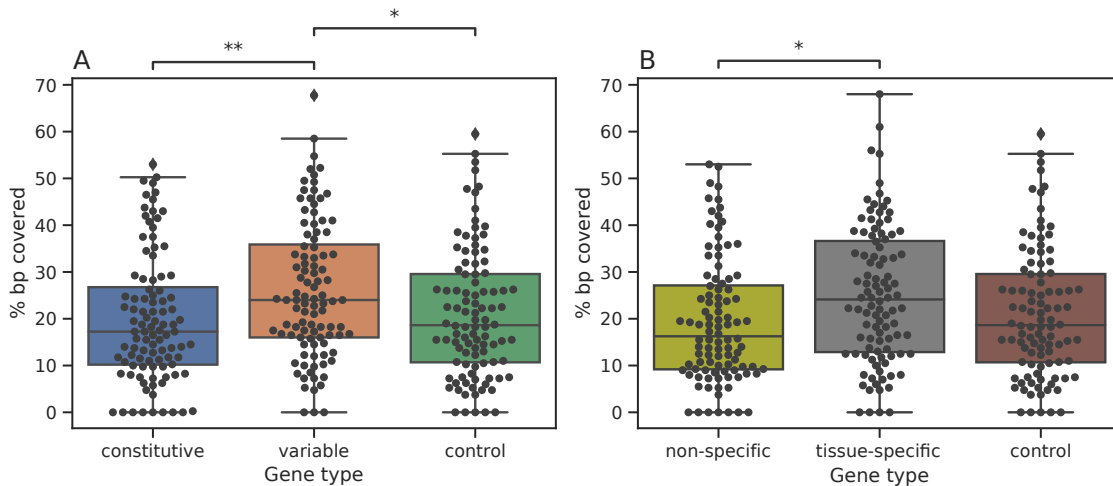


Figure 5.13: **Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. A: Mean percentage TFBS coverage in constitutive, variable and control CRMs. $N=100$. B: Mean percentage TFBS coverage in non-specific, tissue-specific and control CRMs. $N=100$. Significance was calculated using Kruskal-Wallis and Dunn's post hoc tests with Bonferroni correction. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. Control genes are identical in plots A and B. TFBS motif data was obtained from DAP-seq¹⁸¹.

When analyses were repeated using the top 300 CRMs in each category, most findings supported those found with the top 100 CRMs (figs. S11 to S14). However, at the whole CRM level, there was a significant difference in TFBS coverage between non-specific, tissue-specific and control CRMs ($N = 300$; Kruskal-Wallis $H = 7.7$; $P < 0.05$) (fig. S10). Dunn's post hoc tests with Bonferroni correction revealed that tissue-specific CRMs had a significantly higher TFBS coverage ($19.0\% \pm 8.8$) than non-specific CRMs ($17.6\% \pm 9.5$; $P < 0.05$).

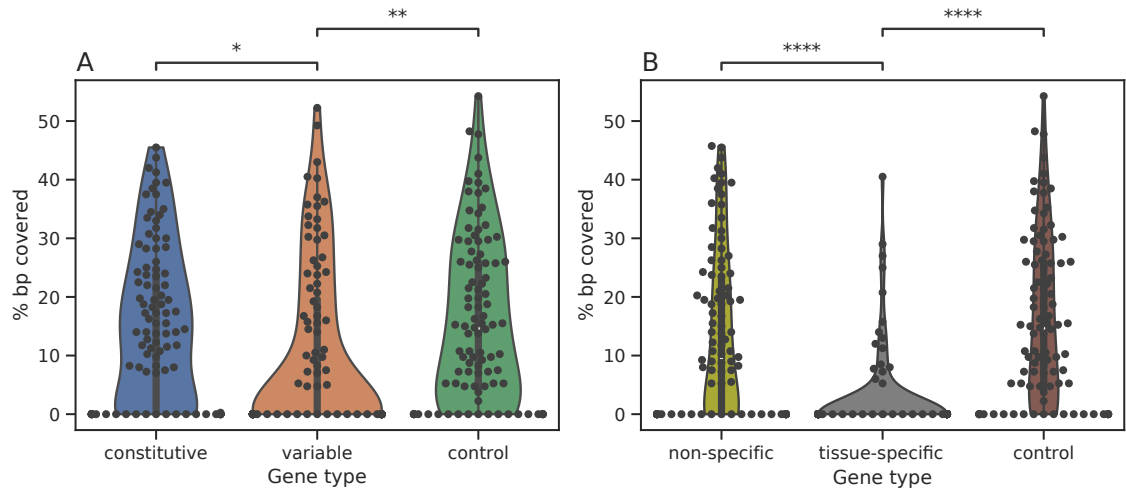


Figure 5.14: **Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions.** A: Mean percentage coverage of TFBS falling within open chromatin in 400 bp constitutive, variable and control CRMs. $N=100$. B: Mean percentage coverage of TFBS falling within open chromatin in 400 bp non-specific, tissue-specific and control CRMs. $N=100$. Significance was calculated using Kruskal-Wallis and Dunn's post hoc tests with Bonferroni correction. Control genes are identical in plots A and B. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. TFBS motif data was obtained from DAP-seq¹⁸¹.

5.4.6 TF diversity

The top 100 genes of each category were analysed to determine the diversity of TFs likely to bind, and to investigate if variable genes have a higher diversity of TFs than constitutive genes (see section 2.7.8). There was no significant difference in TF (Kruskal-Wallis $H = 0.9$, $P > 0.05$) or TF family Shannon diversity (Kruskal-Wallis $H = 0.5$, $P > 0.05$) between constitutive, variable or control types (fig. 5.15A and C).

There was also no significant difference in TF (Kruskal-Wallis $H = 2.3$, $P > 0.05$) or TF family diversity (Kruskal-Wallis $H = 2.3$, $P > 0.05$) between non-specific, tissue-specific or control CRMs (fig. 5.15B and D).

There was no significant difference in TF or TF family Shannon diversity using TFBSs falling within open chromatin between constitutive, variable and control CRMs (TF diversity: $N = 54$, Kruskal-Wallis $H = 3.0$, $P > 0.05$; TF family diversity: $N = 54$, Kruskal-Wallis $H = 0.9$, $P > 0.05$) or non-specific, tissue-specific or control CRMs (TF diversity: $N = 37$, Kruskal-Wallis $H = 4.0$, $P > 0.05$; TF family diversity: $N = 37$, Kruskal-Wallis $H = 4.8$, $P > 0.05$).

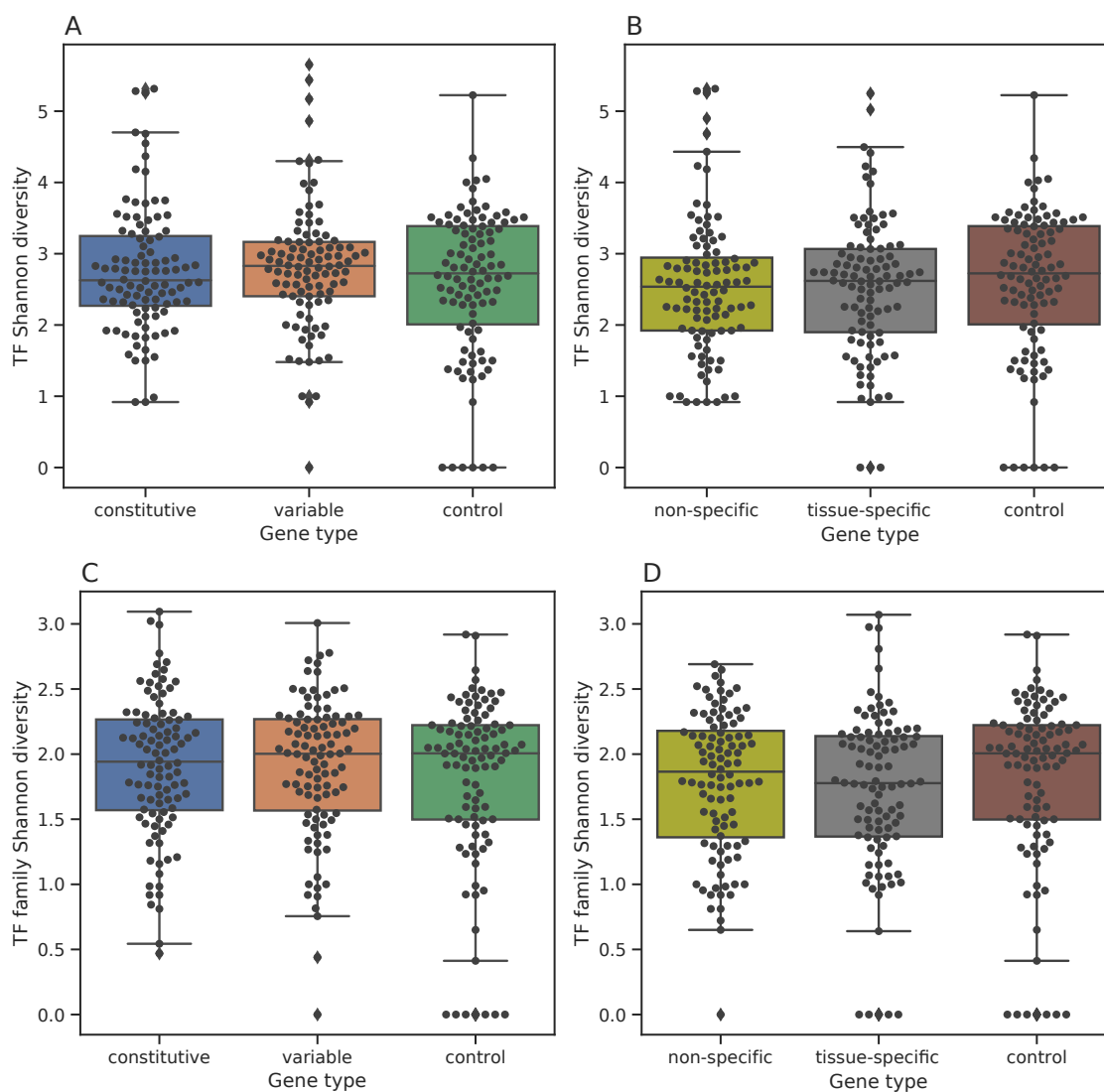


Figure 5.15: **Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Shannon diversity of individual TFs in A: 100 constitutive, 100 variable and 100 control genes and B: 100 non-specific, 100 tissue-specific and 100 control genes. Shannon diversity of TF families in C: 100 constitutive, 100 variable and 100 control genes and in D: 100 non-specific, 100 tissue-specific and 100 control genes. Control genes are identical in all plots.

A sliding window analysis (see section 2.7.4) revealed that in the constitutive open chromatin region 50–150 bp upstream of the ATG start codon variable promoters had slightly higher Shannon diversity than constitutive promoters (fig. 5.16A) and tissue-specific promoters had a higher Shannon diversity than non-specific promoters (fig. 5.16B). Median TF family Shannon diversity did not differ from 0 at any point along the CRM for any gene category.

TF Shannon diversity analysis was repeated restricting analysis to the 400 bp region upstream of the TSSs. Again, no significant difference in TF or TF family diversity was found between any CRM types whether using all TFBSs or only including those falling within open chromatin (All Kruskal-Wallis tests, $P > 0.05$).

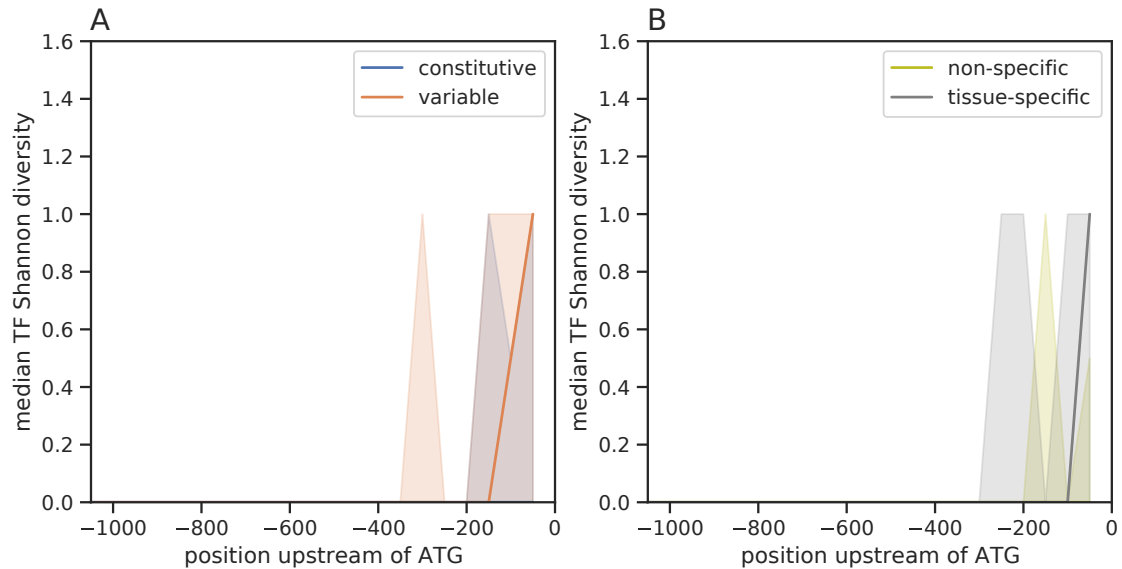


Figure 5.16: **Sliding window analysis of TF Shannon diversity of 100 constitutive (blue) and 100 variable (orange) Arabidopsis CRMs.** Data points are positioned in the centre of each 100 bp window. Windows are offset by 50 bp. The median value for each bin is displayed at the central bp of the bin. Shading represents 95 % confidence intervals estimated using 10000 bootstraps. A: Median Shannon diversity of individual TFs sliding windows in constitutive and variable CRMs. $N=100$. B: Median Shannon diversity of individual TFs sliding windows in non-specific and tissue-specific CRMs. $N=100$.

A principal component analysis was run using the Shannon diversity of TF families binding promoters in each category to look for an association between specific TF families and promoter category. To test the hypothesis that constitutive genes are bound by a different set of TF families than variable genes, Kmeans clustering was conducted. Kmeans clustering of the diversity of TF families binding to 400 bp CRMs did not correspond to gene type (fig. 5.17).

To test the hypothesis that non-specific genes are bound by a different set of TF families than tissue-specific genes, Kmeans clustering was conducted. Kmeans clustering of the diversity of TF families binding 400 bp CRMs did not correspond to gene type (fig. 5.18).

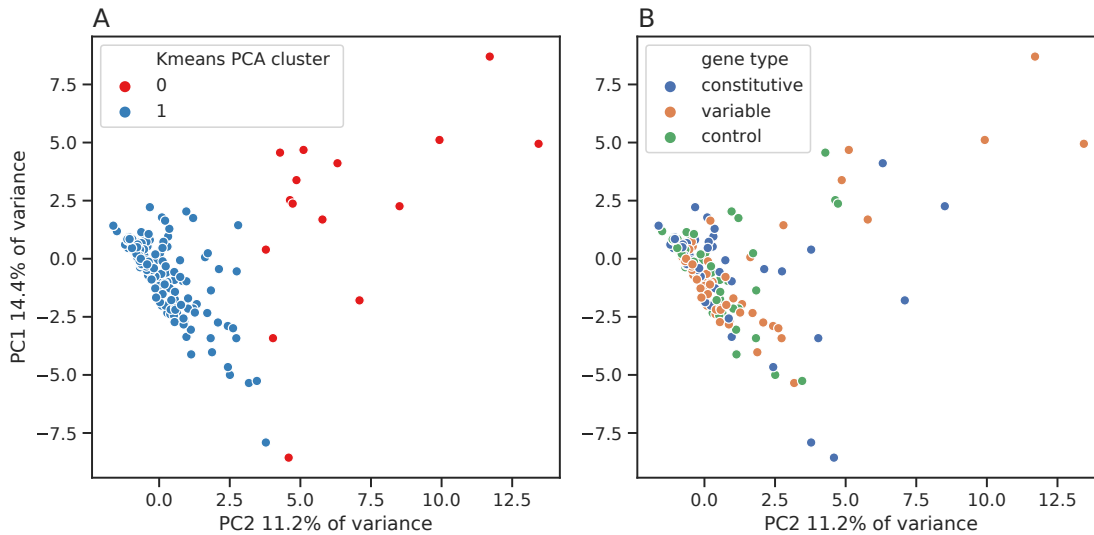


Figure 5.17: **Two PCA components accounting for the highest percent of variation for Shannon diversity of TF families binding promoters of constitutive and variable genes.** A principal component analysis was run where 95% of the variation was maintained with 22 components. Hierarchical clustering was used to estimate the number of clusters, K , using the silhouette method³⁷⁵ which was then used as K in Kmeans clustering. A: points coloured with two Kmeans clusters. B: points coloured by gene type. Constitutive genes, blue. Variable genes, orange. Control genes, green.

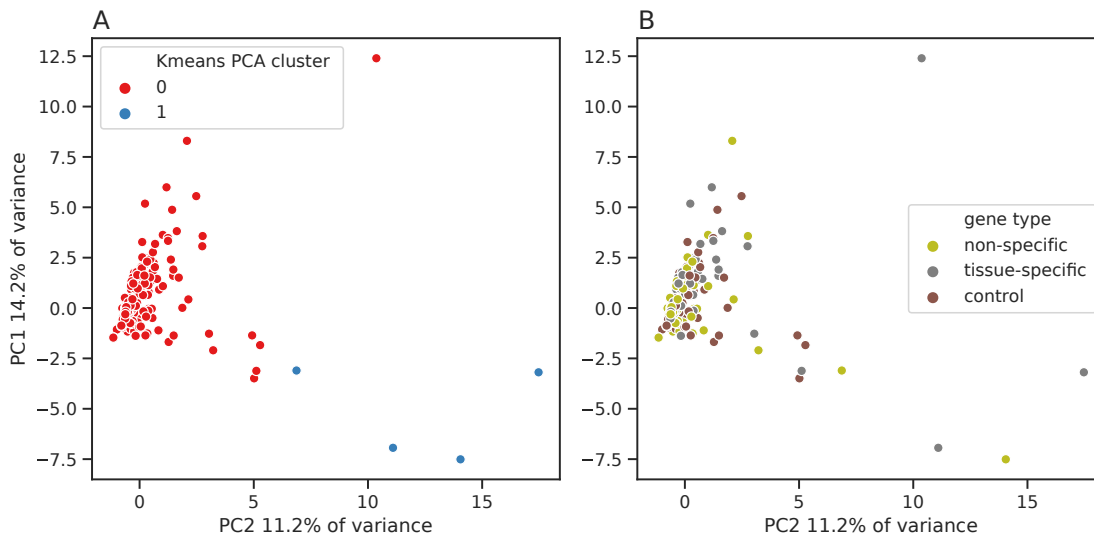


Figure 5.18: **Two PCA components accounting for the highest percent of variation for Shannon diversity of TF families binding promoters of non-specific and tissue-specific genes.** A principal component analysis was run where 95% of the variation was maintained with 22 components. Hierarchical clustering was used to estimate the number of clusters, K , using the silhouette method³⁷⁵ which was then used as K in Kmeans clustering. A: points coloured with two Kmeans clusters. B: points coloured by gene type. Non-specific genes, green. Tissue-specific genes, grey. Control genes, brown.

The number of genes included in the analyses of TF and TF family diversity using only TFBSs falling within open chromatin was low (ranging from $N = 20$ to $N = 54$) because genes with no TFBSs in open chromatin were filtered out. To increase sample size, all analyses were repeated using the top 300 CRMs in each category (figs. S15, S17 and S19). The findings mainly supported the results found with the top 100 CRMs in each category except for the following. Using only TFBSs falling within open chromatin in the whole CRM there was a significant difference in TF diversity between non-specific, tissue-specific and control CRMs ($N = 112$, Kruskal-Wallis $H = 10.2$; $P < 0.01$) (fig. S16B). Dunn's post hoc tests with Bonferroni correction revealed that tissue-specific CRMs (2.1 ± 1.2) had a significantly lower TF diversity than non-specific (2.5 ± 1.0 ; $P < 0.05$) and control CRMs (2.6 ± 0.9 ; $P < 0.01$). There was also a significant difference in TF family diversity in the 400 bp CRM between non-specific, tissue-specific and control CRMs ($N = 300$; Kruskal-Wallis $H = 8.1$, $P < 0.05$) (fig. S18D). Dunn's post hoc tests with Bonferroni correction revealed that non-specific CRMs (1.3 ± 0.8) had a significantly lower TF diversity than control CRMs (1.5 ± 0.7 ; $P < 0.05$) and tissue-specific CRMs (1.5 ± 0.7 ; $P < 0.05$).

5.4.7 TATA box enrichment

Enrichment of 15 bp TATA boxes was compared between constitutive and variable genes to test the hypothesis that variable genes are enriched in TATA boxes (see section 2.7.9). Variable promoters were found to be enriched in TATA boxes (total TATA boxes = 53; observed bp=840; expected bp=643; $\log_2\text{fold}=0.39$; $P < 0.01$) compared to the background of all 200 constitutive and variable promoters (fig. 5.19A). Conversely, constitutive promoters had significantly fewer TATA boxes compared to the background of all 200 constitutive and variable promoters (total TATA boxes = 29; observed bp=457; expected bp=656; $\log_2\text{fold}=-0.52$; $P < 0.01$). Enrichment of 15 bp TATA boxes was also compared between non-specific and tissue-specific genes to test the hypothesis that tissue-specific genes are enriched in TATA boxes over non-specific genes.

Tissue-specific promoters were not significantly enriched in TATA boxes (total TATA boxes = 33; observed bp=560; expected bp=484; $\log_2\text{fold}=0.21$; $P > 0.05$) compared to the background of all 200 non-specific and tissue-specific promoters. (fig. 5.19B). Non-specific promoters were also not significantly enriched in TATA boxes (total TATA boxes = 27; observed bp=432; expected bp=508; $\log_2\text{fold}=-0.23$; $P > 0.05$) compared to the

background of all 200 non-specific and tissue-specific promoters.

Analyses were repeated for the top 300 CRMs in each category to increase the sample size. As for the top 100 CRMs, variable promoters were enriched in TATA boxes compared to background CRMs (total TATA boxes = 177; observed bp=2956; expected bp=2376; $\log_2\text{fold}=0.31$; $P < 0.01$) and constitutive CRMs had significantly fewer TATA boxes compared to background CRMs (total TATA boxes = 94; observed bp=1492; expected bp=2082; $\log_2\text{fold}=-0.48$; $P < 0.01$) (fig. S20A). Interestingly, tissue-specific promoters were significantly enriched in TATA boxes compared to the background of all 600 non-specific and tissue-specific promoters (total TATA boxes = 144; observed bp=2417; expected bp=1833; $\log_2\text{fold}=0.40$; $P < 0.01$) (fig. S20B). Non-specific promoters had significantly fewer TATA boxes compared to the background of all 600 promoters (total TATA boxes = 81; observed bp=1303; expected bp=1885; $\log_2\text{fold}=-0.53$; $P < 0.01$).

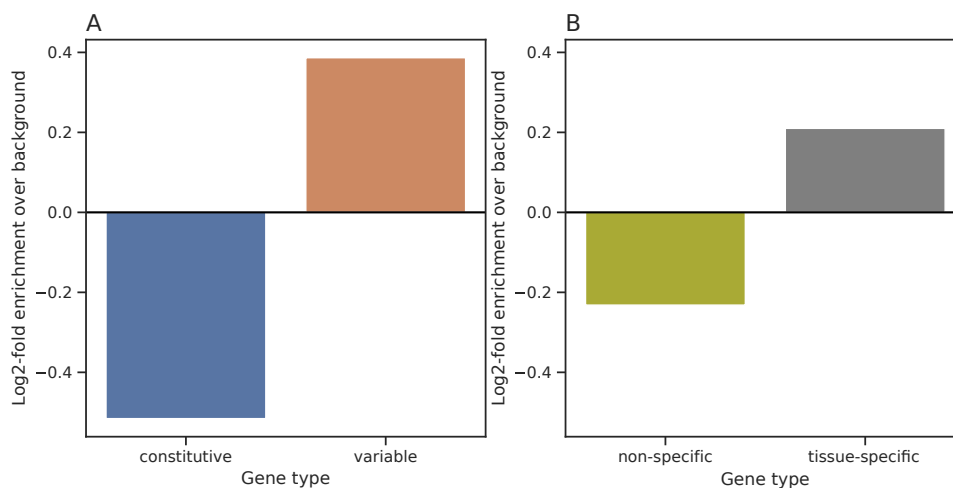


Figure 5.19: **\log_2 -fold enrichment of 15 bp TATA boxes in Arabidopsis CRMs.** A: \log_2 -fold enrichment of 15 bp TATA boxes in 100 variable (blue) and 100 constitutive (orange) Arabidopsis CRMs compared to a background of constitutive and variable promoters combined. B: \log_2 -fold enrichment of 15 bp TATA boxes in 100 non-specific (green) and 100 tissue-specific (grey) Arabidopsis CRMs compared to a background of non-specific and tissue-specific promoters combined. TATA box locations within 50 bp upstream of the EPD TSS were downloaded from EPD⁴⁹⁸. Gat software³⁷⁷ was used to calculate enrichment.

5.5 Discussion

The genes in the four gene categories (constitutive, variable, tissue-specific and non-specific) had a range of expression patterns, with the constitutive and non-specific categories being stably expressed across different conditions/tissues, and the variable and tissue-specific genes being expressed only in certain conditions or tissues. Gene ontology enrichment analysis showed that the constitutive genes were associated with intracellular protein transport and endocytosis. Non-specific genes were associated with the mitochondrion, cytosol, and Golgi apparatus. These functions support the expectation that broadly expressed genes are likely to be involved in cellular processes required in all cells. The tissue-specific genes were associated with the pollen tube, extracellular region, chloroplast, protein binding and pectinesterase activity. These functions also fall in line with the literature due to their association with specific tissues such as pollen tubes, and the association with chloroplasts which are only found in certain cell types. Pectinesterases are plant cell wall-associated enzymes that are involved in the degradation of pectin, a major component of the plant cell wall⁴⁹⁹. The tissue-specific genes were also associated with protein binding so they could be involved in signalling, and there was an association with pollen exine development (GO:0010584). Das and Bansal⁴⁹⁴ found that tissue-specific or narrowly expressed genes were associated with the cell wall, signalling and developmental processes, in agreement with the associations of the tissue-specific genes in this chapter. Also in support, Einarsson et al.⁴⁸³ found that in humans, constitutive genes were associated with housekeeping processes such as metabolism, while variably expressed genes were associated with dynamics functions such as signalling, stimulus response and developmental processes. A limitation of the analyses in this chapter is that as the tissue-specific genes were highly expressed in a few pollen related tissues along with chloroplasts, it is difficult to generalise the results to other tissue-specific genes. What is true in pollen may not be true in other tissues. In the future, a similar analysis should be performed on other tissue-specific genes expressed in a range of tissues to see if the results are consistent across different tissues.

5.5.1 The CRMs of constitutive and non-specific genes contain more open chromatin than variable and tissue-specific genes

Percentage coverage of open chromatin was higher in constitutive genes than variable genes, and higher in non-specific than tissue-specific genes. This is consistent with a previous study in Arabidopsis, where Cortijo et al.⁴⁹⁶ found that genes with stochastic expression contained a more compact chromatin state than constitutively expressed genes. Another study in Arabidopsis showed similar results, with constitutively expressed genes having increased DNase I sensitivity than tissue-specific genes⁴⁹⁴. This is also consistent with a study using mammalian hESC cell lines where local chromatin density was higher in variably expressed genes (with high CV) than constitutive genes (with low CV)⁵⁰⁰. However, the authors also found that chromatin accessibility was not related to gene expression variation. They suggest that this is because chromatin remodelling proteins locally condense chromatin in variably expressed genes⁵⁰⁰. Percentage open chromatin coverage increased around the TSS, which was expected, as TSSs have been shown to generally be chromatin accessible across eukaryotes^{501–503}. Since only a limited number of tissues were present in the tissue-specific gene set, the analysis should be repeated in the future with a wider range of tissue-specific genes to see if the results are consistent across different tissues.

5.5.2 Constitutive and non-specific genes contain increased GC content in 400 bp upstream region than variable and tissue-specific genes

In the 400 bp region upstream of the ATG start codon, the percentage GC content was higher in constitutive genes than variable genes, and higher in non-specific than tissue-specific genes. This is in agreement with previous work in which mammalian constitutive promoters were shown to have a higher GC content than tissue-specific promoters^{485,486}. In 1000 bp upstream regions, the percentage GC content was higher in constitutive genes than variable genes but was not significantly different between non-specific and tissue-specific genes. This suggests that the 400 bp region is important for explaining differences in expression variability between the gene categories. In plants, genes with GC-rich promoters have a lower expression than AT-rich promoters²⁷⁵. This is consistent with the finding that constitutive genes have more GC-rich promoters than variable genes, and that

constitutive genes are known to have a weaker expression in general than variable genes in Arabidopsis^{157,372}.

The GC content was found to gradually increase from upstream regions towards the ATG start codon. This is consistent with studies in animals which showed a peak in GC content around the TSS, and a drop off in GC content throughout the gene with a drop in GC content at the transcription stop site⁵⁰⁴.

5.5.3 Variable and tissue-specific genes contain more TFBSs in the 400 bp region upstream of the ATG start codon than constitutive and non-specific genes

The percentage of bp covered by at least one TFBS was higher in variably expressed than constitutively expressed genes, and higher in tissue-specific than non-specific genes. This is consistent with a previous study in Arabidopsis, where stochastically expressed genes were found to be targeted by a larger number of TFs than constitutive genes⁴⁹⁶. However, in this chapter, there was no significant difference in the percentage of nucleotides covered by at least one TFBS in 1000 bp upstream regions between the gene categories. When the regions compared were narrowed to those falling within open chromatin, the opposite was found, with a higher percentage of nucleotides covered by TFBSs in constitutive genes than variable genes, and in non-specific genes than tissue-specific genes. This could be because constitutively expressed genes contain more open chromatin than variable genes, and chromatin remodelling proteins may locally condense chromatin in variably expressed genes⁵⁰⁰, obscuring TFBSs and making variably expressed genes more reliant on pioneer transcription factors to displace chromatin to enable transcription.

Median percentage TFBS coverage increased from upstream regions towards the ATG start codon. This is in agreement with a previous study in human tissues in which the distribution of TFBSs peaked around the TSS⁵⁰⁵.

5.5.4 TF diversity did not differ between gene categories

The Shannon diversity of TFs and TF families binding to a gene did not differ between the gene categories. However, when comparing regions falling within open chromatin, the top 300 tissue-specific genes had a lower TF Shannon diversity than non-specific and control

genes. In contrast, TF family Shannon diversity was higher in the top 300 tissue-specific genes than the top 300 non-specific genes in the 400 bp upstream regions. This was supported by a previous study in humans in which variably expressed promoters were found to bind a higher diversity of TFs than constitutively expressed promoters⁴⁸³.

Kmeans clustering of the diversity of TF families binding to 400 bp upstream regions did not correspond to gene type, suggesting that constitutive and non-specific genes were not bound by a different set of TF families to variable or tissue-specific genes. This was supported by a previous study in *Arabidopsis* which found that 500 bp promoters had no significant GO term enrichment or gene family associations with constitutive or variable genes categories⁴⁹³. However, these findings were contradicted in another study in *Arabidopsis*, where certain TF families were enriched in open chromatin regions of specific cell types, for example, WRKY family motifs were enriched in epidermal and cortex cell chromatin accessible sites⁵⁰⁶. This could be due to the increased resolution of comparing regulatory landscapes of genes in single cells (integration of single-cell ATAC-seq and single-cell RNA-seq) as opposed to comparisons between whole tissues, with potentially increased noise in heterogeneous tissue data masking any enriched TF families. In humans, TFs binding to highly variable promoters were mostly associated with tissue-specific or developmental regulation, while TFs binding to constitutive promoters were mostly associated with ubiquitous activity across cell types⁴⁸³.

5.5.5 TATA boxes were enriched in variable and tissue-specific promoters

15 bp TATA boxes were enriched in variable promoters compared to the background of combined constitutive and variable promoters. Additionally, TATA boxes were enriched in tissue-specific promoters compared to the background of all 600 non-specific and tissue-specific promoters). This is consistent with a previous study in *Arabidopsis*, in which TATA boxes were found to be enriched in variable genes compared to constitutive genes⁴⁹⁴. Similarly, previous research in animals showed that TATA boxes are enriched in variable promoters^{507,508}. Genes containing TATA boxes are associated with stress responses while TATA-less promoters are associated with constitutive expression^{508,509}. Promoters containing TATA boxes were shown to be up to 4x stronger than TATA-less promoters²⁷⁵, which is consistent with the higher expression of variable genes compared to constitutive genes^{157,372}.

5.5.6 Analysis of promoter architecture is useful for learning design features for designing synthetic promoters

Various design features can be extracted from the results in this chapter which can be used to design synthetic nitrate (N)-responsive promoters for use in genetic feedback controllers in chapter 6. For example, TATA boxes were enriched in variably expressed genes, and are known to improve the strength of genes²⁷⁵. Therefore, TATA boxes should be included and tested in synthetic nitrate-responsive promoter designs. Additionally, N-responsive synthetic promoters should contain a low GC content, since the GC content of variable promoters was significantly lower than constitutive promoters and a lower GC content is linked with higher expression²⁷⁵. Variably expressed and tissue-specific genes had a higher percentage of nucleotides covered by TFBSs than constitutive/non-specific genes, suggesting that synthetic N-responsive promoters could be compact and include several TFBSs and not too much spacer sequence. Indeed, endogenous variably expressed genes were found to be shorter than constitutively expressed genes⁴⁹⁴.

Four categories of genes (constitutive, variable, non-specific, tissue-specific) were selected using CV and Tau distributions. Microarray expression data was used to rank genes according to their CV. A limitation of microarrays is that they do not discriminate between alternative TSSs. It would be useful to repeat the analyses in this chapter using RNA-seq data, which can discriminate between alternative TSSs. Although splitting genes into categories is useful for extracting architectural features from promoter regions, it is important to note that the expression patterns of genes are on a spectrum, and even constitutive genes show significant cell to cell variability in Arabidopsis⁴⁷⁶. To account for this, control CV and Tau gene categories were used which included genes from across the spectrum of expression patterns. As expected, the distribution of features such as GC content and TFBS coverage of these control categories tended to lie centrally in between the other gene categories (constitutive and variable, non-specific and tissue-specific). The analyses in this chapter should be repeated in the future with a wider range of tissue-specific genes to see if the results are consistent across different tissues. Since only a limited number of tissues were present in the tissue-specific gene category, it is possible that the results are biased towards the specific tissues present in the dataset. As the selection of tissue-specific genes was based on the Tau distribution, it is possible that the results are

biased towards genes with high expression in a single tissue. It might be useful to analyse genes that are tissue-specific with lower expression too, to see if the results are consistent across different expression levels.

When extracting promoters from the *A. thaliana* genome, 1000 bp upstream of the annotated TSS was used, along with the 5' UTRs. This was to ensure that the whole TSR was included, as genes can have more than one TSS. Since some promoters were very short due to upstream genes, there could be potential biases in the data. For example, the chromatin state of the upstream gene could affect the expression of the downstream gene. Additionally, enhancers were not included in the analysis, which could affect the expression of the gene. Therefore, only some of the gene expression variation will have been captured in the promoter sequences analysed in this chapter. In the future, it would be interesting to extend the analysis to include enhancers and the chromatin state of surrounding genes, to determine how they affect gene expression patterns. Overall, the differences in promoter architecture between constitutive and variable gene categories in this chapter are more reliable for designing synthetic promoters with different expression patterns than the differences between non-specific and tissue-specific gene categories, since the tissue-specific gene category only represented a small number of tissues.

The analyses in this chapter demonstrated that TFBS density was higher and TATA boxes were enriched in variably expressed genes compared to constitutive genes, and GC content and chromatin percentage coverage was higher in constitutively expressed genes than variably expressed genes. In general, these results were supported by previous studies in *Arabidopsis* and other species, which is encouraging for the use of the results in designing synthetic promoters.

5.5.7 Conclusions

In summary, this chapter has shown that the architecture of promoters/5' UTRs is different between constitutive and variably expressed genes, and between non-specific and tissue-specific genes. Useful design features could be extracted from the results, which are used in chapter 6 to design synthetic minimal N-responsive promoters for use in genetic feedback controllers. The analysis pipelines used in this chapter could be adapted for use in other organisms in the future, to compare whether similar promoter features are

found in other species, and to allow the design of synthetic promoters capable of similar expression patterns in multiple species. A wider range of tissue-specific genes should be used in the future to see if the results are consistent across different tissues.

Chapter 6

Engineering of a genetic feedback loop

6.1 Preface

This chapter describes the design and construction of synthetic genetic elements that introduce positive and negative synthetic genetic feedback into the nitrogen subnetwork. I designed all synthetic promoters and assembled all plasmids for transient testing of synthetic promoters and designed/assembled all initial CRISPR synthetic TFs constructs. Dr. Tufan Oz assembled the final CRISPR synthetic TFs constructs, synthetic promoter LucN::YFP reporters, and the final feedback controllers. I performed all transient protoplast co-expression luciferase assays assessing performance of synthetic promoters and synthetic TFs, and all root LucN luciferase assays assessing nitrate-responsiveness of synthetic promoters. Dr. Tufan Oz transformed all Arabidopsis lines for generation of transgenic seeds.

6.2 Introduction

Plants regulate their metabolism and growth in response to changes in environmental conditions. Complex GRNs composed of suites of genes which interact with each other are responsible for mediating these changes³. GRNs are comprised of nodes (genes) and edges which represent regulatory relationships between nodes. Network motifs such as

feedback loops are common features of GRNs⁷. Positive feedback, where a downstream node Z activates an upstream node X which regulates Z directly or indirectly often leads to bistability, with genes involved in either the on or off state¹⁰ (see fig. 1.1). Negative feedback, where a downstream node Z represses an upstream node X which regulates Z, can lead to oscillations⁹, and improve the robustness of the network⁷. It can therefore be hypothesised that the addition of feedback into an existing GRN might confer new phenotypes. Synthetic biology approaches provide the opportunity to engineer GRNs.

Synthetic biology aims to predictably engineer biological systems either to confer new functions, such as the ability to produce new molecules, or to tune existing functions. These aims are typically achieved by engineering existing genetic elements or by the introduction of synthetic genetic circuits. While synthetic feedback might be engineered into a network by engineering an existing promoter region to include new binding sites for a given TF, the introduction of new sequence data into an exact genomic location remains technically challenging in plants. Further, different TFs would most likely be required to activate and repress the target promoter, making it difficult to compare positive and negative feedback from the same network node. An alternative approach is to insert synthetic genetic circuitry to create the new network edges. Synthetic genetic circuits generally contain synthetic regulators that either control the expression of other genes in the circuit, or that modify the expression of host genes^{253,510,511}. One of the reasons that it remains challenging to design synthetic genetic circuits that behave predictably is because it is unknown how and if host elements such as metabolites, TFs and regulatory RNAs will interact with and affect the behaviour of synthetic genetic elements⁵¹². To minimise such interactions, it is desirable to use orthogonal parts, which have minimal crosstalk with host regulatory systems. These include synthetic TFs that only activate synthetic promoters and are reasonably unaffected by host molecules or sequences. Several orthogonal regulatory systems have been demonstrated in plants. These include the glucocorticoid-inducible chimeric transcriptional activator GVG, containing the yeast Gal4 DNA-binding domain, the VP16 activation domain and the GR domain, and an inducible promoter containing 4×UAS TFBSs⁵¹³. Another example is the estradiol-inducible activator, XVE (LexA DNA-binding domain, VP16 activation domain and human oestrogen receptor) and a target synthetic promoter containing 8×LexA TFBSs⁵¹⁴. Recently, a copper-inducible synthetic promoter containing four motifs called copper-binding sites, paired with the yeast

copper responsive factor CUP2 fused to the Gal4 DNA-binding domain was demonstrated in *Nicotiana benthamiana*²⁹¹. Programmable TFs systems such as TALE and synthetic TALE-activated promoter (STAP)²⁸⁴, and Cas9-based systems with the *AtuNOS* promoter target⁵¹⁵ have also been demonstrated. A particular advantage of programmable synthetic TFs such as TALEs and Cas9 is that they can also be used to control the expression of host genes (see section 1.6.4). For example, dCasEV2.1-mediated transcriptional activation of genes in the flavonoid pathway in *N. benthamiana* enabled the re-routing of metabolic fluxes towards the accumulation of metabolites of interest⁵¹⁶. This feature provides opportunities for engineering the behaviour of GRNs.

To enable synthetic TFs to activate their downstream synthetic promoter, their expression is often controlled using either a natural or synthetic promoter. This is often a constitutive promoter, although promoters known to express in the desired cells or tissues of interest can also be used^{259,297}. The use of minimal synthetic promoters that respond to a limited range of known host elements (*e.g.*, TFs^{278,281,285,286}), might allow greater predictability over the spatial and/or temporal conditions under which expression of the synthetic TF is activated.

In this chapter, the design and construction of synthetic genetic elements that introduce positive and negative synthetic genetic feedback into the nitrogen subnetwork is described. To design these synthetic genetic feedback controllers, first information about the structure of the subnetwork (chapter 3) was used to identify a target TF, ARF18, to activate/repress. Next, synthetic TFs that bind to the upstream regulatory region of *ARF18* and either activate or repress its transcription were built and tested. To enable these elements to function in response to TFs lower in the hierarchy, thus creating feedback, a suite of synthetic minimal N-responsive promoters was designed with binding sites for NLP7. Additional synthetic promoters with binding sites for other early responders to nitrate³¹⁸ were also built. A schematic depicting the feedback loop via a synthetic promoter and synthetic TF to *ARF18* is shown in fig. 6.1.

This chapter describes the design, building and testing of all synthetic elements present in the feedback controllers. At the time of writing, the feedback controllers have been assembled and transgenic plant lines containing these controllers have been produced. However, analysis of these lines is outstanding.

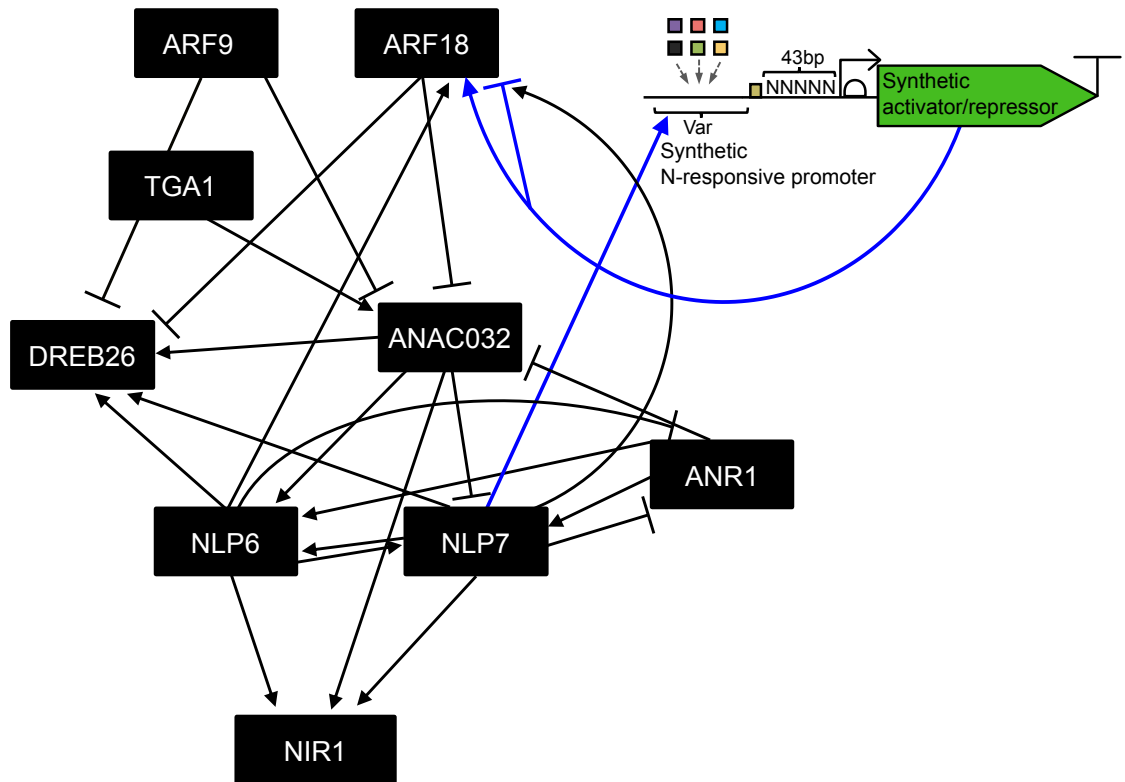


Figure 6.1: **Schematic of new edges added by synthetic feedback controller.** A synthetic minimal N-responsive promoter responds to NLP7 (or other early responders to nitrate) and activates a synthetic CRISPR TF that binds to the upstream regulatory region of *ARF18* and activates or represses its transcription. This adds a genetic feedback loop to *ARF18*.

6.3 Aims

The aims of this chapter are to design, build and quantitatively test (i) synthetic transcriptional activators and repressors that alter the expression of *ARF18* and (ii) minimal synthetic promoters that respond to known TFs and, ideally, are activated by nitrate. A final aim is to assemble these together to create a synthetic genetic feedback controller.

6.4 Results

6.4.1 Design and construction of programmable synthetic TFs

Synthetic transcriptional activators and repressors were designed to target the promoter of *ARF18*. This network node was selected because it was shown to be capable of binding to and altering the expression of itself, *ANAC032* and *DREB26* in the N-response subnetwork (see chapter 3 and fig. 3.24). Together with the network topology, data from existing

loss-of-function lines suggested that altering the expression of this node would affect the network²⁵.

The Cas9-SunTag system²⁹⁹ was used to construct both synthetic transcriptional activators and repressors (see section 2.1.3). This system was selected because at the time it was shown to have the largest fold-activation (130–4000) of target genes in Arabidopsis⁵¹⁷ (although the recently published CRISPR-Act3.0 system has demonstrated even stronger activation³⁰¹). In the Cas9-SunTag system, an endonuclease deficient Cas9 (dCas9)::NLSGCN4 fusion is co-expressed with a fusion of scFv::sfGFP with either the VP64 transcriptional activator domain or two repeats of the TAD³⁰¹ activator domain, or with three repeats of the SRDX repressor domain²⁹⁴. To target the protein complex to the *ARF18* promoter, three sgRNA guides were designed to target a region close to the TSS, since regions close to the TSS were found in systematic studies to lead to the highest activation of transcription^{300,518}. SpCas9 NGG PAM sites were identified using CRISPOR³⁵¹ with on target efficiency of 40% or more. At least 45 bp of spacing was allowed between protospacers to allow space for multiple guides and CRISPR protein complexes to bind to the promoter. Within this region, three NGG PAM sites at -72 bp, +9 bp or +75 bp relative to the TSS were selected (fig. 6.2).

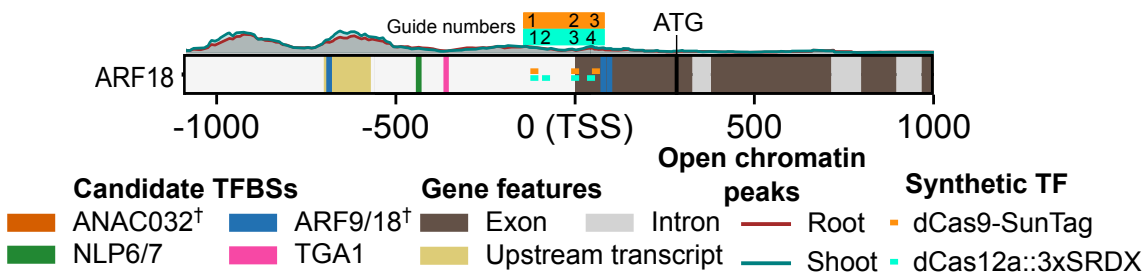


Figure 6.2: **Locations of protospacer targets in *ARF18* of synthetic TFs.** Protospacer locations in *ARF18* upstream region of dCas9 and dCas12a based synthetic programmable TFs. Numbers were assigned to each guide for easy reference. †, TFBS inferred from closely related TF, see section 3.5.1. ATAC-seq data³⁵⁰ (see section 2.7.7) were used to annotate open chromatin peaks.

Target sequences (spacers) were introduced by PCR amplification of a cloned sgRNA scaffold sequence³³¹ using primers with the spacer sequence in a 5' overhang. These were assembled with the *AtU6-26* promoter in a Golden Gate reaction (see section 2.1.3). The dCas9::NLSGCN4 and scFv::sfGFP(VP64/TAD) coding sequences were assembled with

AtUBQ10 promoters to produce transcriptional units with constitutive expression. An insulator was used to separate these transcriptional units (*Petunia hybrida* transformation booster sequence³³⁰). To produce the final constructs, these two transcriptional units were co-assembled with either one or three sgRNA cassettes to produce a multigene construct as well as two selectable marker genes, the fluorescence-accumulating seed technology (FAST-Red) gene^{325,328} and the glufosinate-ammonium tolerance gene (bar gene)³²⁵ (see section 2.1.3). A version of the constructs without any sgRNA expression cassettes was constructed as a negative control. A schematic showing the design of the constructs is shown in fig. 6.3.

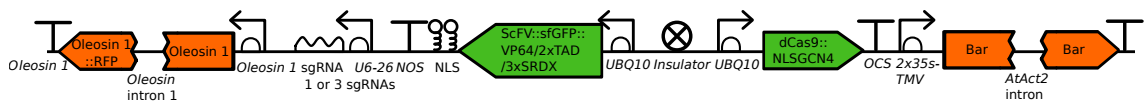


Figure 6.3: **Schematic of SunTag synthetic transcription factors.** This system requires co-expression of a dCas9::NLSGNC4 fusion and a fusion of scFv::sfGFP fused to either a VP64 domain, two repeats of the TAD activation domain³⁰¹, or three repeats of the SRDX repressor domain²⁹⁴. The protein complex is delivered to the target gene by forming a complex with sgRNAs. Constructs also contain the FAST-Red^{325,328} and bar phosphinothricin³²⁵ marker genes to allow the selection of plant lines in which the complete T-DNA has integrated into the genome.

A second system, an endonuclease deficient Cas12a (AsCpf1)²⁹⁴, was also used to construct synthetic transcriptional repressors (see section 2.1.3). This system was selected because it was the strongest published repressor system demonstrated in Arabidopsis at the time (90% repression level)⁵¹⁷. In this system, dCas12a::NLS::3×SRDX fusion protein is used together with a crRNA double hammerhead/hepatitis delta virus ribozyme cassette targeting either one or three protospacers in the promoter region close to the TSS. Within this region, four TTTN PAM sites at -98 bp, -43 bp, +15 bp or +36 bp relative to the TSS were selected (fig. 6.2). These sites were predicted using CRISPOR³⁵¹ to have on target efficiency of 40% or more. Both the dCas12a::NLS::3×SRDX fusion and the crRNA were assembled with *AtUBQ10* promoters. As above, to produce the final constructs, these two transcriptional units were between two selectable marker genes (fig. 6.4). A version of the constructs without any crRNA expression cassettes was constructed as a negative control.

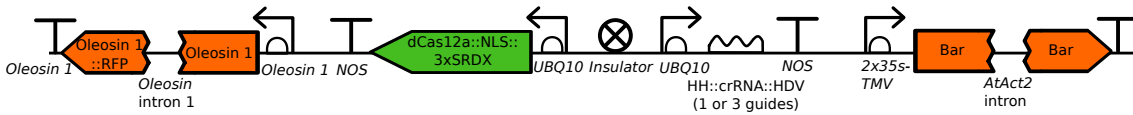


Figure 6.4: **Schematic of synthetic dCas12a transcriptional repressors.** The system requires expression of a dCas12a::SRDX fusion. The protein complex is delivered to the target gene by forming a complex with crRNA, which is encoded between hammerhead/hepatitis delta virus ribozymes. Constructs also contain the FAST-Red^{325,328} and bar phosphinothricin³²⁵ marker genes to allow the selection of plant lines in which the complete T-DNA has integrated into the genome.

6.4.2 Assessing the performance of synthetic TFs

All synthetic TFs were tested in protoplasts. To determine the regulatory effect of a given synthetic TF on the expression of the *ARF18* promoter, constructs encoding p_{ARF18} -LucN were used to transfect Arabidopsis mesophyll protoplasts in the presence and absence of constructs encoding synthetic TFs (see section 2.2.7). Changes in luminescence of the p_{ARF18} -LucN luciferase reporter were quantified relative to an experiment calibrator to assess the effect of each synthetic TF on the expression of *ARF18* (see fig. 2.4).

Co-expression of synthetic transcriptional activators with p_{ARF18} -LucN increased luminescence

Levels of luminescence obtained from p_{ARF18} -LucN were significantly increased following co-expression with transcriptional activators containing the 2×TAD activator domain with one (11.01 ± 0.07 ; $t = -93.9$, $P < 0.001$) and three guides (20.74 ± 0.34 ; $t = -62.9$, $P < 0.001$). Luminescence was also increased by co-expression of transcriptional activators containing the VP64 activator domain with both one (8.88 ± 0.08 ; $t = -44.6$, $P < 0.001$) and three guides (13.28 ± 0.12 ; $t = -83.1$, $P < 0.001$). Increases in expression were assessed by comparing results to the respective no guide controls (2×TAD: 5.40 ± 0.04 ; VP64: 6.27 ± 0.01) (fig. 6.5). Unexpectedly, the luminescence of p_{ARF18} -LucN was also significantly increased, although to a lesser extent, following expression of the no guide controls (2×TAD: 5.40 ± 0.04 ; $t = -27.2$, $P < 0.001$; VP64 6.27 ± 0.01 ; $t = -98.7$, $P < 0.001$), as compared to the no TF control (4.45 ± 0.02). The best activator was the 2×TAD activator with three guides, which increased luminescence by 380% over the no guide control.

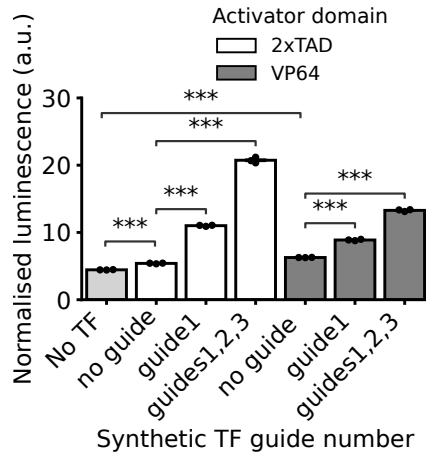


Figure 6.5: **Normalised luminescence from p_{ARF18} -LucN following co-expression with synthetic transcriptional activators.** p_{ARF18} -LucN was co-expressed in Arabidopsis mesophyll protoplasts with dCas9/SunTag²⁹⁹-based synthetic transcriptional activators containing either VP64 (dark grey bars) or TAD³⁰¹ (white bars) activation domains. Luminescence is normalised to a $p_{CaMV35s}$ -LucF experimental calibrator. Luminescence was measured 18 hours after protoplast transfection. Error bars represent two standard errors. $N = 3$ technical replicates. P -values were calculated using Welch's t -tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

Co-expression of synthetic transcriptional repressors with p_{ARF18} -LucN decreased luminescence relative to the no guide control

Levels of luminescence obtained from p_{ARF18} -LucN were significantly decreased following co-expression with transcriptional dCas12a::NLS::3×SRDX repressors containing guide 1 (8.26 ± 0.07 ; $t = 21.0$, $P < 0.001$), 2 (9.06 ± 0.04 ; $t = 12.4$, $P < 0.01$), 3 (8.31 ± 0.11 ; $t = 15.4$, $P < 0.001$), guides 1–3 (7.89 ± 0.08 ; $t = 24.3$, $P < 0.001$) and guides 1, 2 and 4 (9.03 ± 0.15 ; $t = 6.6$, $P < 0.01$) compared to the no guide control (9.83 ± 0.08) (fig. 6.6). There was no significant change in p_{ARF18} -LucN luminescence following co-expression of the SunTag repressor ($p_{AtUBQ10}$ -dCas9::NLSGCN4 fusion and a scFv::sfGFP fusion with 3×SRDX repressor domains) compared to the no guide control ($P > 0.05$). As observed for the synthetic transcriptional activators, co-expression of both the SunTag 3×SRDX repressor (9.83 ± 0.08 ; $t = -91.6$, $P < 0.001$) and the dCas12a::NLS::3×SRDX repressors (12.99 ± 0.25 ; $t = -48.5$, $P < 0.001$) even with no crRNAs significantly increased the luminescence of $ARF18$ -LucN compared to the no TF control (4.45 ± 0.02). The best synthetic repressor was the dCas12a::NLS::3×SRDX repressor containing guides 1–3 which decreased luminescence by 19.7% compared to the no guide control.

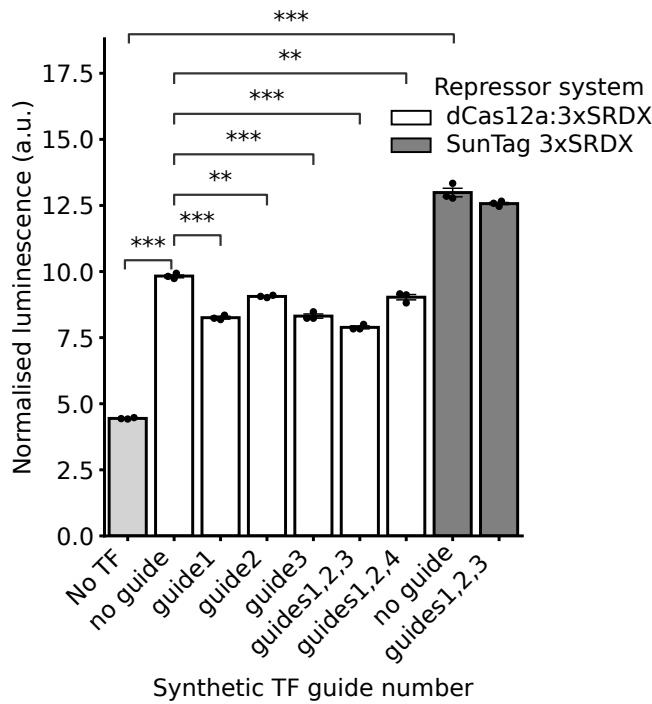


Figure 6.6: **Normalised luminescence from p_{ARF18} -LucN following co-expression with synthetic transcriptional repressors.** p_{ARF18} -LucN was co-expressed in Arabidopsis mesophyll protoplasts with either dCas9/SunTag-based (dark grey bars) or dCas12a-based (white bars) synthetic transcriptional repressors. Luminescence is normalised to a $p_{CaMV35s}$ -LucF experimental calibrator. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

6.4.3 Development of minimal synthetic N-responsive promoters

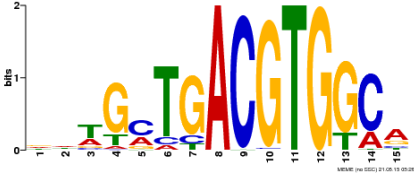
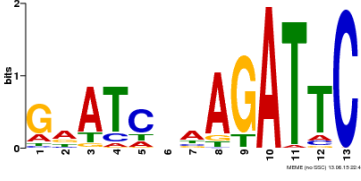
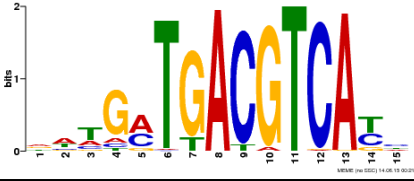
Since ARF18 is also known to be involved in the growth response to shade⁴⁰⁷, constitutive changes to expression might have unintended consequences. It is therefore desirable to limit or control the timing of activation of synthetic TFs. Bringing the expression of synthetic TFs under the control of a TF that is regulated by ARF18 would enable the introduction of synthetic feedback. To enable this, I first designed and built several synthetic promoters. These were tested in protoplasts using a dual luciferase ratiometric assay. To test if these promoters responded to nitrate, they were fused to a reporter and transformed into plants. Stable lines were grown in different nitrate conditions. These results are described below.

Selection of TFBSs

NLP7 (AT4G24020) was a primary candidate TF for controlling the synthetic promoter because it was found to be indirectly regulated by ARF18 through ANAC032 (see chapter 3

and fig. 3.24). Furthermore, NLP7 is an early responder to nitrate and was recently shown to directly sense nitrate⁴⁹. The TFs bZIP3 (AT5G15830), HHO2 (AT1G68670) and TGA1 (AT5G65210) were also selected as they were identified as early responders to nitrate in a previous study³¹⁸. HHO2 is a known repressor of genes such as *NRT2.1*⁵¹⁹. To enable expression in response to the selected TFs, binding sites needed to be identified and encoded into the synthetic promoter. For promoters that respond to NLP7, the NRE was used from NIR1 (AT2G15620), which was previously found to bind NLP7³⁹³. A shortened version of the NRE was also used, containing the core 24 bp which contained the essential NLP7 motifs tested in Konishi and Yanagisawa³⁹⁶ (table 6.1), a sequence which will now be referred to as an NLP7 TFBS. *A. thaliana* TFBS consensus sequences for bZIP3, HHO2 and TGA1 were identified from the plant cistrome database¹⁸¹ (http://neomorph.salk.edu/dev/pages/shhuang/dap_web/pages/browse_table_aj.php) (table 6.1).

Table 6.1: **DAP-seq consensus sequence logos or published sequences used for adding TFBSs in synthetic promoters.** The central region of the NRE motif³⁹³ used in the shortened version is coloured in blue.

TF	Sequence logo
bZIP3	
HHO2	
NRE (contains NLP7)	AGAAACA ACT TGACCCTTTACATTGCTCAAGA
NLP7 (shortened NRE)	GCTCATCTCTT TGACCCTTTACATTGCTCAAGAGC
TGA1	

Design of minimal synthetic N-responsive promoters

The sequences of TFBSs for TFs known to respond to nitrate were added into the variable upstream region of a previously published minimal synthetic promoter²⁷⁸ (fig. 6.7) and the Ω TMV 5' UTR³²⁵. Promoters containing either the complete NRE from *AtNIR1* or the NLP7 TFBS from this element were designed. Five different versions were constructed, each containing variations on the length and identity of sequence between the TFBSs see(fig. 6.8A–E). In addition, previous studies in yeast have shown that placing TATA boxes in between TFBSs can increase induction of transcription and the activation ratio²⁷¹. Further, TATA boxes were enriched in variably expressed Arabidopsis genes (see section 5.4.7), and plant promoters containing TATA boxes were found to be up to 4× stronger than TATA-less promoters²⁷⁵. To test the hypothesis that adding TATA boxes in between TFBSs would increase the activation ratio and induction of promoters, 7 bp TATA boxes (TATATAAA) were added 1 bp downstream of each NLP7 TFBS (fig. 6.8B and C). These were compared to the insertion of 7 bp of random sequence with equal ATCG ratios (fig. 6.8D and E).

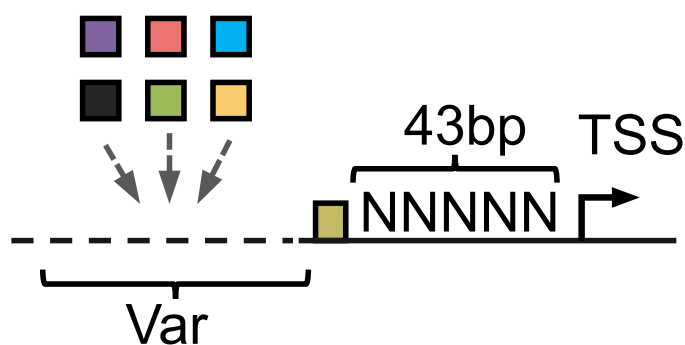


Figure 6.7: **Synthetic promoter design.** Synthetic nitrate-response promoters were designed containing TFBSs for TFs known to respond to nitrate. The TFBSs sequences were added into the upstream variable region of a previously published synthetic promoter consisting of a 43 bp core promoter sequence from *CaMV35s*²⁷⁸.

Synthetic promoters containing either four bZIP3 (fig. 6.8F) or four TGA1 TFBSs were also designed (fig. 6.8H). In addition, promoters containing two copies of two different TFBSs, bZIP3/NLP7 or TGA1/NLP7, were designed (fig. 6.8G and I). As HHO2 is known to be a transcriptional repressor, it was not possible to use the same synthetic promoter design. An alternative design was used in which binding sites were integrated

into a previously published constitutive minimal synthetic promoter (*minsyn_105*)²⁷⁸ to determine if expression would be reduced by HHO2. One design contained three HHO2 TFBSs separated by 20 bp of random spacers followed by *minsyn_105* (fig. 6.8J) and the other design contained the minimal constitutive promoter with three HHO2 TFBSs separated by 20 bp of random spacers immediately upstream of the TSS (fig. 6.8K).

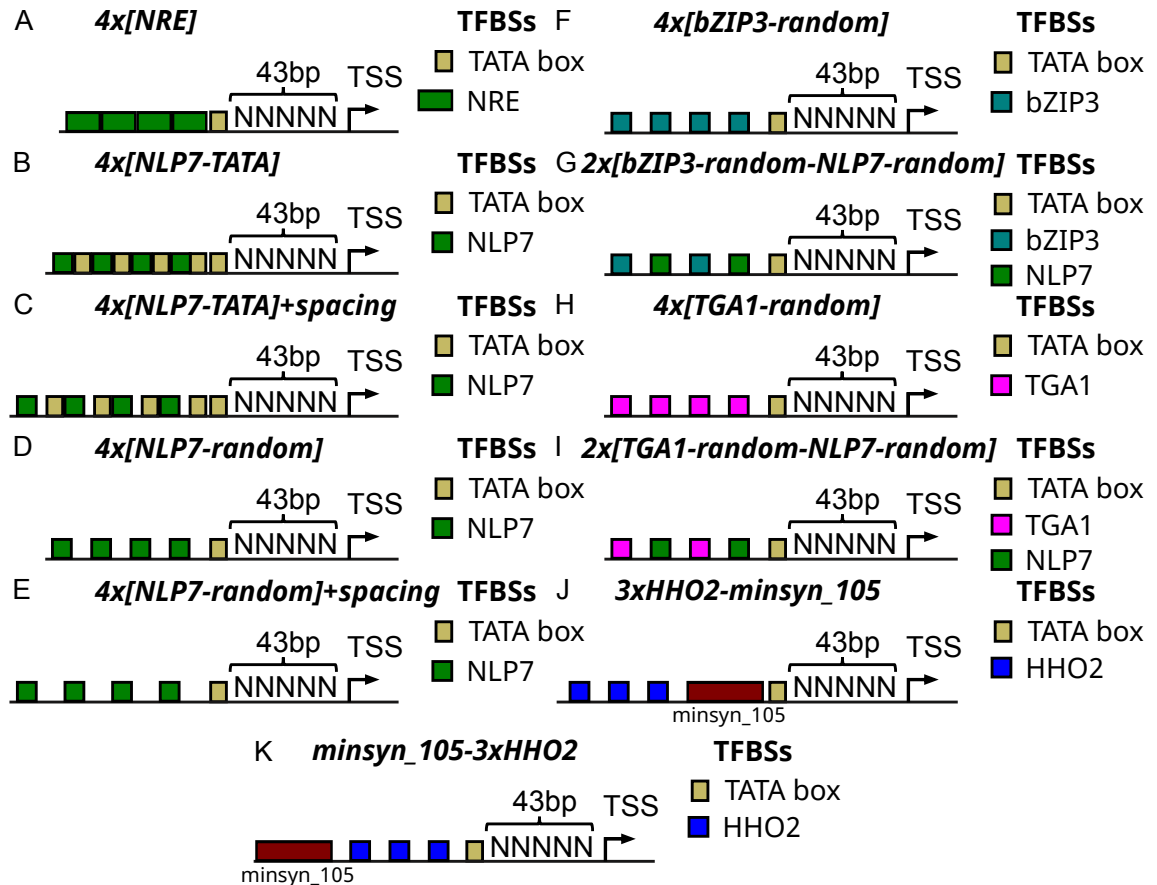


Figure 6.8: Synthetic promoter structure. Various synthetic nitrate-response promoters were designed containing TFBSs for TFs known to respond to nitrate. They included TFBSs for NLP7, bZIP3, TGA1 and HHO2 TFs. The TFBSs sequences were added into the upstream variable region of a previously published synthetic promoter consisting of a 43 bp core promoter sequence from *CaMV35s*²⁷⁸. A: Synthetic promoter containing four copies of NRE from *AtNIR1*, which contain the NLP7 TFBS. B–E: Synthetic promoters containing four copies of NLP7 TFBSs with different spacing and spacers. F: A synthetic promoter containing four copies of bZIP3 TFBSs. G: A synthetic promoter containing a combination of both NLP7 and bZIP3 TFBSs. H: A synthetic promoter containing four copies of TGA1 TFBSs. I: A synthetic promoter containing a combination of both NLP7 and TGA1 TFBSs. J–K: To create synthetic promoters that would be repressed by nitrate, HHO2 TFBSs were added to a constitutive minimal promoter called *minsyn_105* from Cai et al.²⁷⁸ in two different designs.

Each synthetic promoter was fused to a LucN luciferase reporter for use in protoplast

dual luciferase assays as used to assess Arabidopsis promoters in chapter 3. In this assay, luminescence is normalised to an experimental and batch calibrator (see section 2.5.1) and each synthetic promoter is co-expressed with either a constitutively expressed TF or YFP to ensure that the transcriptional load is equal in all experiments (see sections 2.1.2 and 3.5.5).

Regulatory effects of bZIP3 expression on a synthetic promoter-LucN construct containing bZIP3 TFBSs

To test whether the synthetic promoter containing four bZIP3 TFBSs was functional and activated by bZIP3 protein, bZIP3 was co-expressed with the synthetic promoter fused to the coding sequence of LucN. Normalised luminescence of $p_{4 \times [bZIP3\text{-random}]}-LucN$ was significantly higher following co-expression with $p_{CaMV35s}\text{-bZIP3}$ (35.03 ± 0.24 ; $t = -147.9$, $P < 0.001$; fig. 6.9) compared to the no TF control (8.05 ± 0.09).

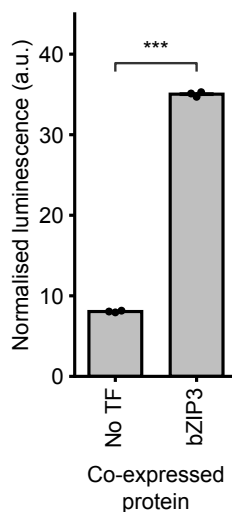


Figure 6.9: **Normalised luminescence of $p_{4 \times [bZIP3\text{-random}]}-LucN$ following co-expression with $p_{CaMV35s}\text{-AtbZIP3}$.** This promoter contains four bZIP3 TFBSs each separated with 20 bp of random sequence with equal ATCG ratios. No TF control = $p_{CaMV35s}\text{-YFP}$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

6.4.4 Regulatory effects of TGA1 expression on a synthetic promoter-LucN construct containing TGA1 TFBSs

To test whether the synthetic promoter containing four TGA1 TFBSs was functional and activated by TGA1 protein, TGA1 was co-expressed with the synthetic promoter fused to the coding sequence of LucN. Normalised luminescence of $p_{4 \times [TGA1-random]}$ -LucN was significantly increased by co-expression of TGA1 (31.04 ± 0.57 ; $t = -47.09$, $P < 0.001$; fig. 6.10) compared to the no TF control (11.85 ± 0.06).

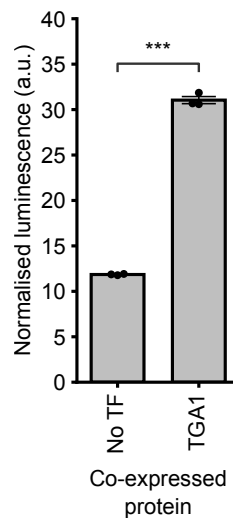


Figure 6.10: **Normalised luminescence of $p_{4 \times [TGA1-random]}$ -LucN following co-expression with $p_{CaMV35s-AtTGA1}$.** This promoter contains four TGA1 TFBSs each separated with 20 bp of random sequence with equal ATCG ratios. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

Regulatory effects of NLP7 expression on synthetic promoter-LucN constructs containing nitrate-responsive *cis*-elements (NREs)

To test whether the synthetic promoter containing four NRE motifs was functional and activated by NLP7 protein, NLP7 was co-expressed with the synthetic promoter fused to the coding sequence of LucN. Normalised luminescence of $p_{4 \times [NRE]}$ -LucN was significantly increased by co-expression of NLP7 (586.52 ± 16.03 ; $t = -11.4$, $P < 0.001$; fig. 6.11) compared to the no TF control (435.77 ± 9.73).

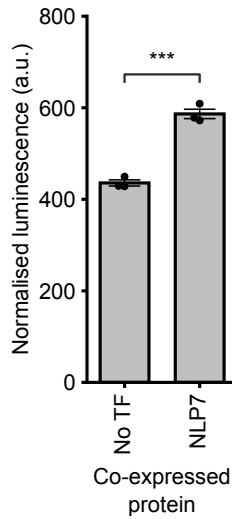


Figure 6.11: **Normalised luminescence of $p_4 \times [NRE]$ -LucN following co-expression with $p_{CaMV35s}$ -AtNLP7.** This promoter contains four copies of the nitrate-responsive *cis*-element NRE from *AtNIR1* (see table 6.1) with no spacer sequences between copies. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

To test if a shortened version of the NRE would be sufficient to drive LucN expression, a version with four copies of the central 24 bp region of the NRE, which still contained the essential conserved NLP7 TFBS motifs described in Konishi and Yanagisawa³⁹³, was used (see table 6.1). 8 bp of random sequence was inserted between each copy of the NRE to prevent steric hindrance. Normalised luminescence of $p_4 \times [NLP7\text{-}random]$ -LucN was significantly increased by co-expression of NLP7 (501.52 ± 1.84 ; $t = -175.3$, $P < 0.001$; fig. 6.12) compared to the no TF control (204.50 ± 1.53). Fold activation of $p_4 \times [NLP7\text{-}random]$ -LucN by NLP7 was increased ($2.45\times$) compared to the promoter containing four complete NRE motifs (fig. 6.11; $1.35\times$).

To test the effect of increased spacing between TFBSs on expression, a synthetic promoter with 28 bp of random sequence between each TFBS was constructed ($p_4 \times [NLP7\text{-}random] + \text{spacing}$ -LucN). Normalised luminescence of $p_4 \times [NLP7\text{-}random] + \text{spacing}$ -LucN was significantly increased by co-expression of NLP7 (576.28 ± 12.14 ; $t = -37.1$, $P < 0.001$; fig. 6.13) compared to the no TF control (248.00 ± 3.11). Co-expression of NLP7 resulted in a 2.32-fold increase in the expression of $p_4 \times [NLP7\text{-}random] + \text{spacing}$ -LucN compared to a 2.45-fold increase of $p_4 \times [NLP7\text{-}random]$ -LucN.

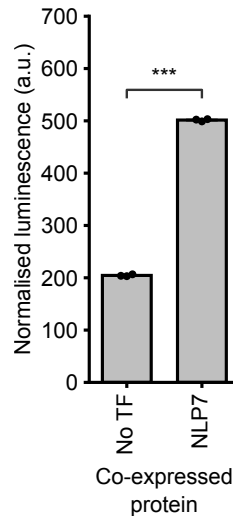


Figure 6.12: **Normalised luminescence of $p_4 \times [NLP7\text{-}random]\text{-LucN}$ following co-expression with $p_{CaMV35s}\text{-}AtNLP7$.** This promoter contains four NRE motifs, truncated to only include NLP7 TFBSs. Each element is separated by 7 bp of random spacer sequence with equal ATCG ratios. No TF control = $p_{CaMV35s}\text{-}YFP$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

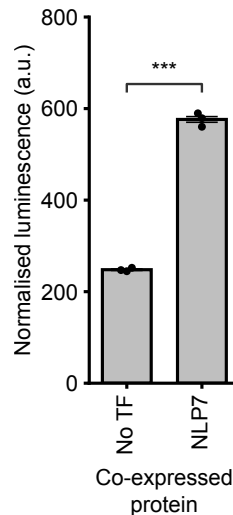


Figure 6.13: **Normalised luminescence of $p_4 \times [NLP7\text{-}random]\text{-}spacing\text{-LucN}$ following co-expression with $p_{CaMV35s}\text{-}AtNLP7$.** This promoter contains four NRE motifs, truncated to only include NLP7 TFBSs, with 28 bp of random spacer sequence with equal ATCG ratios downstream of each binding site. No TF control = $p_{CaMV35s}\text{-}YFP$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

To test whether the addition of TATA boxes would increase the fold activation of luminescence following co-expression with NLP7, a synthetic promoter with a 7 bp TATA box (TATATAA) 1 bp downstream of each NLP7 TFBS was constructed ($p_4 \times [NLP7\text{-}TATA]\text{-}$

LucN). Normalised luminescence of $p_4 \times [NLP7-TATA]$ -LucN was significantly increased by co-expression of NLP7 (73.48 ± 0.37 ; $t = -39.3$, $P < 0.001$; fig. 6.14) compared to the no TF control (23.28 ± 1.77). Co-expression of NLP7 resulted in a 3.16-fold increase of $p_4 \times [NLP7-TATA]$ -LucN compared to a 2.45-fold increase of $p_4 \times [NLP7-random]$ -LucN.

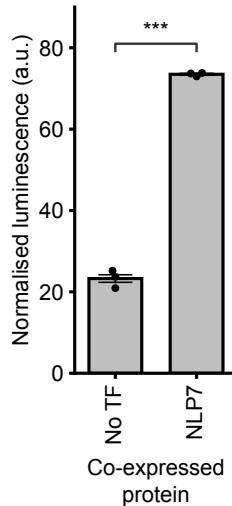


Figure 6.14: **Normalised luminescence of $p_4 \times [NLP7-TATA]$ -LucN following co-expression with $p_{CaMV35s-AtNLP7}$.** This promoter contains four NRE motifs, truncated to only include NLP7 TFBSs, with 7 bp TATA box (TATATAA) 1 bp downstream of each binding site. No TF control = $p_{CaMV35s-YFP}$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

To test the effect of spacing, a synthetic promoter with 21 bp of random sequence followed by a 7 bp TATA box downstream of each TFBSs was constructed ($4 \times [NLP7-TATA] + spacing$ -LucN). Normalised luminescence of $p_4 \times [NLP7-TATA] + spacing$ -LucN was significantly increased by co-expression of NLP7 (522.25 ± 4.34 ; $t = -82.0$, $P < 0.001$; fig. 6.15) compared to the no TF control (264.63 ± 1.00). Co-expression of NLP7 resulted in a 1.97-fold increase of $p_4 \times [NLP7-TATA] + spacing$ -LucN compared to a 3.16-fold increase of $p_4 \times [NLP7-TATA]$ -LucN. However, the luminescence of $p_4 \times [NLP7-TATA] + spacing$ -LucN was higher (522.25) than that of $p_4 \times [NLP7-TATA]$ -LucN (73.48) following co-expression with NLP7.

A previously published synthetic N-responsive promoter (*NRP*) known to respond to NLP7 was also tested as a positive control (p_{NRP} -LucN)¹. Normalised luminescence of p_{NRP} -LucN was significantly increased by co-expression of NLP7 (11.60 ± 0.46 ; $t = -25.0$, $P <$

0.01; fig. 6.16) compared to the no TF control (3.28 ± 0.07), a fold-change of 3.54.

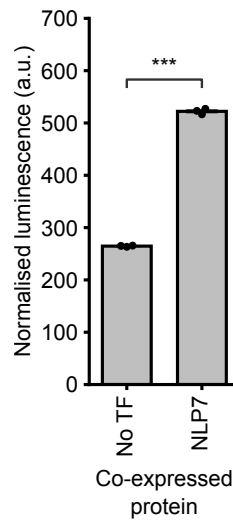


Figure 6.15: **Normalised luminescence of $p_{4 \times [NLP7-TATA]} + spacing-LucN$ following co-expression with $p_{CaMV35s-AtNLP7}$.** This promoter contains four NRE motifs, truncated to only include NLP7 TFBSs, with 21 bp of random spacer sequence with equal ATCG ratios downstream of each binding site, followed by a 7 bp TATA box (TATATAA). No TF control = $p_{CaMV35s-YFP}$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

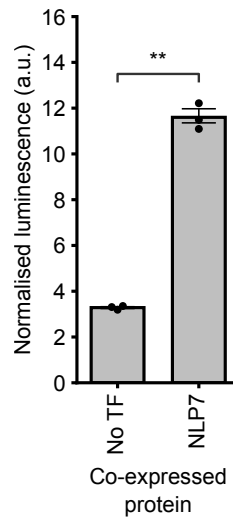


Figure 6.16: **Normalised luminescence of $p_{NRP-LucN}$ following co-expression with $p_{CaMV35s-AtNLP7}$.** No TF control = $p_{CaMV35s-YFP}$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

Regulatory effects of TF expression on the luminescence of a synthetic promoter-LucN construct containing bZIP3 and NLP7 TFBSs

To test whether a synthetic promoter containing both bZIP3 and NLP7 TFBSs would respond to co-expression with both TFs, a synthetic promoter (${}_{\text{p}}2\times[\text{bZIP3-random-NLP7-random}]\text{-LucN}$) containing two copies of each TFBS in alternating positions was tested. Normalised luminescence of ${}_{\text{p}}2\times[\text{bZIP3-random-NLP7-random}]\text{-LucN}$ was significantly increased by co-expression of bZIP3 (161.24 ± 2.29 ; $t = -3.9$ $P < 0.05$), NLP7 (291.15 ± 5.10 ; $t = -32.9$ $P < 0.001$), and both bZIP3/NLP7 (297.00 ± 5.10 ; $t = -54.1$, $P < 0.001$; fig. 6.17) compared to the no TF control (150.28 ± 3.26).

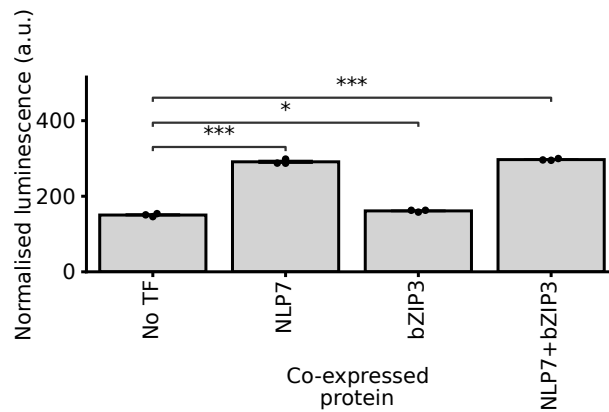


Figure 6.17: **Normalised luminescence of ${}_{\text{p}}2\times[\text{bZIP3-random-NLP7-random}]\text{-LucN}$ following co-expression with bZIP3 and/or NLP7.** This promoter contains two copies each of bZIP3 and NLP7 TFBSs with 20 bp of random sequence with equal ATCG ratios downstream of each binding site. No TF control = ${}_{\text{p}}\text{CaMV35s-YFP}$. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

6.4.5 Regulatory effects of TF expression on the luminescence of a synthetic promoter-LucN construct containing TGA1 and NLP7 binding sites

To test whether a synthetic promoter containing both TGA1 and NLP7 TFBSs would respond to co-expression with both TFs, a synthetic promoter (${}_{\text{p}}2\times[\text{TGA1-random-NLP7-random}]\text{-LucN}$) containing two copies of each TFBS in alternating positions was tested.

Normalised luminescence of $p_{2 \times [TGA1-random-NLP7-random]}$ -LucN was significantly increased by co-expression of TGA1 (100.33 ± 2.38 ; $t = -14.5$ $P < 0.01$), NLP7 (299.0 ± 1.98 ; $t = -149.7$ $P < 0.001$), and both TGA1/NLP7 (390.21 ± 10.20 ; $t = -43.6$, $P < 0.001$; fig. 6.18) compared to the no TF control (74.80 ± 0.76).

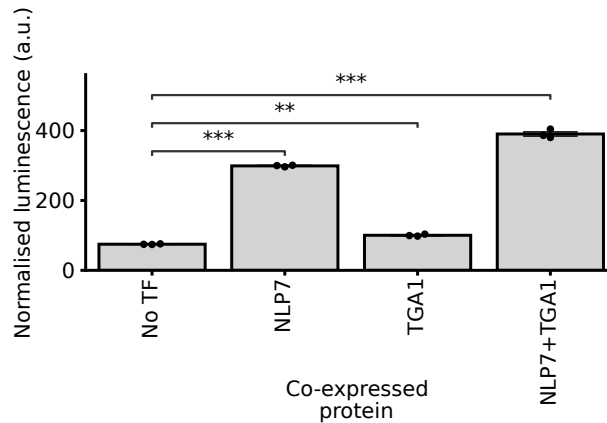


Figure 6.18: **Normalised luminescence of $p_{2 \times [TGA1-random-NLP7-random]}$ -LucN following co-expression with TGA1 and/or NLP7.** This promoter contains two copies each of TGA1 and NLP7 TFBSs with 20 bp of random sequence with equal ATCG ratios downstream of each binding site. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

Regulatory effects of TF expression on two synthetic promoter-LucN constructs containing the HHO2 TFBS

To create a synthetic promoter that was repressed by HHO2, two promoter designs were tested. The first encoded three HHO2 TFBSs followed a minimal synthetic constitutive promoter from Cai et al.²⁷⁸ ($3 \times HHO2-minsyn_{105}$). The second encoded the minimal constitutive promoter with three HHO2 TFBSs added immediately upstream of the TSS ($minsyn_{105}-3 \times HHO2$). Unexpectedly, normalised luminescence of $p_{3 \times HHO2-minsyn_{105}}$ -LucN was significantly increased by co-expression of HHO2 (429.18 ± 2.58 ; $t = -18.4$, $P < 0.001$) compared to the no TF control (347.94 ± 5.70 ; fig. 6.19).

However, normalised luminescence of $p_{minsyn_{105}-3 \times HHO2}$ -LucN was significantly reduced by co-expression of HHO2 (28.59 ± 0.25 ; $t = 18.44$, $P < 0.001$) compared to the no TF control (32.73 ± 0.19 ; fig. 6.20).

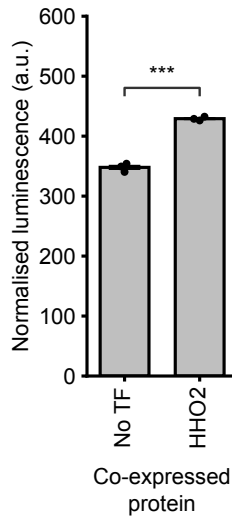


Figure 6.19: **Normalised luminescence of $p_{3\times HHO2}$ -*minsyn_105*-LucN following co-expression with $p_{CaMV35s}$ -*AtHHO2*.** Promoter contains three HHO2 TFBSs separated by 20 bp of random spacers with equal ATCG ratios followed by a minimal synthetic constitutive promoter (*minsyn_105*) from Cai et al.²⁷⁸. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

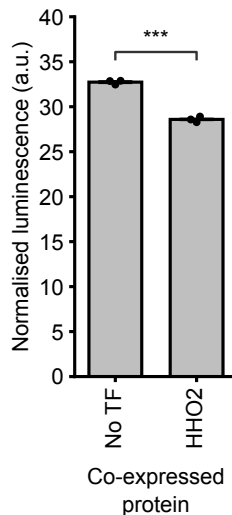


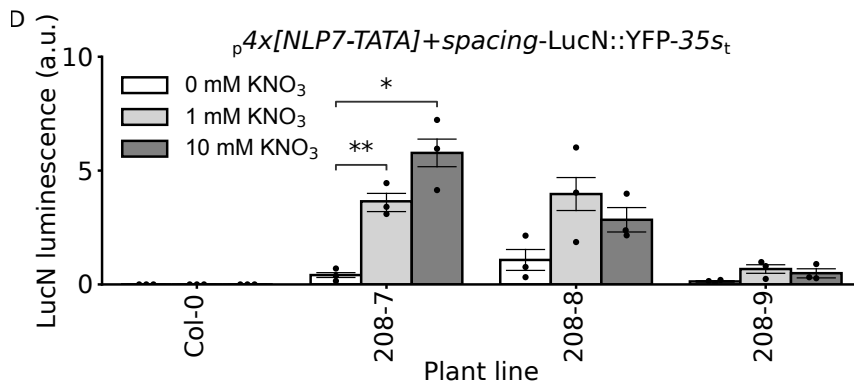
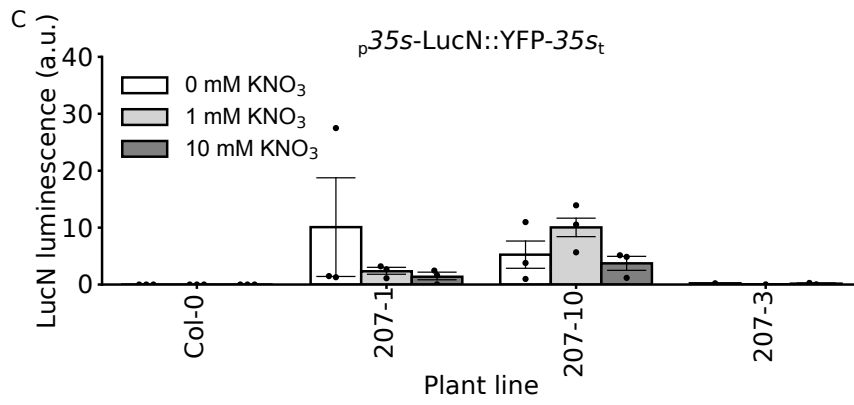
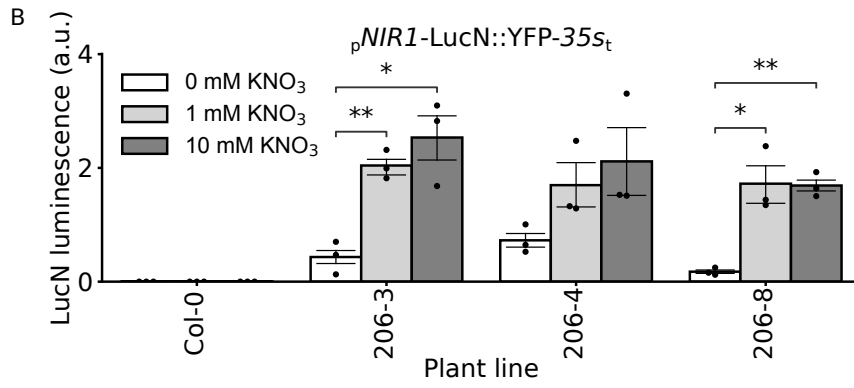
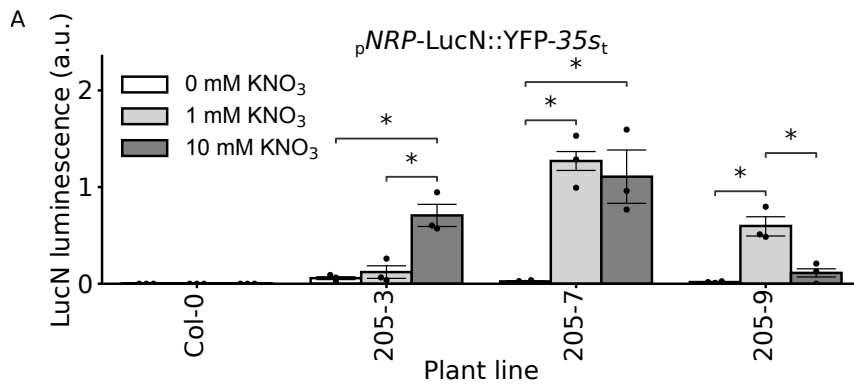
Figure 6.20: **Normalised luminescence of p_{minsyn_105} - $3\times HHO2$ -LucN following co-expression with $p_{CaMV35s}$ -*AtHHO2*.** Promoter contains a minimal synthetic constitutive promoter (*minsyn_105*) from Cai et al.²⁷⁸ with three HHO2 TFBSs separated by 20 bp of random spacers with equal ATCG ratios immediately upstream of the TSS. No TF control = $p_{CaMV35s}$ -YFP. Error bars represent two standard errors. Luminescence was measured 18 hours after protoplast transfection. $N = 3$ technical replicates. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

Investigating the nitrate responsiveness of synthetic minimal promoters in transgenic plants

Two synthetic promoters ($4 \times [NLP7-TATA] + spacing$ and $4 \times [bZIP3-random]$) were selected to investigate their nitrate responsiveness because they were strongly activated when co-expressed with their respective cognate TFs. *NIR1* and *NRP* were used as positive controls as these promoters have previously been shown to respond to nitrate^{1,396}. An additional promoter (*CaMV35s*) was included as an experimental control as it was expected to express in all tissues and conditions. To determine whether the synthetic promoters would respond to nitrate, synthetic promoters were fused to a LucN::YFP reporter to enable both visualisation and quantification of responses to nitrate in stable transgenic plants. Constructs containing the LucN::YFP reporter were co-assembled between FAST-Red^{325,328} and bar phosphinothricin³²⁵ selectable marker genes and final multigene constructs were used to transform *Arabidopsis* plants (see section 2.1.2). FAST-Red positive T1 seeds were collected and grown on MS media then soil. Three lines were selected for each construct and T2 seeds were collected. 12 FAST-Red positive T2 seeds of each line were grown on 0, 1 and 10 mM KNO₃ on vertical square 12 mm plates (Greiner Bio-One 688161). Roots from 12-day old plants were harvested from square plates (see section 2.2.2). Each biological replicate contained the combined roots from three plants. Protein was extracted and LucN luciferase assays were conducted, with normalisation of transgene-LucN luminescence activity to protein absorbance at 660 nm (see section 2.5.2).

Lines containing $p_{NRP-LucN::YFP-CaMV35s_t}$ had increased luminescence on higher nitrate concentrations

Luminescence from line plntEPSWT20205-3 containing $p_{NRP-LucN::YFP}$ was significantly increased on 10 mM KNO₃ (0.71 ± 0.17) compared to either 1 mM (0.12 ± 0.01 ; $t = -4.2$, $P < 0.05$) or 0 mM KNO₃ (0.06 ± 0.03 ; $t = -5.3$, $P < 0.05$; fig. 6.21A). Similarly, luminescence was significantly lower from line plntEPSWT20205-7 containing $p_{NRP-LucN::YFP}$ on 0 mM KNO₃ (0.02 ± 0.01) compared to 1 mM KNO₃ (1.27 ± 0.22 ; $t = -8.0$, $P < 0.05$) and 10 mM KNO₃ (1.11 ± 0.35 ; $t = -4.3$, $P < 0.05$; fig. 6.21A). However, luminescence from line plntEPSWT20205-9 containing $p_{NRP-LucN::YFP}$ was higher on 1 mM KNO₃ (0.60 ± 0.14) compared to 0 mM KNO₃ (0.02 ± 0.01 ; $t = -5.8$, $P < 0.05$) or 10 mM KNO₃ ($0.11 \pm$



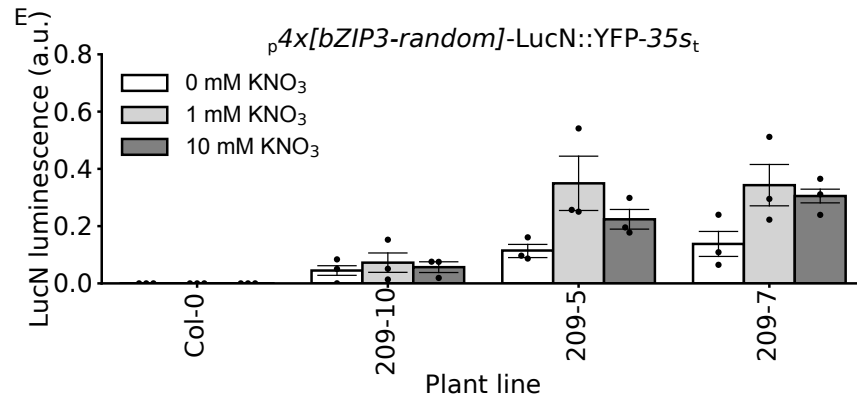


Figure 6.21: **Normalised luminescence from the roots of transgenic lines containing synthetic promoter-LucN::YFP constructs.** LucN luminescence activity was normalised to protein absorbance at 660 nm. A-E: Luminescence of extracted root protein from lines containing promoter:LucN constructs when grown on 0, 1 and 10 mM KNO₃. Col-0 showed no luminescence activity. Error bars represent two standard errors. $N = 3$ biological replicates of 3 pooled plants each. P -values were calculated using Welch's t-tests³⁴⁹. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$.

0.09; $t = -4.16$, $P < 0.05$; fig. 6.21A).

Lines containing $pNIR1$ -LucN::YFP-*CaMV35s*_t had increased luminescence on higher nitrate concentrations

Luminescence from lines plntEPSWT20206-3 and plntEPSWT20206-8 containing $pNIR1$ -LucN::YFP was significantly higher on both 10 mM KNO₃ (plntEPSWT20206-3: 2.53 ± 0.61 , $t = -4.5$, $P < 0.05$; plntEPSWT20206-8: 1.69 ± 0.18 , $t = -11.7$, $P < 0.01$) and 1 mM KNO₃ (plntEPSWT20206-3: 2.04 ± 0.21 , $t = -7.3$, $P < 0.01$; plntEPSWT20206-8: 1.72 ± 0.47 , $t = -4.6$, $P < 0.05$) compared to 0 mM KNO₃ (plntEPSWT20206-3: 0.43 ± 0.24 ; plntEPSWT20206-8: 0.18 ± 0.05 ; fig. 6.21B). Although line plntEPSWT20206-4 showed a similar trend, there was no significant difference between conditions ($P > 0.05$; fig. 6.21B).

Lines containing $p4 \times [NLP7-TATA]+spacing$ -LucN::YFP-*CaMV35s*_t had increased luminescence on higher nitrate concentrations

Luminescence from line plntEPSWT20208-7 containing $p4 \times [NLP7-TATA]+spacing$ -LucN::YFP was significantly higher when grown on 10 mM KNO₃ (5.78 ± 1.26 ; $t = -5.9$, $P < 0.05$) and 1 mM KNO₃ (3.65 ± 0.58 ; $t = -7.4$, $P < 0.01$) compared to 0 mM KNO₃ (0.41 ± 0.22 ; fig. 6.21D). In line plntEPSWT20208-8, $p4 \times [NLP7-TATA]+spacing$ -LucN::YFP luminescence was higher on 1 and 10 mM KNO₃ than 0 mM KNO₃, although this was not

significant ($P > 0.05$; fig. 6.21D). Luminescence from line plntEPSWT20208-9 containing $p_{4 \times [NLP7-TATA]+spacing}$ -LucN::YFP was low regardless of nitrate concentration and was not significantly different between conditions ($P > 0.05$; fig. 6.21D).

Lines containing $p_{4 \times [bZIP3-random]}$ -LucN::YFP-*CaMV35s_t* had increased luminescence on higher nitrate concentrations

Luminescence from lines plntEPSWT20209-5, plntEPSWT20209-7, and plntEPSWT20209-10, which contained $p_{4 \times [bZIP3-random]}$ -LucN::YFP, was higher on 1 and 10 mM KNO₃ than 0 mM KNO₃, however, this was not significant ($P > 0.05$; fig. 6.21E).

Lines containing $p_{CaMV35s}$ -LucN::YFP-*CaMV35s_t* did not increase with higher nitrate

As expected, there were no significant differences in luminescence between conditions from lines plntEPSWT20207-1, plntEPSWT20207-3, or plntEPSWT20207-10 which contained $p_{CaMV35s}$ -LucN::YFP ($P > 0.05$; fig. 6.21C).

6.4.6 Construction of synthetic genetic feedback controllers

To introduce synthetic genetic feedback into the Arabidopsis N-responsive subnetwork, the expression of *ARF18* must be regulated by the expression of a TF further down the transcriptional cascade, such as NLP7 (see fig. 6.1). To achieve this, the best performing NLP7-responsive synthetic promoter ($4 \times [NLP7-TATA]+spacing$) was assembled with the synthetic activators (SunTag-2 \times TAD and SunTag-VP64 each with three sgRNAs) and repressors (dCas12a::NLS::3 \times SRDX with three crRNAs) (fig. 6.22). In addition, equivalent constructs were assembled using the known N-responsive promoter, *NRP¹*, as well as $4 \times [bZIP3-random]$, in place of the $4 \times [NLP7-TATA]+spacing$ promoter (fig. 6.22).

These expression cassettes were co-assembled between FAST-Red and bar selectable marker genes, and the final multigene constructs were used to transform Arabidopsis Col-0 plants (see section 2.2.4).

T1 seeds were collected and those showing either RFP fluorescence or resistance to phosphinothricin were grown for the generation of T2 seeds (see section 2.2.5). At the time of writing, analysis of these lines is still outstanding.

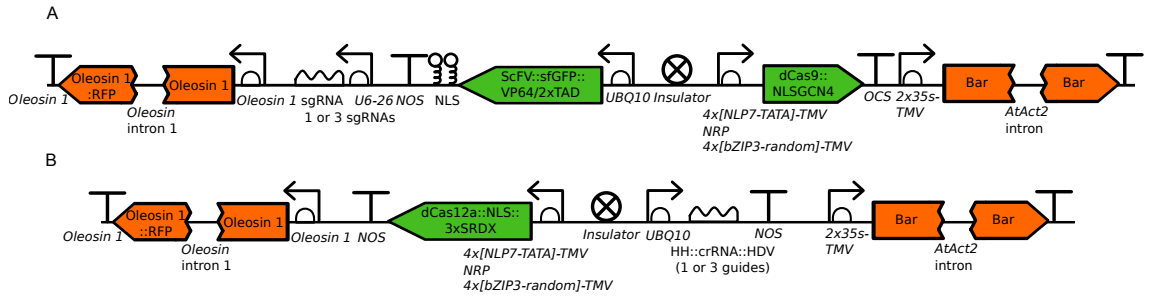


Figure 6.22: **Schematic of synthetic feedback controllers.** A: This synthetic transcriptional activator feedback controller requires co-expression of a dCas9::NLSGCM4 fusion under the control of either a synthetic minimal NLP7-responsive promoter, the NRP^1 promoter, or a synthetic minimal $4 \times [bZIP3-random]$ promoter, and a fusion of scFv::sfGFP fused to either a VP64 domain or two repeats of the TAD activation domain³⁰¹. The protein complex is delivered to the target gene by forming a complex with sgRNAs. B: This synthetic transcriptional repressor feedback controller requires co-expression of a dCas12a::SRDX fusion under the control of either a synthetic minimal NLP7-responsive promoter, the NRP^1 promoter, or a synthetic minimal $4 \times [bZIP3-random]$ promoter. The protein complex is delivered to the target gene by forming a complex with crRNA, which is encoded between hammerhead/hepatitis delta virus ribozymes. Both constructs also contain the FAST-Red^{325,328} and bar phosphinothricin³²⁵ marker genes to allow the selection of plant lines in which the complete T-DNA has integrated into the genome.

6.5 Discussion

6.5.1 Synthetic transcription factors regulate the expression of p_{ARF18} -LucN

Luminescence obtained from p_{ARF18} -LucN increased when co-expressed with Cas9/SunTag-based synthetic transcriptional activators with $2 \times$ TAD or VP64 activator domains. Luminescence was highest when three guides were used, and luminescence was higher with $2 \times$ TAD than with VP64. These results are in agreement with other studies in which the $2 \times$ TAD activator domain also showed stronger activation than the VP64 domain³⁰¹.

Only the dCas12a::NLS:: $3 \times$ SRDX repressor system showed a decrease in luminescence of p_{ARF18} -LucN. A previous study also showed repression of Arabidopsis genes using the dCas12a::NLS:: $3 \times$ SRDX system²⁹⁴. However, for both Cas9/SunTag-based synthetic transcriptional activators and the dCas12a::NLS:: $3 \times$ SRDX repressors, luminescence of p_{ARF18} -LucN was increased when compared to the no TF control. Increases in luminescence were greater for SunTag constructs which recruited multiple copies of the repressor domains as compared to just one set of $3 \times$ SRDX repressor domains in dCas12a systems. The reasons for the increases in luminescence are not clear, however, the SRDX domain is

known to recruit chromatin remodellers to repress gene expression^{293,295}. In addition, chromatin in protoplasts has been shown to be more open than in whole plants⁴¹⁷. Therefore, the SRDX domain may function differently in protoplasts.

In this study, co-expression of p_{ARF18} -LucN was used to test the activity of synthetic TFs because luminescence measurements are rapid and require only a few protoplasts. This was a significant consideration due to the large number of constructs being compared. An alternative assay measuring the expression of the endogenous *ARF18* gene in response to the expression of synthetic TFs using qRT-PCR might give more reliable performance predictions of integrated transgenes. A recent study which generated synthetic activators and repressors also had difficulty engineering reliable synthetic repressors²⁵⁹.

It will be important in the future to repeat the protoplast co-expression experiments with separate transfections to ensure that the results are reproducible. Additionally, it would be useful to repeat these experiments in whole plants to determine if the results obtained in protoplasts are representative of the activity of the synthetic TFs on the native *arf18* expression.

The promoters driving the experimental calibrator *CaMV35s::luciferase* were shown to respond to co-expressed transcription factors in chapter 3. It would be useful in the future to test whether the co-expression of the synthetic TFs affected the activity of the experimental calibrator. As the synthetic TFs use sgRNAs to specifically target the ARF18 promoter, it is unlikely that the synthetic TFs would affect the activity of the experimental calibrator. Overall, a range of synthetic transcriptional activators and repressors of different strengths were designed and tested transiently in protoplasts. It will require further testing to determine which synthetic TFs are most suitable for use in stable lines.

6.5.2 Synthetic minimal promoters that respond to specific TFs

The aim to build synthetic minimal promoters that respond to specific TFs was successful, with several novel functional promoters being built. The $p_{4 \times [bZIP3-random]}$ -LucN promoter was activated by bZIP3, with a 4.35-fold increase in LucN activity in response to bZIP3. This was as expected, since bZIP3 mostly activates genes containing TFBSs in their promoters³¹⁸. Similarly, the $p_{4 \times [TGA1-random]}$ -LucN promoter was activated by TGA1, with a 2.62-fold increase in LucN activity in response to TGA1. This was

also expected, since TGA1 mostly activates genes containing TFBSs in their promoters³¹⁸. TGA1 and bZIP3 motifs are similar, with a conserved TGACGT core (table 6.1), and they are both in the bZIP family⁵²⁰. In the $4\times[bZIP3-random]$ and $4\times[TGA1-random]$ synthetic promoters, the motifs were separated by 20 bp of random sequence. In Mehrotra and Mehrotra²⁸⁷, synthetic promoters were tested that contained ACGT motifs, which are known to bind bZIP proteins⁵²¹. The ACGT motifs were separated by 5, 10 and 25 bp random sequences. The promoter containing the 5 bp spacers was found to be strongly activated by salicylic acid, while the one with 25 bp was only activated by abscisic acid, suggesting that different TFs regulated the promoter depending on the spacing²⁸⁷. In the future it would be useful to test further bZIP3 and TGA1 synthetic promoters with different spacing between motifs, to find the optimal spacing for activation by these TFs.

The NRE motif in the *NIR1* 5' UTR is a well-known 43 bp nitrate response element which binds NLP TFs, which are known to activate *NIR1*³⁹⁶. The $p_{4\times[NRE]}$ -LucN promoter was activated by NLP7, with a 1.35-fold increase in LucN activity. The baseline activity of this promoter without co-expression of NLP7 was relatively high compared to other tested synthetic promoters. This might have been because of the presence of additional TFBS within the NRE that attract additional TFs. To reduce the background activity of the promoter, the NRE motif was truncated to contain only the essential core NLP7 motifs as described in Konishi and Yanagisawa³⁹³ (see table 6.1). This promoter was modified by the addition of 8 bp random spacers with equal ATCG ratios between each NLP7 TFBS to make the $4\times[NLP7-random]$ promoter. The luminescence of $p_{4\times[NLP7-random]}$ -LucN showed a 2.45-fold increase in LucN activity when co-expressed with NLP7, compared to the no TF control. This was a greater fold induction than obtained from $p_{4\times[NRE]}$ -LucN (1.35 \times). The expression level of $p_{4\times[NLP7-random]}$ -LucN in the absence of NLP7 was reduced as compared to $p_{4\times[NRE]}$ -LucN indicating that the NRE motif is likely to be attracting additional TFs. Since the length and identity of sequence between TFBSs can affect the hierarchical arrangement of TFs and the recruitment of the RNA polymerase complex^{281,287}, a variant of this synthetic promoter was tested that contained a 28 bp spacer between motifs. The hypothesis was that increasing the spacing between motifs might reduce steric hindrance and allow high-affinity binding. However, although the $p_{4\times[NLP7-random]+spacing}$ -LucN responded to co-expression of NLP7, with a 2.32-fold increase in LucN activity compared to the no TF control, this promoter had a slightly

lower fold increase as compared to $p_{4 \times [NLP7\text{-}random]}$ -LucN. This suggests that 8 bp of spacing between TFBSs is enough for correct binding of NLP7.

The TATA box motif supports the binding of TATA box binding proteins and the formation of the pre-initiation complex. Although most plant genes do not contain a TATA motif¹⁵⁷, in chapter 5 it was shown that genes that have variable expression patterns are enriched for TATA boxes. Additionally, early stress response genes are more likely to contain a TATA box than late response genes⁴⁹³. Kotopka and Smolke²⁷¹ found that, in yeast, synthetic promoter designs containing a TATA box had increased activity compared to those without a TATA box. The authors also found that their CNN model placed TATA-like motifs at the 5' end of TFBSs in the best performing (strongest) promoters. Therefore, the hypothesis that synthetic promoters containing TATA boxes would be more responsive to NLP7 was tested, with 7 bp TATA boxes (TATATAA) placed in between each binding site. The luminescence of $p_{4 \times [NLP7\text{-}TATA]}$ -LucN was increased by NLP7, with a 3.16-fold increase in LucN activity in response to NLP7. This 3.16-fold increase was a slightly higher induction than the 2.45-fold increase in luminescence by $p_{4 \times [NLP7\text{-}random]}$ -LucN following co-expression with NLP7, suggesting that the TATA boxes increased the induction of the promoter to NLP7. However, the luminescence obtained from $p_{4 \times [NLP7\text{-}random]}$ -LucN was 6.83 \times higher than $p_{4 \times [NLP7\text{-}TATA]}$ -LucN following co-expression with NLP7, suggesting that the TATA box lowered the overall expression of the promoter, potentially due to TATA-box binding proteins competing for space with NLP7 while binding $4 \times [NLP7\text{-}TATA]$. A version of the promoter with 21 bp spacing upstream of each TATA box was also tested, with the hypothesis that the increased spacing would allow more room for NLP7 and TATA-box binding proteins to bind. The luminescence obtained from $p_{4 \times [NLP7\text{-}TATA]+spacing}$ -LucN was increased by NLP7, with a 1.97-fold increase in LucN activity in response to NLP7. Although this was a lower fold increase in luminescence compared to that of $p_{4 \times [NLP7\text{-}TATA]}$ -LucN (3.16 \times), the luminescence obtained from $p_{4 \times [NLP7\text{-}TATA]+spacing}$ -LucN was 7.11 \times higher than $p_{4 \times [NLP7\text{-}TATA]}$ -LucN following co-expression with NLP7, suggesting that the increased spacing allowed for more NLP7 binding and increased the overall expression of the promoter.

Luminescence from p_{NRP} -LucN increased following co-expression with NLP7, in line with previously published data¹. The luminescence of p_{NRP} -LucN (11.60) was 45 \times lower than

that of $p_4 \times [NLP7-TATA] + spacing$ -LucN (522.25) following co-expression of NLP7. A range of synthetic promoter strengths is useful for the tuning of gene expression and will be useful for modulating the expression of genes in the N-response subnetwork, including through the introduction of genetic feedback. To further increase the range of promoter strengths, the sequence of the TATA boxes in $p_4 \times [NLP7-TATA] + spacing$ -LucN could be modified, as in a previous study where *CaMV35s* promoter activity was tuned by mutating the TATA box⁵²².

To test whether synthetic promoters that contain TFBSs for two different TFs would respond to both TFs, motifs were arranged in the same promoter in an alternating pattern, separated by 20–21 bp random spacers. The $p_2 \times [bZIP3-random-NLP7-random]$ -LucN promoter was activated by both NLP7 and bZIP3, although expression of NLP7 increased luminescence by a greater fold difference (1.94-fold) as compared to bZIP3 (1.07-fold). The $p_2 \times [TGA1-random-NLP7-random]$ -LucN promoter was also activated by both NLP7 and TGA1; NLP7 also increased luminescence by a greater fold difference (4.00-fold) than TGA1 (1.34-fold). The activating effect of co-expressing both NLP7 and TGA1 was additive, resulting in a 5.22-fold increase in luminescence.

Luminescence obtained from expression of $p_{minsyn_105-3} \times HHO2$ -LucN decreased with co-expression of HHO2. This supports the finding of an earlier study in which HHO2 was shown to directly repress the NRT2.1 promoter⁵¹⁹. The reduction in luminescence in this promoter is likely to be due to the positioning of the HHO2 TFBSs close to the TSS. The binding of HHO2 proteins may prevent the formation of the pre-initiation complex. Interestingly, Brooks et al.³¹⁸ found that the 5' UTRs of genes repressed by HHO2 were enriched in HHO2 TFBSs. In the future, it would be interesting to add HHO2 TFBSs to the 5' UTRs of synthetic promoters. Surprisingly, luminescence obtained from the expression of $p_3 \times HHO2-minsyn_105$ -LucN increased with co-expression of HHO2, suggesting that the location of these TFBSs may modulate the activity of the proteins they bind.

It would be useful in the future to repeat all protoplast co-expression experiments with separate transfections of each plasmid, to ensure that similar results are obtained. There is expected to be some variability between batches despite having a batch calibrator, as this was observed with other protoplast co-expression experiments. In chapter 3, the promoters driving the experimental calibrator *CaMV35s::luciferase* were shown to respond

to co-expressed transcription factors. Identifying a promoter that does not respond to the co-expression of the TFs would be useful in future experiments to use in the experimental calibrator, to ensure any observed difference in luminescence is due to the regulation of the synthetic promoter. However, as the fold-change in luminescence with and without co-expression of the respective TFs was high in the majority of experiments, it is unlikely that the significant changes were due to the experimental calibrator. Additionally, later experiments (see section 6.5.3) showed that the synthetic promoters responded to nitrate so were likely functional.

6.5.3 Synthetic promoters respond to nitrate

Several synthetic constructs were used to produce stable transgenic lines to investigate if expression in plants roots could be regulated by nitrate. To do this, T2 plants of three independent lines for each construct were grown in three different nitrate conditions and changes in expression were assessed by quantifying luminescence.

As expected, there was no luminescence in Col-0 control plants that did not contain transgenes and the luminescence obtained from lines containing the constitutively expressed $p_{CaMV35s}$ -LucN::YFP construct was not significantly affected by nitrate. One line (plntEPSWT20207-3) containing this construct had low values in all conditions, likely due to transgene silencing. For all constructs, there was variation between plant lines likely due to differences in the transgene insertion site, which has been observed to influence expression⁵²³. Variation could have also been caused by differences in the copy number of the transgenes. It will be important in the future to determine the copy number of transgenes in the lines to rule this out. As roots from three T2 plants of each line were pooled, variation within lines is difficult to assess. However, some variation might be expected because individual T2 progeny of selfed T1 transgenics, in which the transgenes are expected to be hemizygous, are likely to be segregating. Although the presence of the transgene in T2 seeds was assessed by RFP fluorescence from the FAST-red marker^{325,328}, zygosity was not determined.

Luminescence from two lines (plntEPSWT20205-3 and 7) containing p_{NRP} -LucN::YFP was increased on 1 and 10 mM KNO_3 compared to 0 mM KNO_3 . This supports existing data showing that *NRP* expression is induced by nitrate¹. Expression of *NIR1* is also

known to be activated by nitrate^{393,396}. Supporting this, luminescence of lines containing the p_{NIR1} -LucN::YFP construct also increased with nitrate concentration. Both *NIR1* and *NRP* contain NRE motifs that bind NLP7. Recently, NLP7 was shown to be a direct sensor of nitrate and the mutation of all seven NLP TFs abolishes primary nitrate responses in Arabidopsis⁴⁹. Although the expression of *NIR1* and *NRP* in response to nitrate may also be regulated by additional TFs, the presence and potentially the concentration of NLP7 (which may be regulated by TFs in the subnetwork elucidated in chapter 3) is likely to have a substantial effect on the nitrate-responsiveness of *NIR1* and *NRP*. Therefore, the expression of minimal synthetic promoters that only attract NLP7 are expected to respond to nitrate. Previously, a $4 \times NRE$ synthetic promoter was shown to be activated by nitrate³⁹³, which was supported by the luminescence values obtained from two plant lines containing $p_{4 \times [NLP7-TATA]} + spacing$ -LucN::YFP. Luminescence increased when plants were grown on 1 and 10 mM KNO₃ compared to 0 mM KNO₃. Luminescence of one line (plntEPSWT20208-9) was very low in all nitrate conditions, potentially due to transgene silencing.

Out of all tested lines, line plntEPSWT20205-7 containing p_{NRP} -LucN::YFP had the largest fold increase (63.5 \times) in luminescence from 0 mM KNO₃ to 1 mM KNO₃. Line plntEPSWT20205-7 also had a 55.5-fold increase in luminescence from 0 mM KNO₃ to 10 mM KNO₃. The $p_{4 \times [NLP7-TATA]} + spacing$ -LucN::YFP construct gave the next highest increase in luminescence in line plntEPSWT20208-7, with a 14.1-fold increase from 0 mM KNO₃ to 10 mM KNO₃, and an 8.9-fold increase in luminescence from 0 mM KNO₃ to 1 mM KNO₃. The p_{NIR1} -LucN::YFP construct gave a 9.56-fold increase in luminescence in line plntEPSWT20206-8 from 0 mM KNO₃ to 1 mM KNO₃, and a 9.39-fold increase from 0 mM KNO₃ to 10 mM KNO₃.

Luminescence from two plant lines (plntEPSWT20209-5 and 7) containing $p_{4 \times [bZIP3-random]}$ -LucN::YFP was higher on 1 and 10 mM KNO₃ compared to 0 mM KNO₃, although this was not a significant increase. Out of lines containing $p_{4 \times [bZIP3-random]}$ -LucN::YFP, line plntEPSWT20209-5 had the largest fold change in luminescence with only a 3.00-fold increase from 0 mM KNO₃ to 1 mM KNO₃, and 1.95-fold increase from 0 mM KNO₃ to 10 mM KNO₃. One reason that larger fold-changes (4.35-fold) were observed in transient assays where bZIP3 was co-expressed but were not observed in stable transgenic lines

might be because the levels of endogenous bZIP3 are relatively low. Expression of bZIP3 is only induced 2.84-fold after 30 minutes of nitrate treatment⁵²⁴.

Although *NLP7* expression is not induced by nitrate, NLP7 is post-translationally regulated by nitrate and rapidly accumulates in the nucleus following nitrate treatment⁴⁰. This rapid accumulation of NLP7 in the nucleus may result in higher concentrations of NLP7 protein explaining why synthetic promoters with NLP7 binding sites have a greater fold-change in response to nitrate compared to those that attract bZIP3. Further, gene targets of NLP7 continue to be actively transcribed even after the TFs dissociates from its targets after transient-binding²⁶, so a lower concentration of NLP7 protein might be required than bZIP3 protein to induce transcription of target genes. Knowledge of relative protein levels of these TFs would be helpful in informing the design of synthetic promoters. However, a previous study which carried out intensity-based absolute quantification (iBAQ) of protein content showed iBAQ values of 4.56 for NLP7 and 5.35 for bZIP3 in Arabidopsis roots⁵²⁵. Although the protein level of bZIP3 was higher than NLP7 in this study, some bZIP3 protein might be inactive due to the formation of heterodimers with other proteins, which can modify the DNA-binding domain of bZIP proteins⁵²⁶. Additionally, TFs from different families such as bZIPs and basic helix-loop-helices (bHLHs) can compete to bind the same TFBS¹⁶⁷ and may be capable of binding to the TFBSs in the $4 \times [bZIP3\text{-random}]$ synthetic promoter, competing with bZIP3. This hypothesis could be tested by either testing the ability of other bZIP family proteins to bind to the $4 \times [bZIP3\text{-random}]$ promoter or measuring the effects of co-expressing multiple proteins on expression in transient assays.

In the future it would be interesting to do time course analyses to determine if the fold-change in luminescence from the synthetic promoters changes over time after nitrate treatment. It would also be interesting to test the synthetic promoters in mutant lines where the TF of interest is knocked out to determine if the synthetic promoters are still induced by nitrate. If they still responded to nitrate, this would suggest that other TFs are capable of binding to the TFBSs in the synthetic promoters. It would also be interesting to test the synthetic promoters in other plant species to determine if they are also induced by nitrate.

6.5.4 Towards the development of synthetic genetic feedback controllers

The final aim of this chapter was to build genetic feedback controllers composed of minimal synthetic N-responsive promoters controlling synthetic CRISPR TFs which modulate the expression of *ARF18*. This was to introduce positive and negative genetic feedback loops to the N-response subnetwork from genes further down the transcriptional cascade, which would allow for the control of *ARF18* expression in response to N-availability. These expression changes could potentially increase the robustness of the network⁷ and/or lead to oscillations⁹ in the case of negative feedback or could lead to bistability in the case of positive feedback¹⁰.

To achieve this, an NLP7-responsive synthetic promoter ($4 \times [NLP7-TATA] + spacing$) was assembled with either synthetic activators (SunTag-2 \times TAD and SunTag-VP64 each with three sgRNAs) or synthetic repressors (dCas12a::NLS::3 \times SRDX with three crRNAs). Lines containing these constructs were produced, with the expectation that the feedback controllers would add positive or negative feedback, respectively, into the network. Control lines in which the synthetic activator/repressor are constitutively expressed were also produced.

In addition, equivalent constructs were assembled and transformed into stable lines in which expression of synthetic TFs is controlled by the known N-responsive promoter, *NRP*, as well as $4 \times [bZIP3-random]$, both of which were also shown to respond to nitrate. Since *NRP* was shown to respond to NLP7, this construct can also be considered to be a feedback controller. Additional work is needed to determine the position of bZIP3 in the N-response network. However, the expression of bZIP3 is known to change in response to nitrate^{318,524}. Therefore, coupling the expression of *ARF18* to bZIP3 expression might lead to a similar, albeit weaker, response to nitrate treatment as the NLP7-controlled feedback controllers. This response is expected to be slightly slower, since bZIP3 responds within ten minutes of nitrate treatment³¹⁸, whereas NLP7 responds post-translationally within three minutes of nitrate treatment⁴⁰.

To date, these plant lines have not yet been analysed. Measuring changes in the expression of *ARF18* and other genes in the N-response subnetwork, as well as N-induced changes to root system architecture phenotypes, will allow these hypotheses to be tested. Based on the data obtained from protoplast co-expression assays, a range of transcriptional effector

strengths are to be expected while testing in whole plants, with *NRP* controlled feedback controllers causing the weakest regulation of *ARF18*, and effectors controlled by $4\times[NLP7-TATA]+spacing$ causing the largest change in expression of *ARF18*. If stronger activation of *ARF18* is needed for observable phenotypes, additional transcriptional activators could be tested such as the recently reported CRISPR-Act3.0 system³⁰¹. If stronger repression of *ARF18* is required, different repressor domains such as DLN144, DLS and MIX domains, which have recently proved effective in *N. benthamiana*, could be tested⁵²⁷. DLN144, a DLN hexapeptide motif from ZOS5-09 in rice, was 40% more effective than the SRDX domain at repressing in yeast⁵²⁸. Combinations or multiple copies of repressor domains could also be tested to potentially increase repression strength.

6.5.5 Conclusions

The aim of building and testing synthetic transcriptional activators and repressors was accomplished. Further experiments are needed to test the performance of these synthetic transcriptional effectors in whole plants. This chapter shows that synthetic nitrate responsive promoters can be engineered using knowledge of TFBSs, fulfilling the second aim. These promoters can be designed to be activated or repressed by one or more TFs. Transient protoplast co-expression assays are useful for rapidly assessing the performance of synthetic promoters including their responsiveness to cognate TFs. These protoplast assays are also useful for testing the performance of synthetic transcriptional activators and repressors. However, due to the disrupted chromatin state in protoplasts, the results may not reflect the behaviour of the synthetic TFs in whole plants, so further experiments are needed to test whether these synthetic TFs are functional on endogenous genes. The evaluations of lines in which these elements are constitutively expressed will enable this. The $4\times[NLP7-TATA]+spacing$ and $4\times[bZIP3-random]$ promoters responded to increasing nitrate in root luciferase assays, so, provided that the synthetic TFs are functional, it is likely that the genetic feedback controllers will be functional in whole plants. The genetic feedback controllers were assembled and transformed into stable lines, fulfilling the third aim. These lines will be tested in the future to determine whether they can be used to tune the response of *ARF18* to nitrate, and potentially alter NUE.

Chapter 7

Discussion

The predictable engineering of biological systems either to tune existing functions such as the NUE of crops, or engineer new functions, is one of the aims of synthetic biology. The engineering of regulatory elements and the introduction of synthetic genetic circuits into crops are powerful approaches to achieve this goal. However, to be able to reliably predict the behaviour of either changes to existing sequences or of new circuits, it is necessary to understand the underlying biological mechanisms. The overall aim of this thesis was to identify approaches and technologies for predictable engineering of plant GRNs using a subnetwork that regulates plant responses to nitrate. To this aim, a N-response subnetwork of TFs that were predicted to cross-regulate each other was identified and characterised²⁵. These TFs directly or indirectly regulated many downstream genes involved in nitrate metabolism and growth. In chapter 3, edges in this subnetwork were characterised by the identification of promoter motifs that indicated likely interactions between TFs and target promoters, and by using protoplast co-expression assays, to determine whether TFs activate or repress their targets. 29 edges between eight TFs and the *NIR1* gene were found to significantly alter the expression of target genes, which supported their direct regulation when integrated with TRAM data, which measured the relative binding affinity of TFs to sites in the promoters of target genes. These data will be useful for understanding the overall GRN and for predicting the effects of edge disruption on the expression of other genes in the subnetwork. These data will also be useful for testing and identifying network motifs. One such motif was a feedforward loop in which edges from ARF18 to *NLP7* via ANAC032, and ARF18 to *DREB26* were supported. An edge from DREB26

to *NLP7* could not be tested using co-expression assays because the candidate TFBS for DREB26 was in the coding region of *NLP7*. To date, most published GRNs only provide partial support for regulatory interactions since they only integrate one or two types of data, such as TARSET or Y1H data. Combining proof of direct regulation with TRAM and confirming the direction of regulation with protoplast co-expression data provides a more complete picture of the GRN, and similar assays could be used to expand the number of tested edges in the wider N-responsive GRN. Once edges in GRNs are elucidated, it will be possible to make predictions about how they might be manipulated to alter the expression of downstream genes and improve beneficial traits such as NUE. For example, protoplast co-expression data from chapter 3 was used in chapter 6 to design and engineer synthetic genetic feedback controllers that are expected to alter the expression of *ARF18* in response to nitrate, the behaviour and dynamics of the network and, potentially, the growth of plants in response to nitrate.

Perturbing GRNs by disrupting edges within the network is another approach to investigate network function, including the roles of network motifs. Individual edges could be disrupted using genome engineering tools to introduce mutations that disrupt individual binding motifs. However, this is likely to be laborious especially as it is difficult to identify individual TFBSs that are important for regulatory function and a given TF often binds to multiple sites within a target promoter. Ideally, this approach also requires knowledge of all TFs and other proteins that bind to the target promoters. Although progress was made at characterising the N-response subnetwork, the dataset is still likely to be incomplete. An alternative way to disrupt edges in GRNs is to use CRISPR library approaches to add variation into upstream regulatory regions, bypassing the need to understand all edges targeting that promoter. This was achieved in chapter 4 in which a pooled CRISPR library was constructed targeting the promoters/5' UTRs of *ARF9/18*, *DREB26* and *NLP7*. 96 multigene Cas9 expression constructs targeting these genes were delivered to plants, and 500 T2 plants were sequenced using a pooled barcode amplicon sequencing approach to identify mutations. A range of interesting mutations were found in the CRISPR library, some of which were shown to disrupt the regulation of four TFs involved in the early response to nitrate. Five lines from the resulting library of 327 mutants were analysed by assessing RSA and gene expression using qRT-PCR. Mutations in the *NLP7* promoter disrupted potential edges from ANAC032, ANR1 and *NLP7*, causing

longer and more dense lateral roots. This potentially disrupted the feedforward loop from ARF18 to *NLP7* via ANAC032, altering network dynamics and reducing the robustness of the nitrogen-response subnetwork, leading to the RSA changes. A 5 bp deletion in the *DREB26* 5' UTR caused a similar phenotype, increasing the number and density of lateral roots, potentially due to disruption of the putative feedforward loop from ARF18 to *NLP7* through DREB26. Deletions in the *ARF9* and *ARF18* *cis*-regulatory regions slightly increased total root length, potentially by reducing the overall repression effect of ARF9/18 repressors on genes in the nitrate-response subnetwork in response to auxin. It is undesirable to generate loss-of-function mutants of these and many other TFs because they may have roles in other tissues and processes. For example, ARF18 has a role in shade avoidance in aerial tissues⁴⁰⁷. Despite this, relatively few studies have targeted *cis*-regulatory regions using pooled CRISPR library approaches^{306,307,449}. This demonstrates that CRISPR libraries targeting *cis*-regulatory regions are a relatively untapped resource and may be useful tools for engineering crop traits such as NUE.

In chapter 5, promoter architecture was compared between four gene categories: constitutively expressed, variably expressed, non-specific and tissue-specific genes. Distinctive features were found that may be essential for enabling their expression patterns. For example, constitutively expressed and non-specific genes contained more open chromatin in CRMs than variably expressed and tissue-specific genes. There was also an increased GC content and a lower percentage bp covered by TFBSs in constitutively expressed and non-specific genes compared to variably expressed and tissue-specific genes. Variably expressed genes were enriched for TATA boxes. Useful design features were extracted from these data, which were used in chapter 6 to aid the design of synthetic minimal N-responsive promoters. Similar promoter architecture analyses could be used in other plant species, including crops, to learn features for the design of synthetic promoters for engineering useful traits.

Data from chapter 3 which elucidated the structure of the N-response network were also used to inform the design and engineering of a genetic feedback controller (chapter 6). To construct this controller, promoter design features from chapter 5 were integrated into the design of minimal synthetic promoters that were then shown to respond to their cognate TFs in co-expression assays. Expression from these minimal synthetic promoters was also

activated by nitrate in transgenic plants. To complete the genetic feedback controller, synthetic promoters were coupled to Cas-based programmable transcriptional regulators. Protoplast co-expression assays were also useful for testing the activity of these synthetic activators and repressors on the expression of *ARF18*. The $4 \times [NLP7-TATA]+spacing$ and $4 \times [bZIP3-random]$ promoters responded to nitrate. It is therefore likely that the genetic feedback controllers will be functional. However, it remains to be seen how effective these synthetic genetic feedback controllers will be in changing the plant response to nitrate. If proven effective, equivalents of these controllers as well as additional genetic feedback loops could be engineered into crop species to improve NUE. Synthetic circuits using minimal synthetic promoters controlling synthetic activators and repressors were recently demonstrated in *Arabidopsis* and *N. benthamiana*, and were able to control gene expression across tissues and could successfully reprogram plant root growth²⁵⁹.

Network engineering is a powerful approach for engineering desirable plant traits using specific modifications to the GRN. It can be used to engineer quantitative traits that are difficult to improve using conventional breeding methods, such as yield and nutrient content⁵²⁹. Additionally, engineering of non-coding regions such as promoters can be used to improve traits without disrupting the function of pleiotropic proteins¹¹². Precision genome editing, where precise changes are made at specific loci without incorporating DNA into the genome, can be used to tune the expression of genes in GRNs and is often faster and less labour-intensive than conventional breeding methods⁴²³⁻⁴²⁵. However, network engineering is limited by the current understanding of the GRN. If the GRN is not well understood, it is difficult to predict the outcome of modifications. This is why the work in this thesis is important, as it elucidates the structure of the N-response network and provides tools for engineering it. When not enough is known about the GRN, imprecise forward genetic screens such as physical or chemical mutagenesis and selection, marker assisted breeding, transgenesis, and T-DNA insertion can be used to improve traits. The use of newer technologies such as CRISPR screens can limit the addition of genetic variation to specific genomic regions of interest, which can reduce the number of lines that need to be screened to identify useful mutations compared to other forward genetic screens. Another limitation of network engineering is that not all connections in the GRN are expected to be orthologous in related crop species. For example, in *A. thaliana* ARF18 regulates *NLP7* via ANAC032 whereas in tomato ARF18 regulates *NLP7* directly and through

DREB26 as *ANAC032* is not present in tomato⁵³⁰. Therefore, network engineering in *Arabidopsis* may not be directly applicable to other species. However, the work in this thesis provides a framework for elucidating the structure of GRNs in other species, which can then be used to engineer them. It would be useful to obtain further datasets to facilitate the engineering of GRNs in *A. thaliana* and other species. For example, a more complete DAP-seq dataset containing all TFs would enable the more accurate prediction and testing of additional TF-gene interactions. Extending protoplast co-expression assays and TARSET to additional TFs and target genes would also be useful for elucidating the structure of GRNs. Studying GRNs in specific cell types using single cell technologies such as single cell RNA-seq (scRNA-seq) and single cell ATAC-seq (scATAC-seq) would allow for more fine-grained network engineering in cells and tissues of interest. Once enough data is collected, mathematical models can be used to predict the behaviour of GRNs and guide the design of specific edits in the genome, or of orthogonal parts such genetic feedback controllers to add new edges and functionality into the network. Going forward, it will be important to validate the function of the genetic feedback controllers described in chapter 6 *in planta*. Importantly, phenotypes such as root growth in response to nitrate should be measured in the presence of nitrate to determine if the genetic feedback controllers are functional, and any observed changes in NUE will determine if they can improve the plant response to nitrate. The genetic feedback controllers should be tested and modified to work in other species to determine if they are functional in other crops. Simultaneously, the CRISPR library should be expanded to other genes in the N-response network to determine which additional genes can be targeted to improve the plant response to nitrate. Mutations in promoters of genes which have a change in phenotype in response to nitrate should be identified and characterised to determine if they can be used to improve the plant response to nitrate. Lines containing these mutations could be crossed to test if they have an additive effect on the phenotype.

This thesis demonstrates that a greater understanding of gene regulatory sequences can be used to inform the engineering of quantitative traits such as root growth in response to nitrate. The overall aim of identifying approaches and technologies for predictable engineering of plant GRNs using a subnetwork that regulate plant responses to nitrate was achieved. *Arabidopsis* N-responses were manipulated using a CRISPR library approach and synthetic feedback controllers were designed. These engineering strategies were facilitated

by the identification and characterisation of interactions between TFs and target genes in a N-response network and by the identification of architectural features enriched in genes with condition-responsive patterns of expression. These approaches and tools are widely applicable to elucidating and engineering other plant GRNs.

Appendix A: Supplementary data for Chapter 2

This appendix contains all supplementary materials for chapter [2](#).

File S1 - Table of plasmids used in this thesis (electronic supplementary materials).

File S2 - Table of primers used in this thesis (electronic supplementary materials).

Appendix B: Supplementary data for Chapter 4

This appendix contains all supplementary materials for chapter 4.

Figure S1 - Location of mutation relative to transcription start site in line 69-9

Figure S2 - Location of mutation relative to transcription start site in line 130-4

Figure S3 - Location of mutation relative to transcription start site in line 134-3

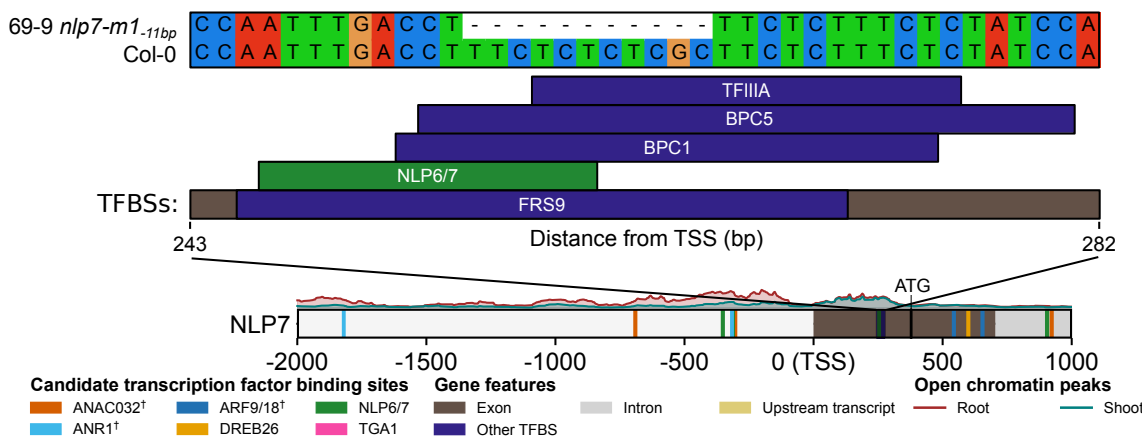


Figure S1: **Location of mutation relative to transcription start site in line 69-9.** All overlapping TFBSs scanned using FIMO³⁶⁴ using DAP-seq motifs¹⁸¹ are shown (purple).

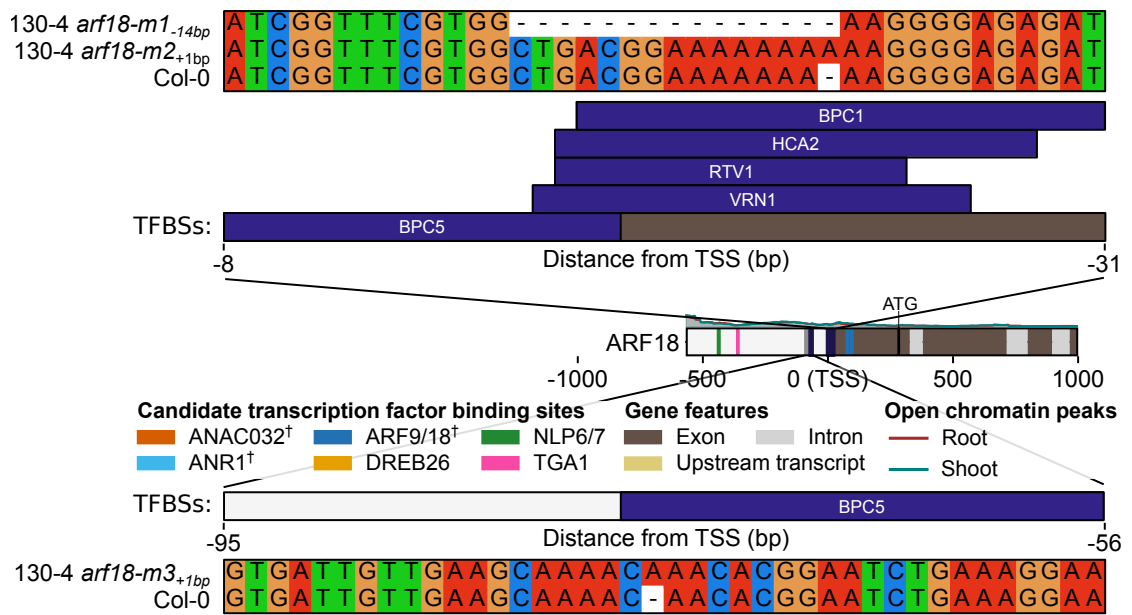


Figure S2: **Location of mutation relative to transcription start site in line 130-4.** All overlapping TFBSs scanned using FIMO³⁶⁴ using DAP-seq motifs¹⁸¹ are shown (purple).

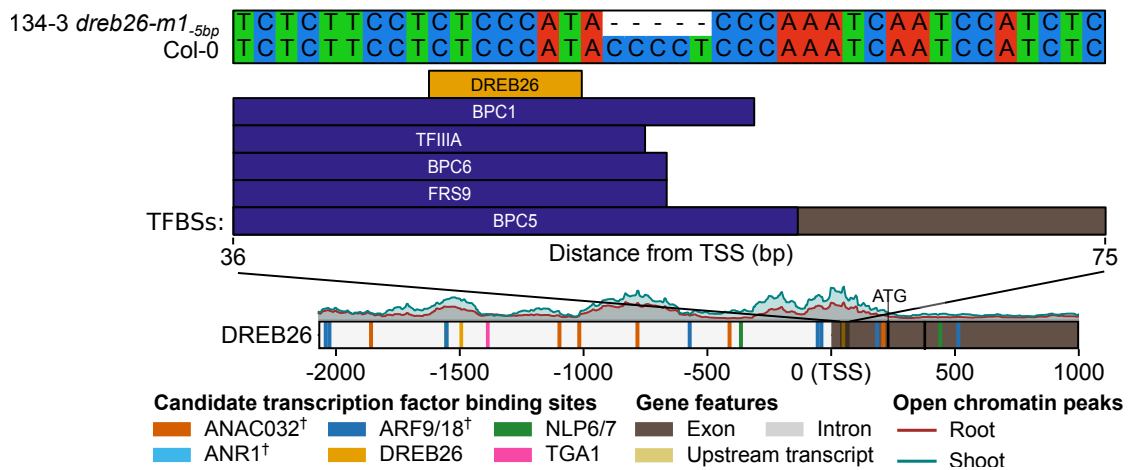


Figure S3: **Location of mutation relative to transcription start site in line 134-3.** All overlapping TFBSs scanned using FIMO³⁶⁴ using DAP-seq motifs¹⁸¹ are shown (purple).

Appendix C: Supplementary data for Chapter 5

This appendix contains all supplementary materials for chapter 5.

Figure S4 - Mean percentage open chromatin in Arabidopsis CRM expression categories

Figure S5 - Sliding window analysis of percentage open chromatin in Arabidopsis CRMs

Figure S6 - Percentage open chromatin in Arabidopsis CRMs centred around the transcription start site (TSS)

Figure S7 - Mean percentage GC content in Arabidopsis CRMs

Figure S8 - Sliding window analysis of percentage GC content in Arabidopsis CRMs

Figure S9 - Mean percentage GC content of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs

Figure S10 - Mean percentage TFBS coverage of Arabidopsis CRMs

Figure S11 - Mean percentage coverage of TFBS within open chromatin regions of Arabidopsis CRMs

Figure S12 - Sliding window analysis of percentage TFBS coverage in Arabidopsis CRMs

Figure S13 - Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs

Figure S14 - Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions

Figure S15 - Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted

to bind to Arabidopsis CRMs

Figure S16 - Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs within open chromatin regions

Figure S17 - Sliding window analysis of Shannon diversity of individual TFs predicted to bind Arabidopsis CRMs

Figure S18 - Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs

Figure S19 - Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions

Figure S20 - \log_2 -fold enrichment of 15 bp TATA boxes in Arabidopsis CRMs

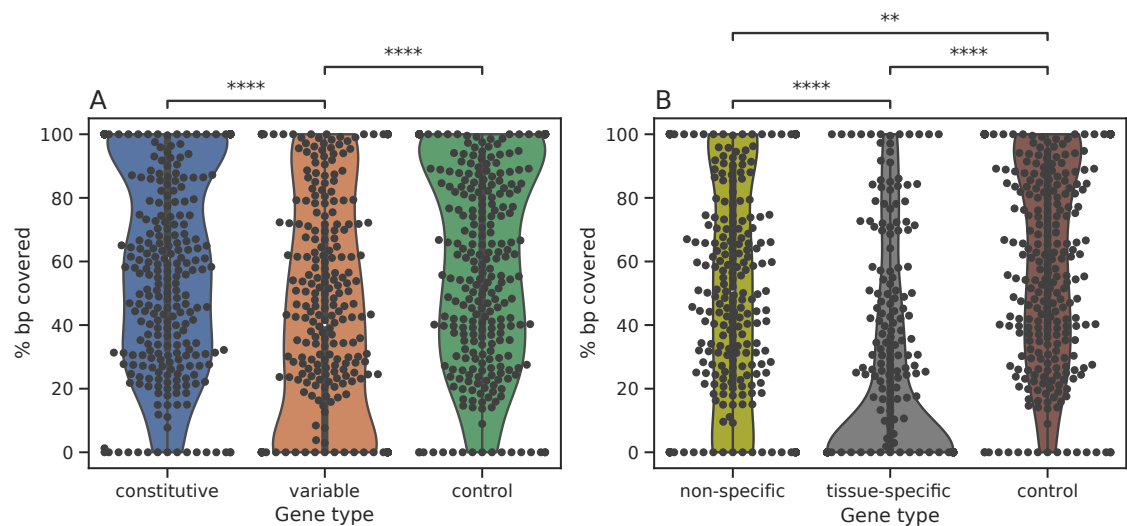


Figure S4: Mean percentage open chromatin in Arabidopsis CRM expression categories. A: Percentage base pairs (bp) within open chromatin in 300 constitutive, variable and control genes. $N=300$. B: Percentage bp within open chromatin in 300 non-specific, tissue-specific and control genes. $N=300$. Significance was calculated using Kruskal-Wallis and Dunn's post hoc tests with Bonferroni correction. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. Control genes are identical in plots A and B.

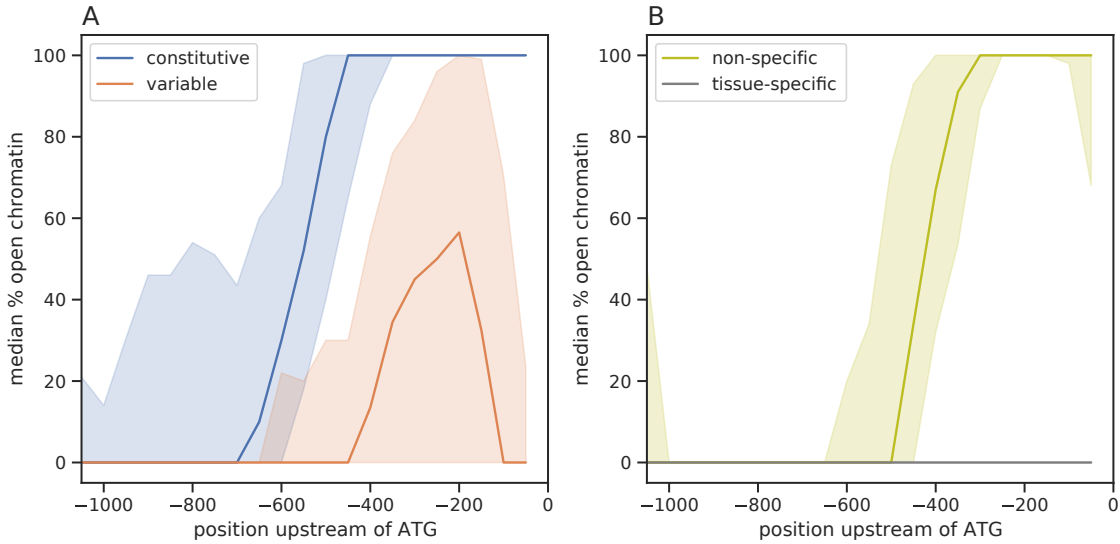


Figure S5: **Sliding window analysis of percentage open chromatin in Arabidopsis CRMs.** The median proportion of nucleotides within open chromatin was scored for 100 bp windows across selected CRMs. The windows were overlapping with a 50 bp offset. The median value for each bin is displayed at the central bp of the bin. Shading represents 95 % confidence intervals estimated using 10000 bootstraps. A: constitutive and variable GC content sliding windows. $N=300$. B: non-specific and tissue-specific sliding windows. $N=300$.

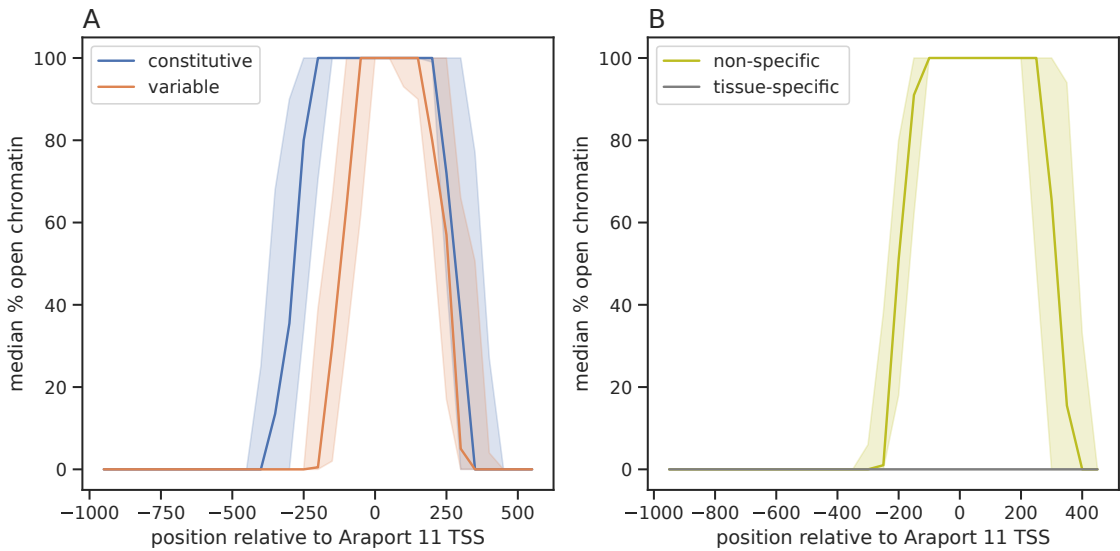


Figure S6: **Percentage open chromatin in Arabidopsis CRMs centred around the transcription start site (TSS).** Windows are offset by 50 bp. Shading represents 95 confidence intervals estimated using 10000 bootstraps. Median percentage of open chromatin peaks overlapping 100 bp windows. Open chromatin peaks derived from the intersect of root and shoot peaks derived from negative control (treated with NaOH) ATAC-seq data by Potter et al.³⁵⁰. A: constitutive and variable GC content sliding windows. $N=300$. B: non-specific and tissue-specific sliding windows. $N=300$.

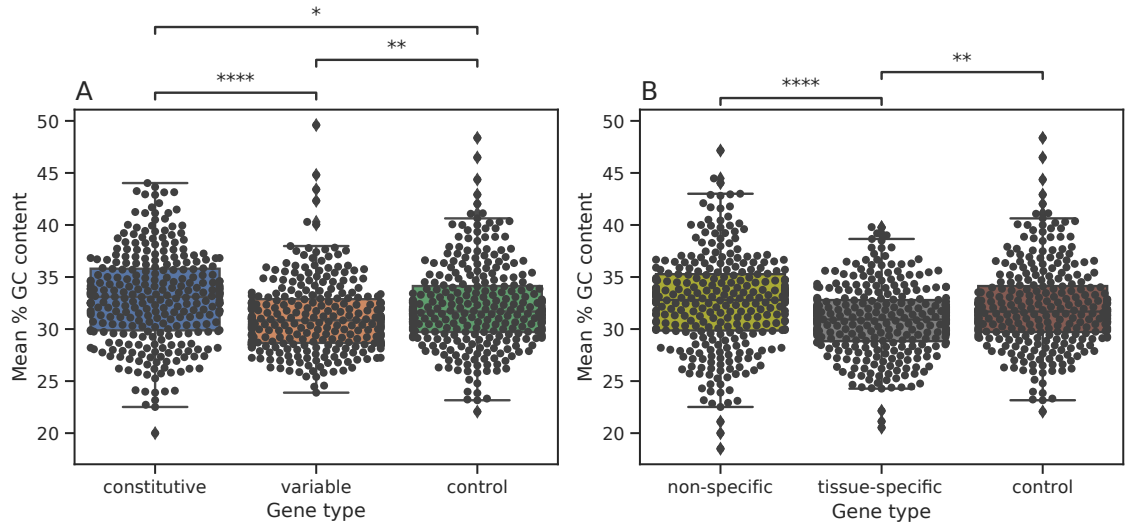


Figure S7: **Mean percentage GC content in Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. A: Percentage GC content in 300 constitutive, variable and control genes. B: Percentage GC content in 300 non-specific, tissue-specific and control genes. There was no significant difference in GC content between categories. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. $N=300$. Control genes are identical in plots A and B.

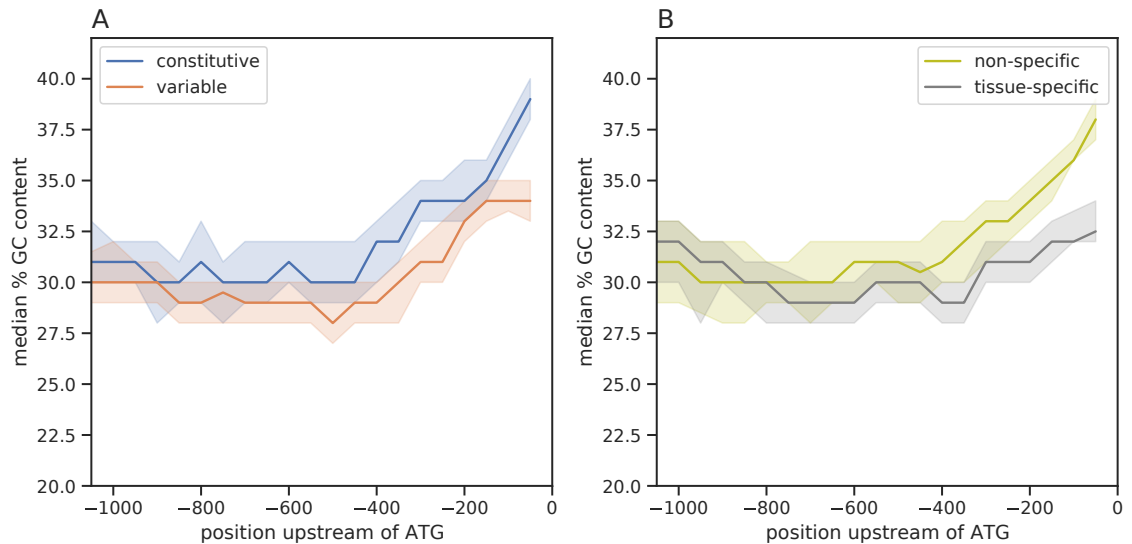


Figure S8: **Sliding window analysis of percentage GC content in Arabidopsis CRMs.** The median percentage GC content was scored for 100 bp windows across selected CRMs. The windows were overlapping with a 50 bp offset. The median value for each bin is displayed here at the central bp of the bin. Shading represents 95 confidence intervals estimated using 10000 bootstraps. A: constitutive and variable GC content sliding windows. $N=300$. B: non-specific and tissue-specific sliding windows. $N=300$.

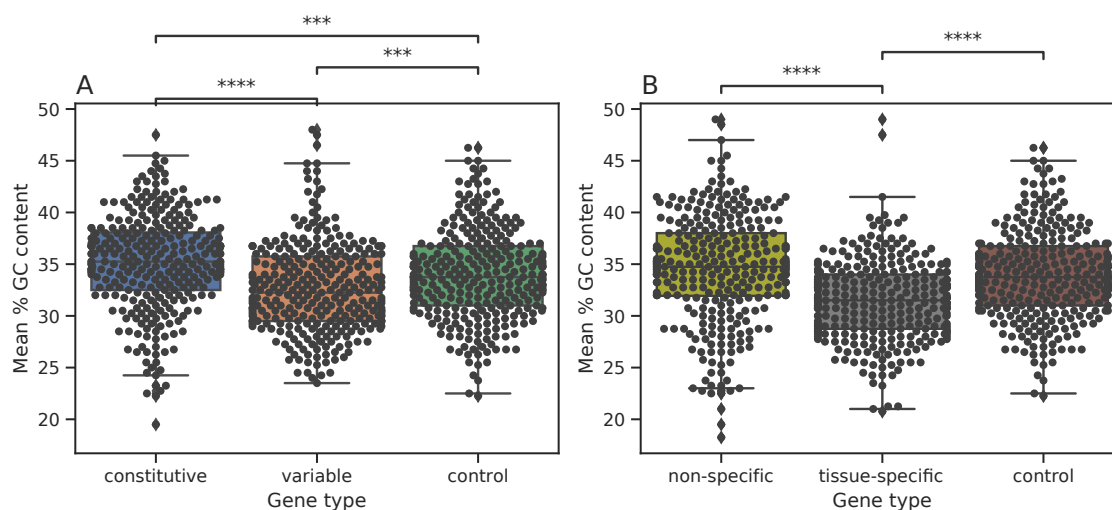


Figure S9: **Mean percentage GC content of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Significance was calculated using Kruskal-Wallis followed by Dunn's post hoc tests with Bonferroni correction. A: Percentage GC content in 300 constitutive, variable and control CRMs. $N=300$. B: Percentage GC content in 300 non-specific, tissue-specific and control genes. $N=300$. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. Control genes are identical in plots A and B.

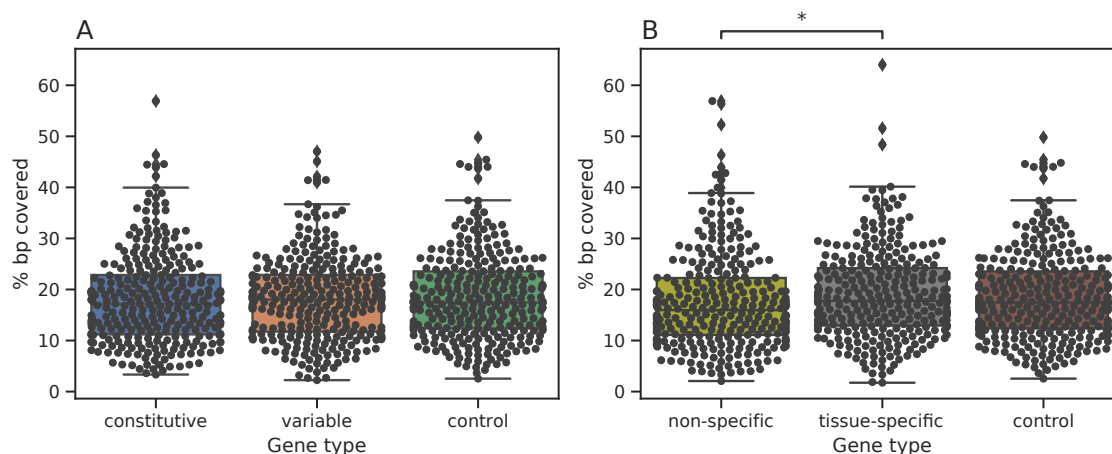


Figure S10: **Mean percentage TFBS coverage of Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Significance was calculated using Kruskal-Wallis. A: There was no significant difference in TFBS coverage between constitutive, variable or control CRMs. $N=300$. B: There was no significant difference in TFBS coverage between non-specific, tissue-specific or control CRMs. *, $P < 0.05$. $N=300$. Control genes are identical in plots A and B.

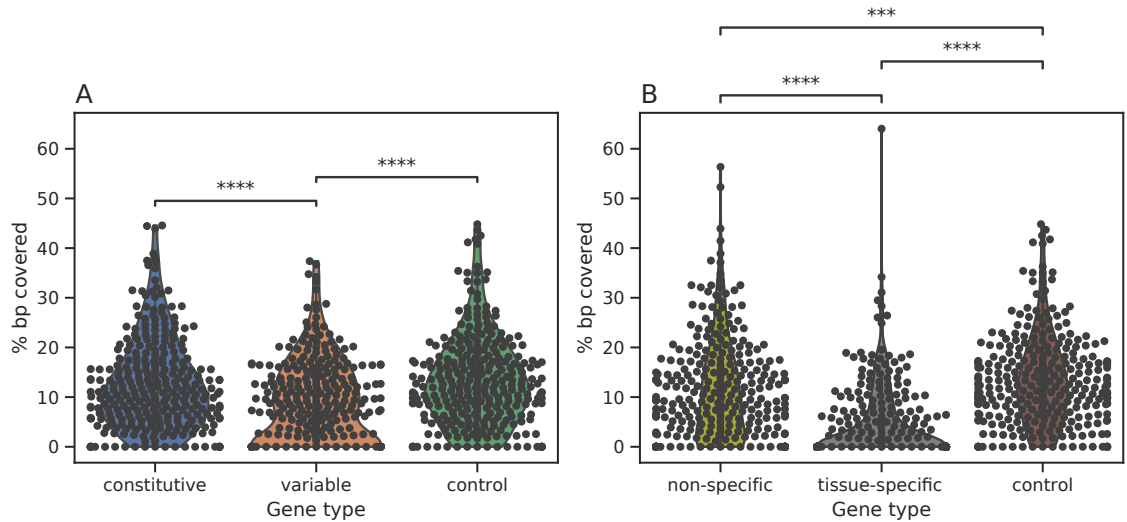


Figure S11: **Mean percentage coverage of TFBS within open chromatin regions of Arabidopsis CRMs.** Significance was calculated using Kruskal-Wallis followed by Dunn's post hoc tests with Bonferroni correction. A: Mean percentage coverage of TFBS falling within open chromatin in constitutive, variable and control CRMs. $N=100$. B: Mean percentage coverage of TFBS falling within open chromatin in non-specific, tissue-specific and control CRMs. $N=100$. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. Control genes are identical in plots A and B.

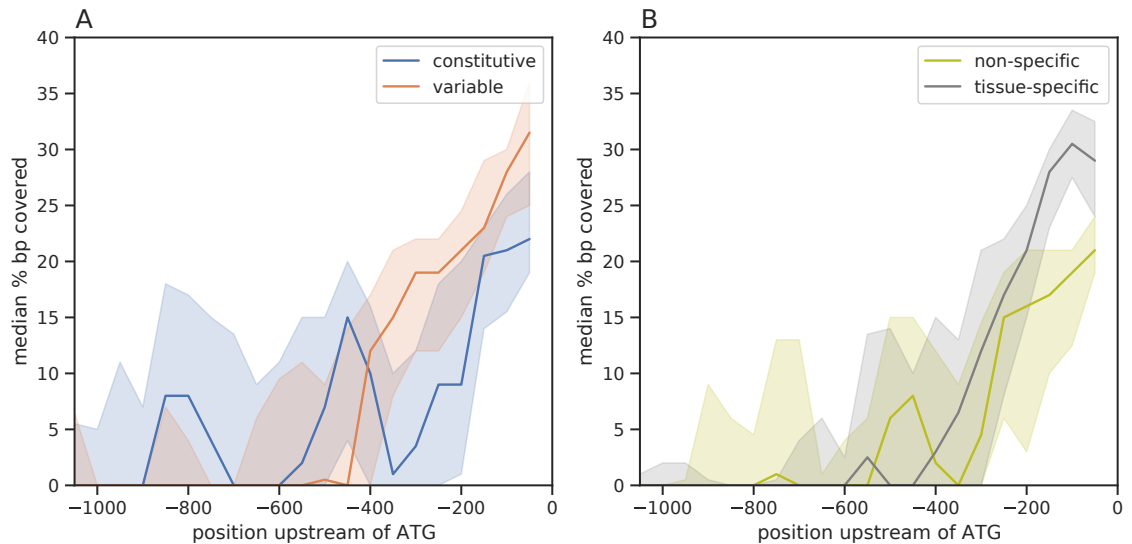


Figure S12: **Sliding window analysis of percentage TFBS coverage in Arabidopsis CRMs.** The median percentage TFBS coverage was scored for 100 bp windows across selected CRMs. The windows were overlapping with a 50 bp offset. The median value for each bin is displayed here at the central bp of the bin. Shading represents 95% confidence intervals estimated using 10000 bootstraps. A: constitutive and variable GC content sliding windows. $N=300$. B: non-specific and tissue-specific sliding windows. $N=300$.

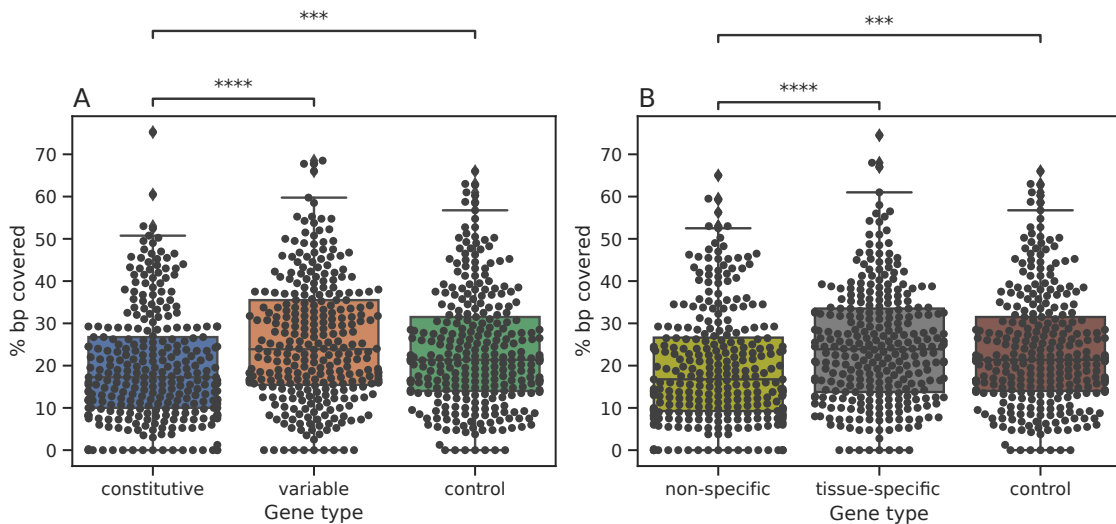


Figure S13: Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs. Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. A: Mean percentage TFBS coverage in constitutive, variable and control CRMs. $N=300$. B: Mean percentage TFBS coverage in non-specific, tissue-specific and control CRMs. $N=300$. Significance was calculated using Kruskal-Wallis and Dunn's post hoc tests with Bonferroni correction. **, $P < 0.01$. *, $P < 0.05$.

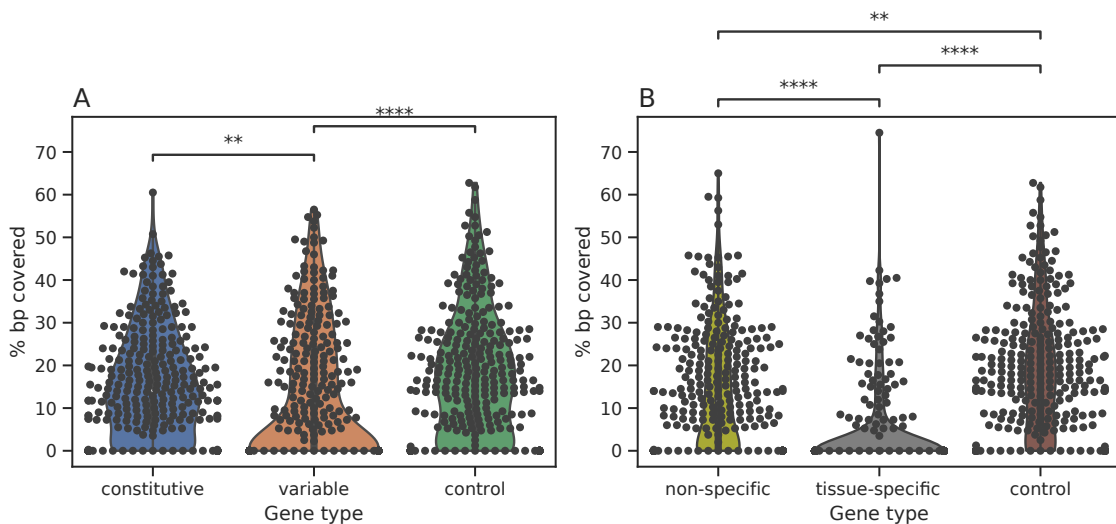


Figure S14: Mean percentage TFBS coverage of the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions. A: Mean percentage coverage of TFBS falling within open chromatin in 400 bp constitutive, variable and control CRMs. B: Mean percentage coverage of TFBS falling within open chromatin in 400 bp non-specific, tissue-specific and control CRMs. ****, $P < 0.0001$. ***, $P < 0.001$. **, $P < 0.01$. *, $P < 0.05$. $N=300$. Significance was calculated using Kruskal-Wallis and Dunn's post hoc tests with Bonferroni correction.

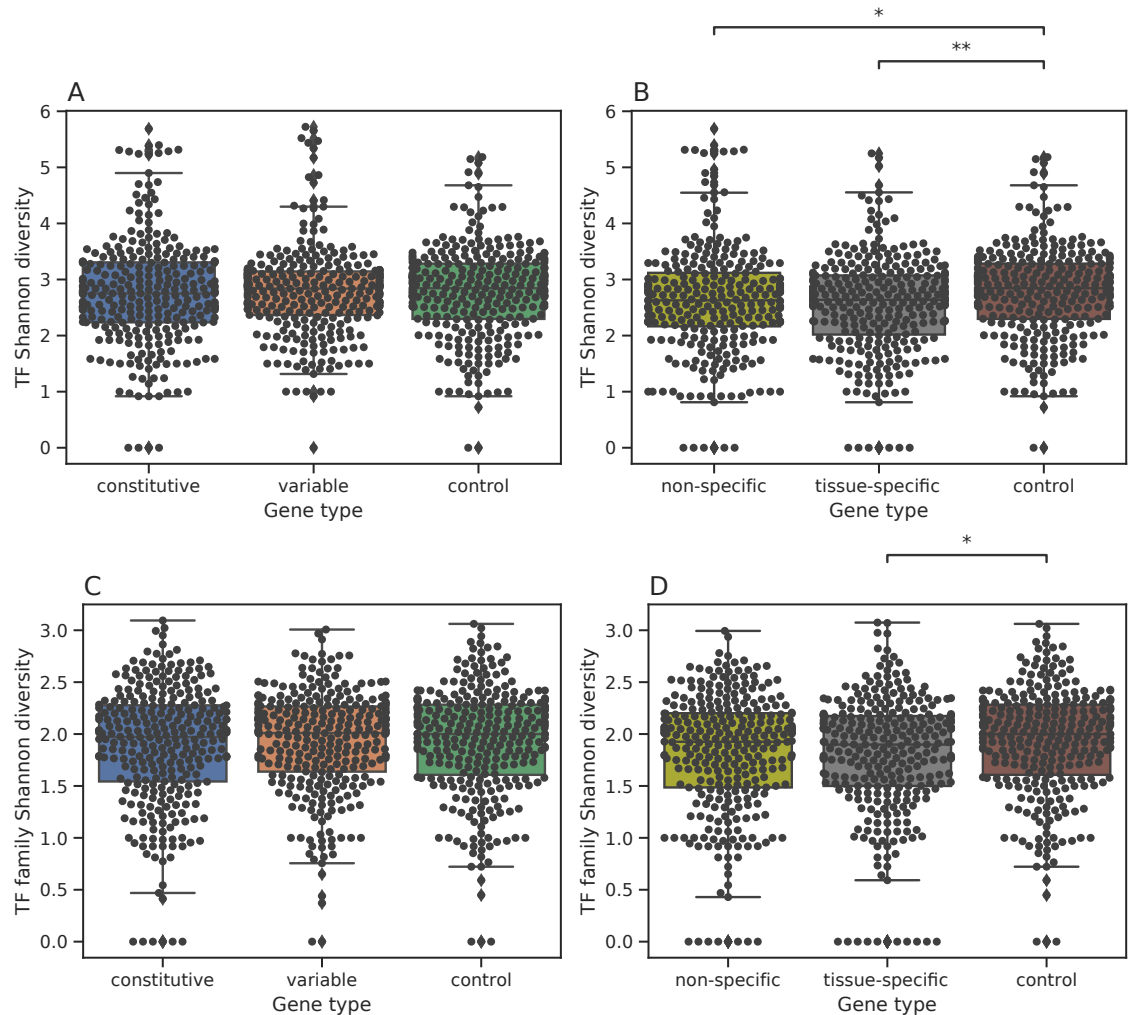


Figure S15: **Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. A: Shannon diversity of individual TFs in 100 constitutive, 300 variable and 300 control genes. B: Shannon diversity of individual TFs in 300 non-specific, 300 tissue-specific and 300 control genes. C: Shannon diversity of TF families in 300 constitutive, 300 variable and 300 control genes. D: Shannon diversity of TF families in 300 non-specific, 300 tissue-specific and 300 control genes. **, $P < 0.01$. *, $P < 0.05$.

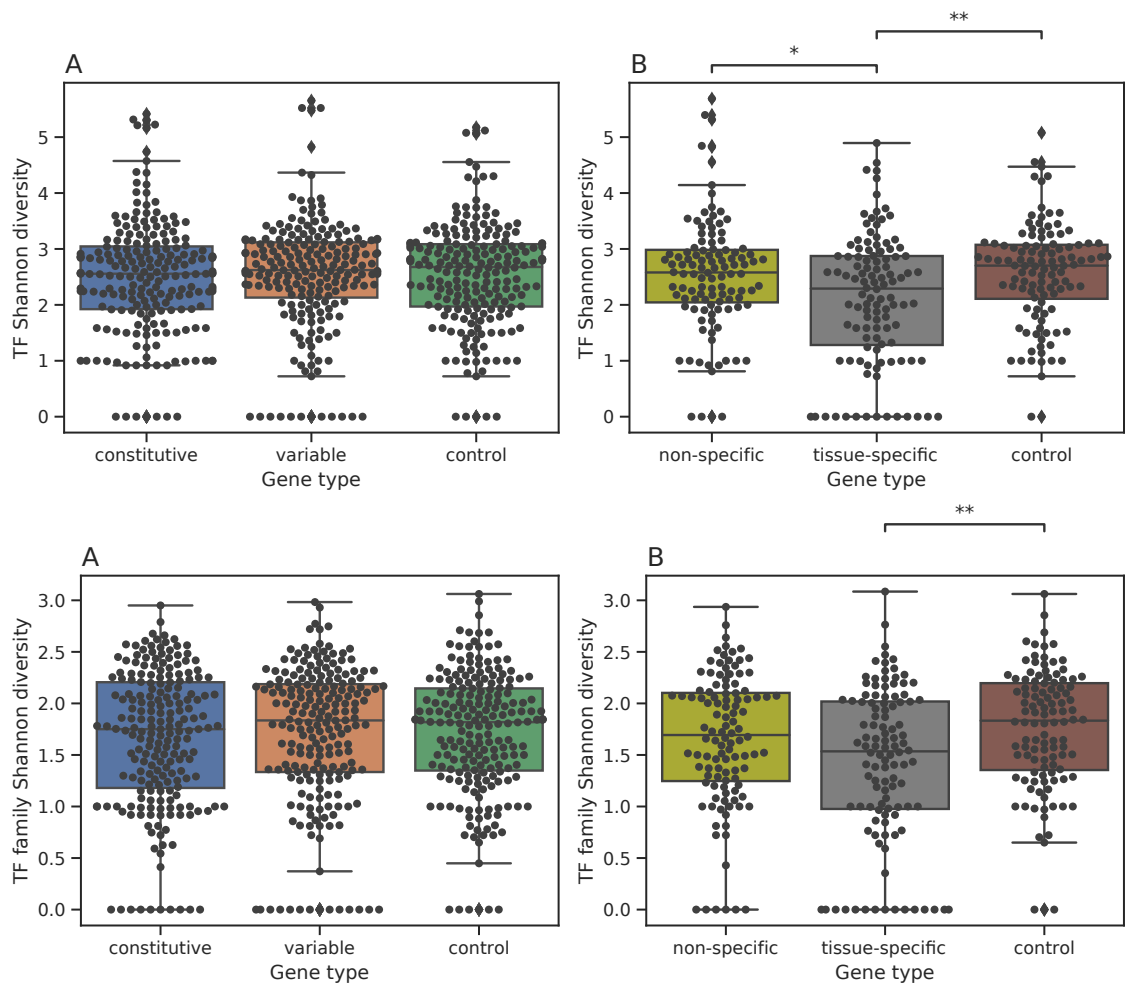


Figure S16: **Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to Arabidopsis CRMs within open chromatin regions.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. A: Shannon diversity of individual TFs in 100 constitutive, 100 variable and 100 control genes. B: Shannon diversity of individual TFs in 100 non-specific, 100 tissue-specific and 100 control genes. C: Shannon diversity of TF families in 100 constitutive, 100 variable and 100 control genes. D: Shannon diversity of TF families in 100 non-specific, 100 tissue-specific and 100 control genes. **, $P < 0.01$. *, $P < 0.05$.

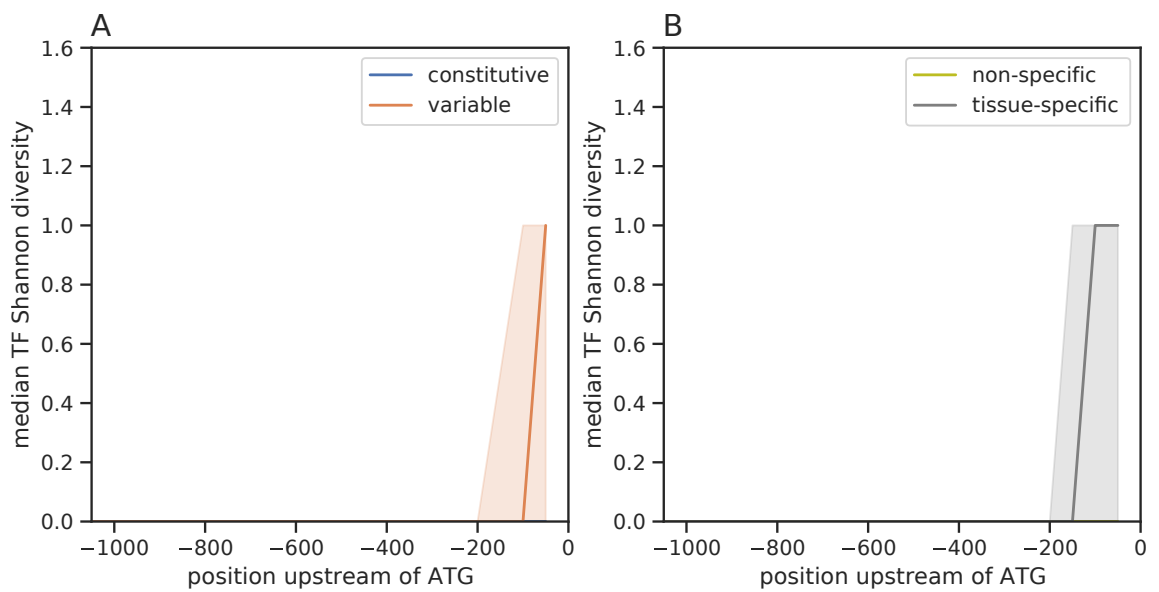


Figure S17: **Sliding window analysis of Shannon diversity of individual TFs predicted to bind Arabidopsis CRMs.** Data points are positioned in the centre of each 100 bp window. Windows are offset by 50 bp. Shading represents 95 confidence intervals estimated using 10000 bootstraps. Median Shannon diversity of individual TFs sliding windows in A: constitutive and variable, and B: non-specific and tissue-specific CRMs. $N=300$,

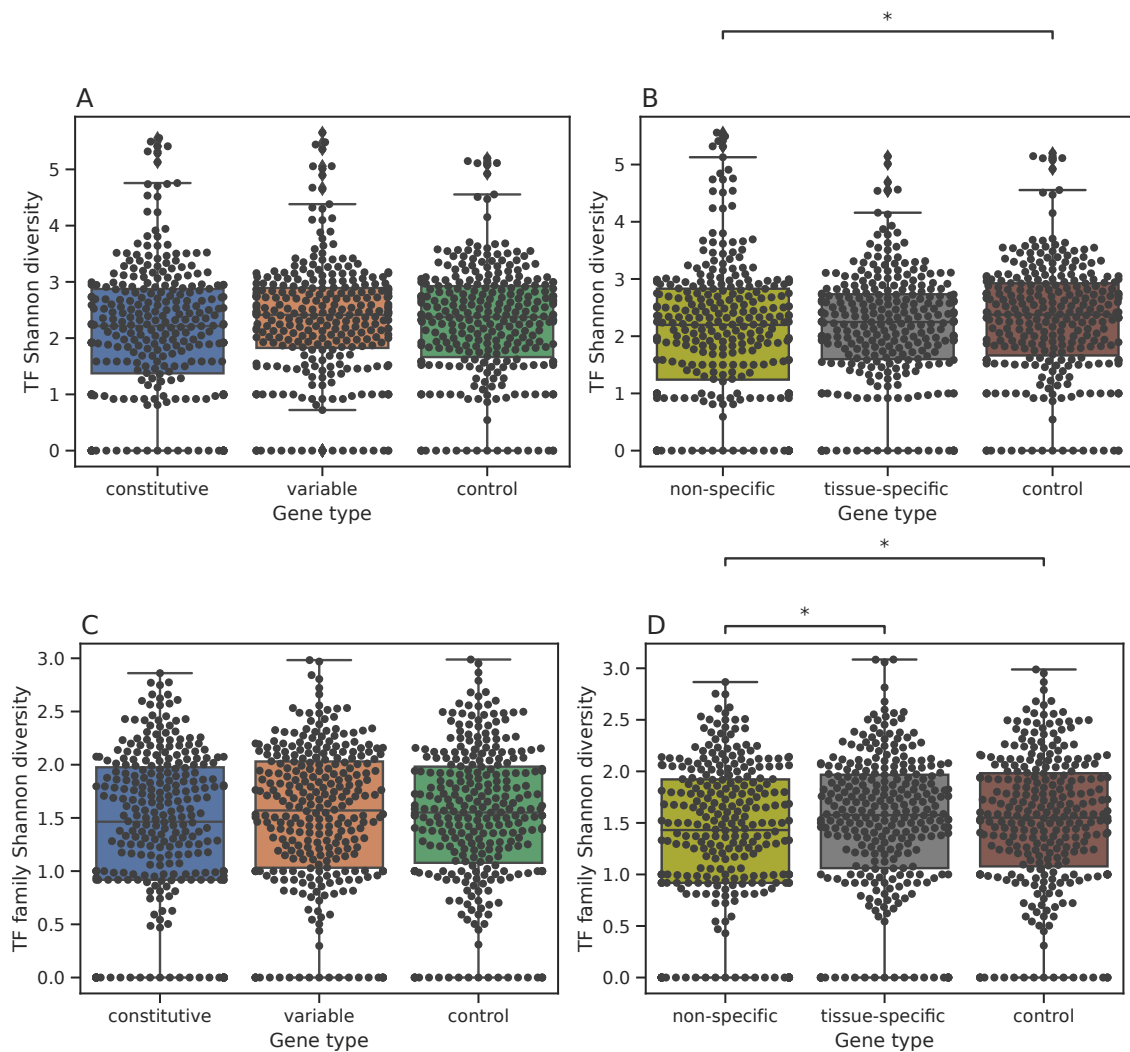


Figure S18: **Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to the 400 nucleotides upstream of the ATG start codon in *Arabidopsis* CRMs.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Shannon diversity of individual TFs in A: 300 constitutive, 300 variable and 300 control genes and B: 300 non-specific, 300 tissue-specific and 300 control genes. Shannon diversity of TF families in C: 300 constitutive, 300 variable and 300 control genes and D: 300 non-specific, 300 tissue-specific and 300 control genes. *, $P < 0.05$.

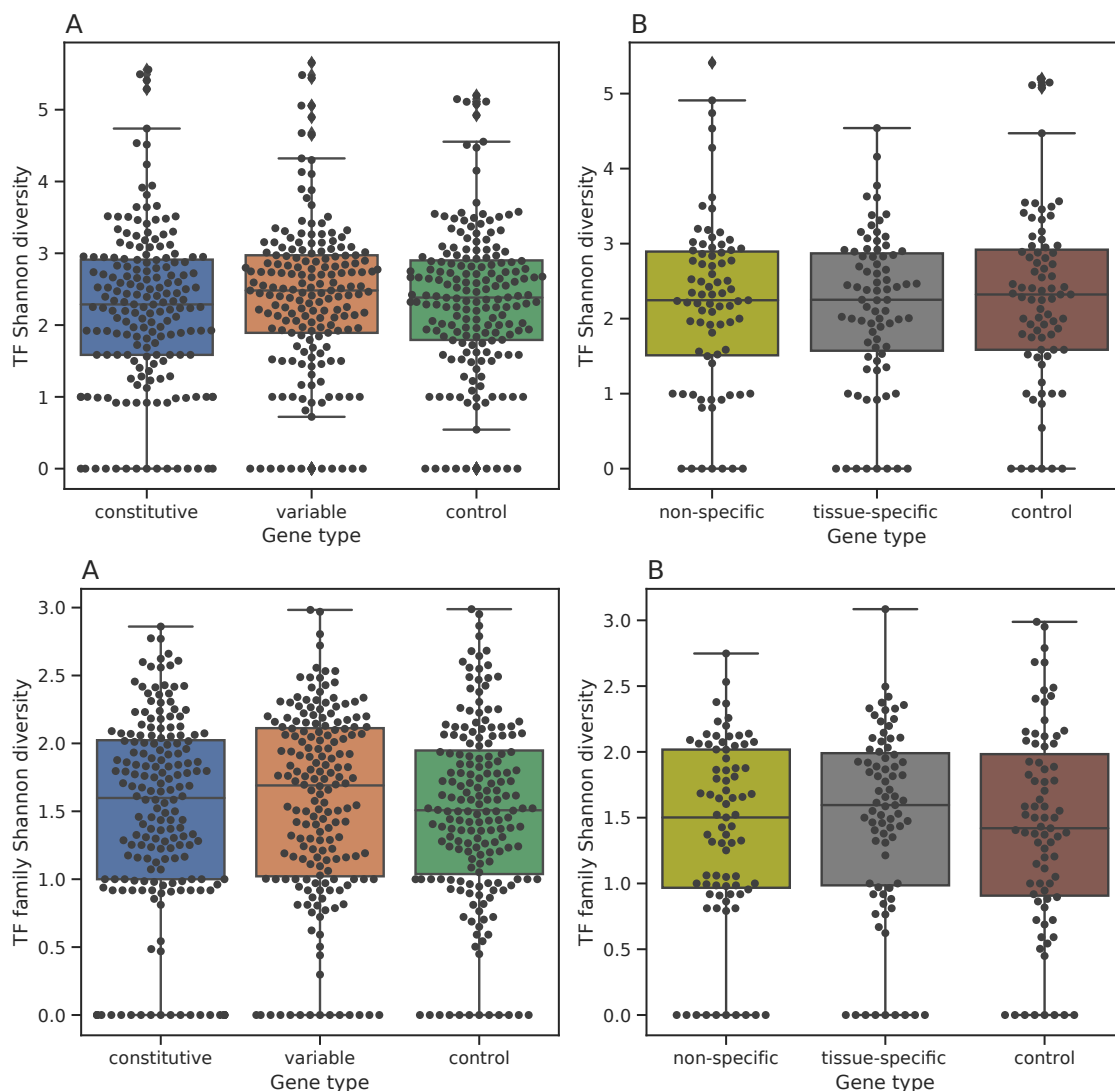


Figure S19: **Shannon diversity of individual TFs (A/B) and TF families (C/D) predicted to bind to the 400 nucleotides upstream of the ATG start codon in Arabidopsis CRMs within open chromatin regions.** Box plots have box boundaries that represent 25th, 50th (median) and 75th percentiles; whiskers are drawn up to the largest or smallest observed point that falls within 1.5 times the interquartile range. Shannon diversity of individual TFs in A: 300 constitutive, 300 variable and 300 control genes and B: 300 non-specific, 300 tissue-specific and 300 control genes. Shannon diversity of TF families in C: 300 constitutive, 300 variable and 300 control genes and D: 300 non-specific, 300 tissue-specific and 300 control genes.

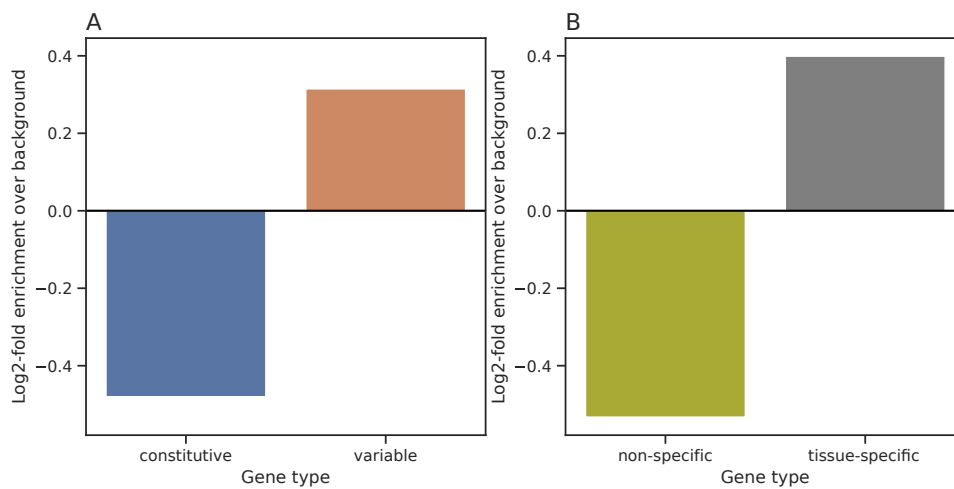


Figure S20: **Log₂-fold enrichment of 15 bp TATA boxes in Arabidopsis CRMs.** A: Log₂-fold enrichment of 15 bp TATA boxes in 300 variable (blue) and 300 constitutive (orange) Arabidopsis CRMs compared to a background of constitutive and variable promoters combined. B: Log₂-fold enrichment of 15 bp TATA boxes in 300 non-specific (green) and 300 tissue-specific (grey) Arabidopsis CRMs compared to a background of non-specific and tissue-specific promoters combined. TATA box locations within 50 bp upstream of the EPD TSS were downloaded from EPD⁴⁹⁸. Gat software³⁷⁷ was used to calculate enrichment.

References

1. Wang, Rongchen et al. (2010). ‘Multiple Regulatory Elements in the Arabidopsis NIA1 Promoter Act Synergistically to Form a Nitrate Enhancer1[W][OA]’. In: *Plant Physiology* 154.1, pp. 423–432 (pages [18](#), [61](#), [238](#), [243](#), [246](#), [247](#), [250](#), [252](#)).
2. Beerli, Roger R. et al. (1998). ‘Toward Controlling Gene Expression at Will: Specific Regulation of the erbB-2/HER-2 Promoter by Using Polydactyl Zinc Finger Proteins Constructed from Modular Building Blocks’. In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14628–14633 (pages [22](#), [62](#)).
3. Hernando, C. Esteban, Andrés Romanowski and Marcelo J. Yanovsky (2017). ‘Transcriptional and Post-Transcriptional Control of the Plant Circadian Gene Regulatory Network’. In: *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* 1860.1, pp. 84–94 (pages [31](#), [56](#), [222](#)).
4. Van den Broeck, Lisa et al. (2020). ‘Gene Regulatory Network Inference: Connecting Plant Biology and Mathematical Modeling’. In: *Frontiers in Genetics* 11 (pages [31](#), [102](#), [156](#), [157](#)).
5. MacNeil, Lesley T. and Albertha J. M. Walhout (2011). ‘Gene Regulatory Networks and the Role of Robustness and Stochasticity in the Control of Gene Expression’. In: *Genome Research* 21.5, pp. 645–657 (pages [31](#), [102](#)).
6. Noman, Nasimul et al. (2015). ‘Evolving Robust Gene Regulatory Networks’. In: *PLOS ONE* 10.1, e0116258 (page [31](#)).
7. Reeves, Gregory T. (2019). ‘The Engineering Principles of Combining a Transcriptional Incoherent Feedforward Loop with Negative Feedback’. In: *Journal of Biological Engineering* 13.1, p. 62 (pages [31](#), [32](#), [223](#), [255](#)).

8. Alon, Uri (2007). ‘Network Motifs: Theory and Experimental Approaches’. In: *Nature Reviews Genetics* 8.6 (6), pp. 450–461 (page 31).
9. Shoval, Oren and Uri Alon (2010). ‘SnapShot: Network Motifs’. In: *Cell* 143.2, 326–326.e1 (pages 32, 223, 255).
10. Maeda, Yusuke T. and Masaki Sano (2006). ‘Regulatory Dynamics of Synthetic Gene Networks with Positive Feedback’. In: *Journal of Molecular Biology* 359.4, pp. 1107–1124 (pages 32, 223, 255).
11. Kaufmann, Kerstin, Alice Pajoro and Gerco C. Angenent (2010). ‘Regulation of Transcription in Plants: Mechanisms Controlling Developmental Switches’. In: *Nature Reviews Genetics* 11.12 (12), pp. 830–842 (page 32).
12. Bilas, Róza et al. (2016). ‘Cis-Regulatory Elements Used to Control Gene Expression in Plants’. In: *Plant Cell, Tissue and Organ Culture (PCTOC)* 127.2, pp. 269–287 (pages 32, 44).
13. Porto, Milena Silva et al. (2014a). ‘Plant Promoters: An Approach of Structure and Function’. In: *Molecular Biotechnology* 56.1, pp. 38–49 (pages 32, 44).
14. Banerji, J., S. Rusconi and W. Schaffner (1981). ‘Expression of a Beta-Globin Gene Is Enhanced by Remote SV40 DNA Sequences’. In: *Cell* 27 (2 Pt 1), pp. 299–308 (pages 32, 44).
15. Schmitz, Robert J, Erich Grotewold and Maike Stam (2022). ‘Cis-Regulatory Sequences in Plants: Their Importance, Discovery, and Future Challenges’. In: *The Plant Cell* 34.2, pp. 718–741 (pages 32, 45, 60).
16. Moore, Lisa D., Thuc Le and Guoping Fan (2013). ‘DNA Methylation and Its Basic Function’. In: *Neuropsychopharmacology* 38.1 (1), pp. 23–38 (pages 32, 47).
17. Zhao, Ting, Zhenping Zhan and Danhua Jiang (2019). ‘Histone Modifications and Their Regulatory Roles in Plant Development and Environmental Memory’. In: *Journal of Genetics and Genomics* 46.10, pp. 467–476 (pages 32, 48).
18. Kelemen, Olga et al. (2013). ‘Function of Alternative Splicing’. In: *Gene* 514.1, pp. 1–30 (page 32).

19. Shyu, Ann-Bin, Miles F Wilkinson and Ambro van Hoof (2008). ‘Messenger RNA Regulation: To Translate or to Degrade’. In: *The EMBO Journal* 27.3, pp. 471–481 (page 32).
20. Merchante, Catharina, Anna N. Stepanova and Jose M. Alonso (2017). ‘Translation Regulation in Plants: An Interesting Past, an Exciting Present and a Promising Future’. In: *The Plant Journal* 90.4, pp. 628–653 (page 33).
21. Lu, Haiqi et al. (2016). ‘Regulation and Role of Post-Translational Modifications of Enhancer of Zeste Homologue 2 in Cancer Development’. In: *American Journal of Cancer Research* 6.12, p. 2737 (page 33).
22. Zhao, Lin et al. (2022). ‘Targeted Protein Degradation: Mechanisms, Strategies and Application’. In: *Signal Transduction and Targeted Therapy* 7.1 (1), pp. 1–13 (page 33).
23. Wu, Ting-Ying et al. (2021a). ‘Evolutionarily Conserved Hierarchical Gene Regulatory Networks for Plant Salt Stress Response’. In: *Nature Plants* 7.6 (6), pp. 787–799 (page 33).
24. Huang, Ling-Zhi et al. (2022a). ‘Gene Networks Involved in Plant Heat Stress Response and Tolerance’. In: *International Journal of Molecular Sciences* 23.19 (19), p. 11970 (page 33).
25. Gaudinier, Allison et al. (2018). ‘Transcriptional Regulation of Nitrogen-Associated Metabolism and Growth’. In: *Nature* 563.7730, pp. 259–264 (pages 33, 36, 50, 63, 64, 103–105, 113, 146, 149, 150, 159, 186–188, 226, 257).
26. Alvarez, José M. et al. (2020). ‘Transient Genome-Wide Interactions of the Master Transcription Factor NLP7 Initiate a Rapid Nitrogen-Response Cascade’. In: *Nature Communications* 11.1 (1), pp. 1–13 (pages 33, 35, 146, 149, 158, 254).
27. Frink, Charles R., Paul E. Waggoner and Jesse H. Ausubel (1999). ‘Nitrogen Fertilizer: Retrospect and Prospect’. In: *Proceedings of the National Academy of Sciences of the United States of America* 96.4, pp. 1175–1180 (page 33).

28. Pingali, Prabhu L. (2012). ‘Green Revolution: Impacts, Limits, and the Path Ahead’. In: *Proceedings of the National Academy of Sciences* 109.31, pp. 12302–12308 (page 33).
29. Khan, M. Nasir and Firoz Mohammad (2014). ‘Eutrophication: Challenges and Solutions’. In: *Eutrophication: Causes, Consequences and Control: Volume 2*. Ed. by Abid A. Ansari and Sarvajeet Singh Gill. Dordrecht: Springer Netherlands, pp. 1–15 (pages 33, 156).
30. Park, S. et al. (2012). ‘Trends and Seasonal Cycles in the Isotopic Composition of Nitrous Oxide since 1940’. In: *Nature Geoscience* 5.4 (4), pp. 261–265 (pages 33, 156).
31. Lehnert, Nicolai et al. (2018). ‘Reversing Nitrogen Fixation’. In: *Nature Reviews Chemistry* 2.10 (10), pp. 278–289 (page 33).
32. Gifford, Miriam L. et al. (2008). ‘Cell-Specific Nitrogen Responses Mediate Developmental Plasticity’. In: *Proceedings of the National Academy of Sciences* 105.2, pp. 803–808 (pages 33, 34).
33. Kiba, Takatoshi and Anne Krapp (2016). ‘Plant Nitrogen Acquisition Under Low Availability: Regulation of Uptake and Root Architecture’. In: *Plant and Cell Physiology* 57.4, pp. 707–714 (page 33).
34. Wang, Ya-Yun et al. (2018a). ‘Nitrate Transport, Signaling, and Use Efficiency’. In: *Annual Review of Plant Biology* 69.1, pp. 85–122 (pages 33, 34).
35. Liu, Kun-Hsiang and Yi-Fang Tsay (2003). ‘Switching between the Two Action Modes of the Dual-Affinity Nitrate Transporter CHL1 by Phosphorylation’. In: *The EMBO Journal* 22.5, pp. 1005–1013 (page 34).
36. Stitt, Mark (1999). ‘Nitrate Regulation of Metabolism and Growth’. In: *Current Opinion in Plant Biology* 2.3, pp. 178–186 (page 34).
37. Zhang, Hanma and Brian G. Forde (2000). ‘Regulation of Arabidopsis Root Development by Nitrate Availability’. In: *Journal of Experimental Botany* 51.342, pp. 51–59 (page 34).

38. Vidal, Elena A. et al. (2014). ‘Nitrogen Control of Developmental Phase Transitions in *Arabidopsis Thaliana*’. In: *Journal of Experimental Botany* 65.19, pp. 5611–5618 (page 34).
39. Wang, Rongchen et al. (2003). ‘Microarray Analysis of the Nitrate Response in *Arabidopsis* Roots and Shoots Reveals over 1,000 Rapidly Responding Genes and New Linkages to Glucose, Trehalose-6-Phosphate, Iron, and Sulfate Metabolism’. In: *Plant Physiology* 132.2, pp. 556–567 (page 34).
40. Marchive, Chloé et al. (2013). ‘Nuclear Retention of the Transcription Factor NLP7 Orchestrates the Early Response to Nitrate in Plants’. In: *Nature Communications* 4.1, pp. 1–9 (pages 34, 35, 188, 254, 255).
41. Liu, Kun-Hsiang et al. (2017). ‘Discovery of Nitrate-CPK-NLP Signalling in Central Nutrient-Growth Networks’. In: *Nature* 545.7654, pp. 311–316 (pages 34, 39, 188).
42. Krouk, Gabriel et al. (2010a). ‘Predictive Network Modeling of the High-Resolution Dynamic Plant Transcriptome in Response to Nitrate’. In: *Genome Biology* 11.12, R123 (page 34).
43. Contreras-López, Orlando et al. (2022). ‘Spatiotemporal Analysis Identifies ABF2 and ABF3 as Key Hubs of Endodermal Response to Nitrate’. In: *Proceedings of the National Academy of Sciences* 119.4, e2107879119 (page 34).
44. Guan, Peizhu et al. (2014). ‘Nitrate Foraging by *Arabidopsis* Roots Is Mediated by the Transcription Factor TCP20 through the Systemic Signaling Pathway’. In: *Proceedings of the National Academy of Sciences* 111.42, pp. 15267–15272 (pages 34–37).
45. Tsay, Yi-Fang (2014). ‘How to Switch Affinity’. In: *Nature* 507.7490, pp. 44–45 (page 34).
46. Wang, Xiaohan et al. (2021). ‘A Transceptor–Channel Complex Couples Nitrate Sensing to Calcium Signaling in *Arabidopsis*’. In: *Molecular Plant* 14.5, pp. 774–786 (pages 34, 36, 37, 39).
47. Remans, Tony et al. (2006). ‘The *Arabidopsis* NRT1.1 Transporter Participates in the Signaling Pathway Triggering Root Colonization of Nitrate-Rich Patches’. In:

- Proceedings of the National Academy of Sciences* 103.50, pp. 19206–19211 (pages [34](#), [142](#)).
48. Chen, Hui-Yu et al. (2021). ‘Potential Transceptor AtNRT1.13 Modulates Shoot Architecture and Flowering Time in a Nitrate-Dependent Manner’. In: *The Plant Cell* 33.5, pp. 1492–1505 (page [35](#)).
 49. Liu, Kun-Hsiang et al. (2022). ‘NIN-like Protein 7 Transcription Factor Is a Plant Nitrate Sensor’. In: *Science* 377.6613, pp. 1419–1425 (pages [35](#), [36](#), [158](#), [186](#), [231](#), [253](#)).
 50. Rubin, Grit et al. (2009). ‘Members of the *LBD* Family of Transcription Factors Repress Anthocyanin Synthesis and Affect Additional Nitrogen Responses in *Arabidopsis*’. In: *The Plant Cell* 21.11, pp. 3567–3584 (pages [35](#), [36](#)).
 51. Xu, Na et al. (2016). ‘The Arabidopsis NRG2 Protein Mediates Nitrate Signaling and Interacts with and Regulates Key Nitrate Regulators’. In: *The Plant Cell* 28.2, pp. 485–504 (pages [35](#), [36](#)).
 52. Zhang, Ting-Ting et al. (2021). ‘NIN-like Protein 7 Promotes Nitrate-Mediated Lateral Root Development by Activating Transcription of TRYPTOPHAN AMINOTRANSFERASE RELATED 2’. In: *Plant Science* 303, p. 110771 (pages [35](#), [39](#), [158](#), [186](#)).
 53. Wang, Honglei et al. (2022a). ‘Regulatory Functions of Cellular Energy Sensor SnRK1 for Nitrate Signalling through NLP7 Repression’. In: *Nature Plants* 8.9 (9), pp. 1094–1107 (pages [35–37](#)).
 54. Wang, Wei et al. (2018b). ‘Expression of the Nitrate Transporter Gene OsNRT1.1A/OsNPF6.3 Confers High Yield and Early Maturation in Rice’. In: *The Plant Cell* 30.3, pp. 638–651 (page [35](#)).
 55. Fang, Zhongming et al. (2013). ‘Altered Expression of the PTR/NRT1 Homologue OsPTR9 Affects Nitrogen Utilization Efficiency, Growth and Grain Yield in Rice’. In: *Plant Biotechnology Journal* 11.4, pp. 446–458 (page [35](#)).

56. Chen, Jingguang et al. (2017). ‘*pOsNAR2.1:OsNAR2.1* Expression Enhances Nitrogen Uptake Efficiency and Grain Yield in Transgenic Rice Plants’. In: *Plant Biotechnology Journal* 15.10, pp. 1273–1283 (page 35).
57. Chen, Jingguang et al. (2016). ‘Agronomic Nitrogen-Use Efficiency of Rice Can Be Increased by Driving *OsNRT2.1* Expression with the *OsNAR2.1* Promoter’. In: *Plant Biotechnology Journal* 14.8, pp. 1705–1715 (page 35).
58. Wu, Jie et al. (2021b). ‘Rice NIN-LIKE PROTEIN 4 Plays a Pivotal Role in Nitrogen Use Efficiency’. In: *Plant Biotechnology Journal* 19.3, pp. 448–461 (pages 35, 63).
59. Bouguyon, Eleonore et al. (2016). ‘Nitrate Controls Root Development through Post-Transcriptional Regulation of the NRT1.1/NPF6.3 Transporter/Sensor’. In: *Plant Physiology*, pp.01047.2016 (pages 36–38).
60. Krouk, Gabriel et al. (2010b). ‘Nitrate-Regulated Auxin Transport by NRT1.1 Defines a Mechanism for Nutrient Sensing in Plants’. In: *Developmental Cell* 18.6, pp. 927–937 (pages 37, 38).
61. Mair, Andrea et al. (2015). ‘SnRK1-triggered Switch of bZIP63 Dimerization Mediates the Low-Energy Response in Plants’. In: *eLife* 4. Ed. by Thorsten Nürnberger, e05828 (page 37).
62. Muralidhara, Prathibha et al. (2021). ‘Perturbations in Plant Energy Homeostasis Prime Lateral Root Initiation via SnRK1-bZIP63-ARF19 Signaling’. In: *Proceedings of the National Academy of Sciences* 118.37, e2106961118 (page 37).
63. Wilmoth, Jill C. et al. (2005). ‘NPH4/ARF7 and ARF19 Promote Leaf Expansion and Auxin-Induced Lateral Root Formation’. In: *The Plant Journal* 43.1, pp. 118–130 (page 37).
64. Guan, Peizhu et al. (2017). ‘Interacting TCP and NLP Transcription Factors Control Plant Responses to Nitrate Availability’. In: *Proceedings of the National Academy of Sciences* 114.9, pp. 2419–2424 (pages 37, 153).
65. Guan, Peizhu (2017). ‘Dancing with Hormones: A Current Perspective of Nitrate Signaling and Regulation in Arabidopsis’. In: *Frontiers in Plant Science* 8 (page 37).

66. Shakeel, Samina N. et al. (2021a). 'Interaction of Cytokinin and Ethylene in the Regulation of Primary Root Growth and Development'. In: *Rhizobiology: Molecular Physiology of Plant Roots*. Ed. by Soumya Mukherjee and František Baluška. Signaling and Communication in Plants. Cham: Springer International Publishing, pp. 195–238 (pages [37](#), [38](#)).
67. Krouk, Gabriel (2016). 'Hormones and Nitrate: A Two-Way Connection'. In: *Plant Molecular Biology* 91.6, pp. 599–606 (pages [37](#), [38](#)).
68. Kurepa, Jasmina and Jan A. Smalle (2022). 'Auxin/Cytokinin Antagonistic Control of the Shoot/Root Growth Ratio and Its Relevance for Adaptation to Drought and Nutrient Deficiency Stresses'. In: *International Journal of Molecular Sciences* 23.4 (4), p. 1933 (pages [38](#), [39](#)).
69. Yang, Qianying et al. (2022). 'ABCG11 Modulates Cytokinin Responses in Arabidopsis Thaliana'. In: *Frontiers in Plant Science* 13 (page [38](#)).
70. Poitout, Arthur et al. (2018). 'Responses to Systemic Nitrogen Signaling in Arabidopsis Roots Involve *Trans*-Zeatin in Shoots'. In: *The Plant Cell* 30.6, pp. 1243–1257 (page [38](#)).
71. Laplaze, Laurent et al. (2008). 'Cytokinins Act Directly on Lateral Root Founder Cells to Inhibit Root Initiation'. In: *The Plant Cell* 19.12, pp. 3889–3900 (page [38](#)).
72. Rivas, M. Ángeles et al. (2022). 'Auxin-Cytokinin Balance Shapes Maize Root Architecture by Controlling Primary Root Elongation and Lateral Root Development'. In: *Frontiers in Plant Science* 13 (page [38](#)).
73. Lamig, Liliana et al. (2022). 'Molecular Mechanisms Underlying Nitrate Responses in Plants'. In: *Current Biology* 32.9, R433–R439 (page [38](#)).
74. Vidal, Elena A. et al. (2010). 'Nitrate-Responsive miR393/AFB3 Regulatory Module Controls Root System Architecture in Arabidopsis Thaliana'. In: *Proceedings of the National Academy of Sciences* 107.9, pp. 4477–4482 (page [38](#)).
75. Ljung, Karin et al. (2005). 'Sites and Regulation of Auxin Biosynthesis in Arabidopsis Roots'. In: *The Plant Cell* 17.4, pp. 1090–1104 (page [38](#)).

76. Benková, Eva et al. (2003). ‘Local, Efflux-Dependent Auxin Gradients as a Common Module for Plant Organ Formation’. In: *Cell* 115.5, pp. 591–602 (page 38).
77. Guo, Fang-Qing, Rongchen Wang and Nigel M. Crawford (2002). ‘The Arabidopsis Dual-affinity Nitrate Transporter Gene AtNRT1.1 (CHL1) Is Regulated by Auxin in Both Shoots and Roots’. In: *Journal of Experimental Botany* 53.370, pp. 835–844 (page 38).
78. Okamoto, Mamoru, J. John Vidmar and Anthony D. M. Glass (2003). ‘Regulation of NRT1 and NRT2 Gene Families of Arabidopsis Thaliana: Responses to Nitrate Provision’. In: *Plant and Cell Physiology* 44.3, pp. 304–317 (page 38).
79. Maghiaoui, Amel et al. (2020). ‘The Arabidopsis NRT1.1 Transceptor Coordinately Controls Auxin Biosynthesis and Transport to Regulate Root Branching in Response to Nitrate’. In: *Journal of Experimental Botany* 71.15, pp. 4480–4494 (page 38).
80. Devi, Loitongbam Lorinda et al. (2022). ‘The Interplay of Auxin and Brassinosteroid Signaling Tunes Root Growth under Low and Different Nitrogen Forms’. In: *Plant Physiology* 189.3, pp. 1757–1773 (page 39).
81. Jia, Zhongtao, Ricardo F.H. Giehl and Nicolaus von Wirén (2020a). ‘The Root Foraging Response under Low Nitrogen Depends on DWARF1-Mediated Brassinosteroid Biosynthesis1 [OPEN]’. In: *Plant Physiology* 183.3, pp. 998–1010 (page 39).
82. Jia, Zhongtao et al. (2019). ‘Natural Variation of BSK3 Tunes Brassinosteroid Signaling to Regulate Root Foraging under Low Nitrogen’. In: *Nature Communications* 10.1 (1), p. 2378 (page 39).
83. Ren, Hong et al. (2019). ‘BRASSINOSTEROID-SIGNALING KINASE 3, a Plasma Membrane-Associated Scaffold Protein Involved in Early Brassinosteroid Signaling’. In: *PLOS Genetics* 15.1, e1007904 (page 39).
84. Song, Xiaoyun et al. (2021). ‘CALMODULIN-LIKE-38 and PEP1 RECEPTOR 2 Integrate Nitrate and Brassinosteroid Signals to Regulate Root Growth’. In: *Plant Physiology* 187.3, pp. 1779–1794 (page 39).

85. Alvarez, José M. et al. (2014). ‘Systems Approach Identifies TGA1 and TGA4 Transcription Factors as Important Regulatory Components of the Nitrate Response of Arabidopsis Thaliana Roots’. In: *The Plant Journal* 80.1, pp. 1–13 (page 40).
86. Canales, Javier et al. (2017). ‘Nitrate Induction of Root Hair Density Is Mediated by TGA1/TGA4 and CPC Transcription Factors in Arabidopsis Thaliana’. In: *The Plant Journal* 92.2, pp. 305–316 (page 40).
87. Kim, Yeong-Woo et al. (2022). ‘Brassinosteroids Enhance Salicylic Acid-Mediated Immune Responses by Inhibiting BIN2 Phosphorylation of Clade I TGA Transcription Factors in Arabidopsis’. In: *Molecular Plant* 15.6, pp. 991–1007 (page 40).
88. Fujita, Yasunari et al. (2009). ‘Three SnRK2 Protein Kinases Are the Main Positive Regulators of Abscisic Acid Signaling in Response to Water Stress in Arabidopsis’. In: *Plant and Cell Physiology* 50.12, pp. 2123–2132 (page 41).
89. Gonzalez-Guzman, Miguel et al. (2012). ‘Arabidopsis PYR/PYL/RCAR Receptors Play a Major Role in Quantitative Regulation of Stomatal Aperture and Transcriptional Response to Abscisic Acid’. In: *The Plant Cell* 24.6, pp. 2483–2496 (page 41).
90. Soon, Fen-Fen et al. (2012). ‘Molecular Mimicry Regulates ABA Signaling by SnRK2 Kinases and PP2C Phosphatases’. In: *Science* 335.6064, pp. 85–88 (page 41).
91. Lee, Won J. et al. (2020). ‘NITROGEN RESPONSE DEFICIENCY 1-Mediated CHL1 Induction Contributes to Optimized Growth Performance during Altered Nitrate Availability in Arabidopsis’. In: *The Plant Journal* 104.5, pp. 1382–1398 (page 41).
92. Su, Hang et al. (2021). ‘Abscisic Acid Signaling Negatively Regulates Nitrate Uptake via Phosphorylation of NRT1.1 by SnRK2s in Arabidopsis’. In: *Journal of Integrative Plant Biology* 63.3, pp. 597–610 (page 41).
93. Yan, Dawei et al. (2016). ‘NIN-like Protein 8 Is a Master Regulator of Nitrate-Promoted Seed Germination in Arabidopsis’. In: *Nature Communications* 7.1 (1), p. 13179 (page 41).

94. Camut, Lucie et al. (2021). 'Nitrate Signaling Promotes Plant Growth by Upregulating Gibberellin Biosynthesis and Destabilization of DELLA Proteins'. In: *Current Biology* 31.22, 4971–4982.e4 (page 41).
95. Jamieson, Fiona et al. (2022). 'Ethylene Signaling Modulates Arabidopsis Thaliana Nitrate Metabolism'. In: *Planta* 255.5, p. 94 (page 41).
96. Binder, Brad M. (2020). 'Ethylene Signaling in Plants'. In: *Journal of Biological Chemistry* 295.22, pp. 7710–7725 (page 42).
97. Zhang, Guo-Bin, Hong-Ying Yi and Ji-Ming Gong (2014). 'The Arabidopsis Ethylene/Jasmonic Acid-NRT Signaling Module Coordinates Nitrate Reallocation and the Trade-Off between Growth and Environmental Adaptation'. In: *The Plant Cell* 26.10, pp. 3984–3998 (page 42).
98. Lin, Shan-Hua et al. (2008). 'Mutation of the Arabidopsis NRT1.5 Nitrate Transporter Causes Defective Root-to-Shoot Nitrate Transport'. In: *The Plant Cell* 20.9, pp. 2514–2528 (page 42).
99. Feng, Ying et al. (2017). 'Ethylene Promotes Root Hair Growth through Coordinated EIN3/EIL1 and RHD6/RSL1 Activity in Arabidopsis'. In: *Proceedings of the National Academy of Sciences* 114.52, pp. 13834–13839 (page 42).
100. Shakeel, Samina N. et al. (2021b). 'Interaction of Cytokinin and Ethylene in the Regulation of Primary Root Growth and Development'. In: *Rhizobiology: Molecular Physiology of Plant Roots*. Ed. by Soumya Mukherjee and František Baluška. Signaling and Communication in Plants. Cham: Springer International Publishing, pp. 195–238 (page 42).
101. Dugardeyn, Jasper and Dominique Van Der Straeten (2008). 'Ethylene: Fine-tuning Plant Growth and Development by Stimulation and Inhibition of Elongation'. In: *Plant Science. Ethylene Biology* 175.1, pp. 59–70 (page 42).
102. Huang, Yupu et al. (2022b). 'Arabidopsis ERF012 Is a Versatile Regulator of Plant Growth, Development and Abiotic Stress Responses'. In: *International Journal of Molecular Sciences* 23.12, p. 6841 (page 42).

103. Feng, Guanqiao et al. (2020). ‘Jasmonate Induced Alternative Splicing Responses in *Arabidopsis*’. In: *Plant Direct* 4.8 (page 42).
104. Sun, Jiaqiang et al. (2009). ‘*Arabidopsis* ASA1 Is Important for Jasmonate-Mediated Regulation of Auxin Biosynthesis and Transport during Lateral Root Formation’. In: *The Plant Cell* 21.5, pp. 1495–1511 (page 42).
105. Waters, Mark T. et al. (2017). ‘Strigolactone Signaling and Evolution’. In: *Annual Review of Plant Biology* 68.1, pp. 291–322 (page 42).
106. Kohlen, Wouter et al. (2012). ‘The Tomato CAROTENOID CLEAVAGE DIOXYGENASE8 (SICCD8) Regulates Rhizosphere Signaling, Plant Architecture and Affects Reproductive Development through Strigolactone Biosynthesis’. In: *New Phytologist* 196.2, pp. 535–547 (page 42).
107. Marro, Nicolás et al. (2022). ‘Strigolactones: New Players in the Nitrogen–Phosphorus Signalling Interplay’. In: *Plant, Cell & Environment* 45.2, pp. 512–527 (page 42).
108. Zhang, Jing et al. (2020a). ‘Strigolactones Inhibit Auxin Feedback on PIN-dependent Auxin Transport Canalization’. In: *Nature Communications* 11.1 (1), p. 3508 (page 42).
109. Ito, Shinsaku et al. (2016). ‘Effects of Strigolactone Signaling on *Arabidopsis* Growth under Nitrogen Deficient Stress Condition’. In: *Plant Signaling & Behavior* 11.1, e1126031 (page 42).
110. Sun, Huwei et al. (2021). ‘SPL14/17 Act Downstream of Strigolactone Signalling to Modulate Rice Root Elongation in Response to Nitrate Supply’. In: *The Plant Journal* 106.3, pp. 649–660 (page 42).
111. Liu, Yansheng, Andreas Beyer and Ruedi Aebersold (2016). ‘On the Dependency of Cellular Protein Levels on mRNA Abundance’. In: *Cell* 165.3, pp. 535–550 (page 42).
112. Wittkopp, Patricia J. and Gizem Kalay (2012). ‘*Cis*-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence’. In: *Nature Reviews Genetics* 13.1, pp. 59–69 (pages 43, 157, 260).

113. Swinnen, Gwen, Alain Goossens and Laurens Pauwels (2016). ‘Lessons from Domestication: Targeting Cis-Regulatory Elements for Crop Improvement’. In: *Trends in Plant Science* 21.6, pp. 506–515 (page 43).
114. Kuo, Chao-Chung et al. (2019). ‘Detection of RNA–DNA Binding Sites in Long Noncoding RNAs’. In: *Nucleic Acids Research* (pages 43, 46).
115. Guo, Haitao and Hongwei Huo (2017). *A New Algorithm for Identifying Cis-Regulatory Modules Based on Hidden Markov Model*. BioMed Research International. URL: <https://www.hindawi.com/journals/bmri/2017/6274513/> (visited on 27/03/2019) (page 43).
116. Jeziorska, Danuta M., Kate W. Jordan and Keith W. Vance (2009). ‘A Systems Biology Approach to Understanding Cis-Regulatory Module Function’. In: *Seminars in Cell & Developmental Biology*. Structure and Function of the Golgi Apparatus and Systems Approaches to Cell and Developmental Biology 20.7, pp. 856–862 (page 43).
117. Tümpel, Stefan et al. (2008). ‘A Regulatory Module Embedded in the Coding Region of Hoxa2 Controls Expression in Rhombomere 2’. In: *Proceedings of the National Academy of Sciences* 105.51, pp. 20077–20082 (page 43).
118. Ostrovsky, Olga et al. (2018). ‘Identification of Strong Intron Enhancer in the Heparanase Gene: Effect of Functional Rs4693608 Variant on HPSE Enhancer Activity in Hematological and Solid Malignancies’. In: *Oncogenesis* 7.6, p. 51 (page 43).
119. Bolle, C. et al. (1994). ‘Segments Encoding 5'-Untranslated Leaders of Genes for Thylakoid Proteins Contain Cis-Elements Essential for Transcription’. In: *The Plant Journal: For Cell and Molecular Biology* 6.4, pp. 513–523 (page 43).
120. Henry, Kelli F. et al. (2018). ‘A Shared Cis-Regulatory Module Activates Transcription in the Suspensor of Plant Embryos’. In: *Proceedings of the National Academy of Sciences* 115.25, E5824–E5833 (page 43).
121. Palmer, Matthew B. et al. (2007). ‘A 3' Enhancer Controls Snail Expression in Melanoma Cells’. In: *Cancer Research* 67.13, pp. 6113–6120 (page 43).
122. Yochum, Gregory S., Ryan Cleland and Richard H. Goodman (2008). ‘A Genome-Wide Screen for Beta-Catenin Binding Sites Identifies a Downstream Enhancer

- Element That Controls c-Myc Gene Expression'. In: *Molecular and Cellular Biology* 28.24, pp. 7368–7379 (page 43).
123. Bien-Willner, Gabriel A., Paweł Stankiewicz and James R. Lupski (2007). 'SOX9 cre1 , a Cis-Acting Regulatory Element Located 1.1 Mb Upstream of SOX9, Mediates Its Enhancement through the SHH Pathway'. In: *Human Molecular Genetics* 16.10, pp. 1143–1156 (page 43).
 124. Oka, Rurika et al. (2017). 'Genome-Wide Mapping of Transcriptional Enhancer Candidates Using DNA and Chromatin Features in Maize'. In: *Genome Biology* 18.1, p. 137 (page 43).
 125. Solovyev, Victor V., Ilham A. Shahmuradov and Asaf A. Salamov (2010). 'Identification of Promoter Regions and Regulatory Sites'. In: *Computational Biology of Transcription Factor Binding*. Ed. by Istvan Ladunga. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 57–83 (page 44).
 126. Roy, Ananda L. and Dinah S. Singer (2015). 'Core Promoters in Transcription: Old Problem, New Insights'. In: *Trends in Biochemical Sciences* 40.3, pp. 165–171 (page 44).
 127. Carter, David et al. (2002). 'Long-Range Chromatin Regulatory Interactions *in Vivo*'. In: *Nature Genetics* 32.4, pp. 623–626 (page 44).
 128. Tolhuis, Bas et al. (2002). 'Looping and Interaction between Hypersensitive Sites in the Active β -Globin Locus'. In: *Molecular Cell* 10.6, pp. 1453–1465 (page 44).
 129. Vilar, Jose MG and Leonor Saiz (2005). 'DNA Looping in Gene Regulation: From the Assembly of Macromolecular Complexes to the Control of Transcriptional Noise'. In: *Current Opinion in Genetics & Development*. Chromosomes and Expression Mechanisms 15.2, pp. 136–144 (page 44).
 130. De Laat, Wouter et al. (2008). 'Chapter 5 Three-Dimensional Organization of Gene Expression in Erythroid Cells'. In: *Current Topics in Developmental Biology*. Vol. 82. Red Cell Development. Academic Press, pp. 117–139 (page 44).

131. Amano, Takanori et al. (2009). ‘Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription’. In: *Developmental Cell* 16.1, pp. 47–57 (page 44).
132. Corless, Samuel and Nick Gilbert (2016). ‘Effects of DNA Supercoiling on Chromatin Architecture’. In: *Biophysical Reviews* 8.3, pp. 245–258 (page 44).
133. Core, Leighton J., Joshua J. Waterfall and John T. Lis (2008). ‘Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters’. In: *Science* 322.5909, pp. 1845–1848 (page 44).
134. Engreitz, Jesse M. et al. (2016). ‘Local Regulation of Gene Expression by lncRNA Promoters, Transcription and Splicing’. In: *Nature* 539.7629, pp. 452–455 (page 44).
135. Andersson, Robin (2015). ‘Promoter or Enhancer, What’s the Difference? Deconstruction of Established Distinctions and Presentation of a Unifying Model’. In: *BioEssays* 37.3, pp. 314–323 (page 44).
136. Ding, Mengting et al. (2018). ‘Enhancer RNAs (eRNAs): New Insights into Gene Transcription and Disease Treatment’. In: *Journal of Cancer* 9.13, pp. 2334–2340 (page 44).
137. Ørom, Ulf Andersson et al. (2010). ‘Long Noncoding RNAs with Enhancer-like Function in Human Cells’. In: *Cell* 143.1, pp. 46–58 (page 44).
138. Gallegos, Jenna E. and Alan B. Rose (2017). ‘Intron DNA Sequences Can Be More Important Than the Proximal Promoter in Determining the Site of Transcript Initiation’. In: *The Plant Cell* 29.4, pp. 843–853 (page 44).
139. Zabidi, Muhammad A. et al. (2015). ‘Enhancer–Core–Promoter Specificity Separates Developmental and Housekeeping Gene Regulation’. In: *Nature* 518.7540 (7540), pp. 556–559 (page 44).
140. Danko, Charles G. et al. (2018). ‘Dynamic Evolution of Regulatory Element Ensembles in Primate CD4 + T Cells’. In: *Nature Ecology & Evolution* 2.3, p. 537 (page 44).

141. Ngan, Chew Yee et al. (2020). 'Chromatin Interaction Analyses Elucidate the Roles of PRC2-bound Silencers in Mouse Development'. In: *Nature Genetics* 52.3 (3), pp. 264–272 (page 45).
142. Pang, Baoxu and Michael P. Snyder (2020). 'Systematic Identification of Silencers in Human Cells'. In: *Nature Genetics* 52.3 (3), pp. 254–263 (page 45).
143. Kurbidaeva, Amina and Michael Purugganan (2021). 'Insulators in Plants: Progress and Open Questions'. In: *Genes* 12.9 (9), p. 1422 (page 45).
144. Heger, Peter and Thomas Wiehe (2014). 'New Tools in the Box: An Evolutionary Synopsis of Chromatin Insulators'. In: *Trends in Genetics* 30.5, pp. 161–171 (page 45).
145. Choudhury, Swarup Roy, Sujit Roy and Dibyendu N. Sengupta (2009). 'A Comparative Study of Cultivar Differences in Sucrose Phosphate Synthase Gene Expression and Sucrose Formation during Banana Fruit Ripening'. In: *Postharvest Biology and Technology* 54.1, pp. 15–24 (page 45).
146. Du, Hwei, Zuxin Zhang and Jiansheng Li (2010). 'Isolation and Functional Characterization of a Waterlogging-Induced Promoter from Maize'. In: *Plant Cell Reports* 29.11, pp. 1269–1275 (page 45).
147. Bian, Chunxiang et al. (2011). 'Cloning, Expression and Characterization of the Putative Nuclear Transport Factor 2 (NTF2) Gene from Moss *Conocephalum Conicum*(L.) Dum'. In: *Molecular Biology Reports* 38.3, pp. 2023–2032 (page 45).
148. Byszewska, Magdalena et al. (2014). 'RNA Methyltransferases Involved in 5' Cap Biosynthesis'. In: *RNA Biology* 11.12, pp. 1597–1607 (page 45).
149. Ramanathan, Anand, G. Brett Robb and Siu-Hong Chan (2016). 'mRNA Capping: Biological Functions and Applications'. In: *Nucleic Acids Research* 44.16, pp. 7511–7526 (page 45).
150. Temperley, Richard J. et al. (2010). 'Human Mitochondrial mRNAs—like Members of All Families, Similar but Different'. In: *Biochimica et Biophysica Acta (BBA) - Bioenergetics*. 16th European Bioenergetics Conference 2010 1797.6, pp. 1081–1085 (page 45).

151. Monde, Rita A, Gadi Schuster and David B Stern (2000). 'Processing and Degradation of Chloroplast mRNA'. In: *Biochimie* 82.6, pp. 573–582 (page 45).
152. Calvo, Sarah E., David J. Pagliarini and Vamsi K. Mootha (2009). 'Upstream Open Reading Frames Cause Widespread Reduction of Protein Expression and Are Polymorphic among Humans'. In: *Proceedings of the National Academy of Sciences* 106.18, pp. 7507–7512 (page 45).
153. Beaudoin, Jean-Denis and Jean-Pierre Perreault (2010). '5'-UTR G-quadruplex Structures Acting as Translational Repressors'. In: *Nucleic Acids Research* 38.20, pp. 7022–7036 (page 45).
154. Jiménez-González, Augusto Samuel et al. (2014). 'Functional and Structural Analysis of Maize Hsp101 IRES'. In: *PLOS ONE* 9.9, e107459 (page 45).
155. Smith, Laura (2008). 'Post-Transcriptional Regulation of Gene Expression by Alternative 5'-Untranslated Regions in Carcinogenesis'. In: *Biochemical Society Transactions* 36 (Pt 4), pp. 708–711 (page 45).
156. Lamesch, Philippe et al. (2012). 'The Arabidopsis Information Resource (TAIR): Improved Gene Annotation and New Tools'. In: *Nucleic Acids Research* 40 (Database issue), pp. D1202–1210 (pages 45, 89, 96).
157. Morton, Taj et al. (2014). 'Paired-End Analysis of Transcription Start Sites in Arabidopsis Reveals Plant-Specific Promoter Signatures'. In: *The Plant Cell* 26.7, pp. 2746–2760 (pages 45, 193, 194, 217, 218, 250).
158. Thieffry, Axel et al. (2020). 'Characterization of Arabidopsis Thaliana Promoter Bidirectionality and Antisense RNAs by Depletion of Nuclear RNA Decay Pathways'. In: *The Plant Cell* (page 45).
159. Watson, James D. et al. (2007). *Recombinant DNA: Genes and Genomes: A Short Course*. W. H. Freeman. 500 pp. (page 45).
160. Zhang, Yong et al. (2009). 'Intrinsic Histone-DNA Interactions Are Not the Major Determinant of Nucleosome Positions *in Vivo*'. In: *Nature Structural & Molecular Biology* 16.8, pp. 847–852 (page 46).

161. Geyer, P. K., M. M. Green and V. G. Corces (1990). 'Tissue-Specific Transcriptional Enhancers May Act in Trans on the Gene Located in the Homologous Chromosome: The Molecular Basis of Transvection in *Drosophila*'. In: *The EMBO journal* 9.7, pp. 2247–2256 (page 46).
162. Xu, Na, Chia-Lun Tsai and Jeannie T. Lee (2006). 'Transient Homologous Chromosome Pairing Marks the Onset of X Inactivation'. In: *Science (New York, N.Y.)* 311.5764, pp. 1149–1152 (page 46).
163. Lomvardas, Stavros et al. (2006). 'Interchromosomal Interactions and Olfactory Receptor Choice'. In: *Cell* 126.2, pp. 403–413 (page 46).
164. Petrenko, Natalia et al. (2019). 'Requirements for RNA Polymerase II Preinitiation Complex Formation in Vivo'. In: *eLife* 8. Ed. by Michael R Green and James L Manley, e43654 (page 46).
165. Soutourina, Julie (2018). 'Transcription Regulation by the Mediator Complex'. In: *Nature Reviews Molecular Cell Biology* 19.4 (4), pp. 262–274 (page 46).
166. Proudfoot, Nick J. (2016). 'Transcriptional Termination in Mammals: Stopping the RNA Polymerase II Juggernaut'. In: *Science (New York, N.Y.)* 352.6291, aad9926 (page 46).
167. Toledo-Ortiz, Gabriela et al. (2014). 'The HY5-PIF Regulatory Module Coordinates Light and Temperature Control of Photosynthetic Gene Transcription'. In: *PLOS Genetics* 10.6, e1004416 (pages 47, 254).
168. Nuruzzaman, Mohammed, Akhter Most Sharoni and Shoshi Kikuchi (2013). 'Roles of NAC Transcription Factors in the Regulation of Biotic and Abiotic Stress Responses in Plants'. In: *Frontiers in Microbiology* 4 (page 47).
169. Shiu, Shin-Han, Ming-Che Shih and Wen-Hsiung Li (2005). 'Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals'. In: *Plant Physiology* 139.1, pp. 18–26 (page 47).
170. Pabo, C O and R T Sauer (1984). 'Protein-DNA Recognition'. In: *Annual Review of Biochemistry* 53.1, pp. 293–321 (page 47).

171. Bewley, C A, A M Gronenborn and G M Clore (1998). ‘Minor Groove-Binding Architectural Proteins: Structure, Function, and DNA Recognition.’ In: *Annual Review of Biophysics and Biomolecular Structure* 27, pp. 105–131 (page 47).
172. Ganji, Mahipal et al. (2016). ‘DNA Binding Proteins Explore Multiple Local Configurations during Docking via Rapid Rebinding.’ In: *Nucleic Acids Research* 44.17, pp. 8376–8384 (page 47).
173. Chiu, Tsu-Pei et al. (2020). ‘TFBSshape: An Expanded Motif Database for DNA Shape Features of Transcription Factor Binding Sites.’ In: *Nucleic Acids Research* 48.D1, pp. D246–D255 (page 47).
174. Li, Jinsen et al. (2017a). ‘Expanding the Repertoire of DNA Shape Features for Genome-Scale Studies of Transcription Factor Binding.’ In: *Nucleic Acids Research* 45.22, pp. 12877–12887 (page 47).
175. Rohs, Remo et al. (2009). ‘The Role of DNA Shape in Protein–DNA Recognition.’ In: *Nature* 461.7268 (7268), pp. 1248–1253 (page 47).
176. Sielemann, Janik et al. (2021). ‘Local DNA Shape Is a General Principle of Transcription Factor Binding Specificity in Arabidopsis Thaliana.’ In: *Nature Communications* 12.1 (1), p. 6549 (page 47).
177. Gordân, Raluca et al. (2013). ‘Genomic Regions Flanking E-Box Binding Sites Influence DNA Binding Specificity of bHLH Transcription Factors through DNA Shape.’ In: *Cell Reports* 3.4, pp. 1093–1104 (page 47).
178. Dantas Machado, Ana Carolina et al. (2015). ‘Evolving Insights on How Cytosine Methylation Affects Protein–DNA Binding.’ In: *Briefings in Functional Genomics* 14.1, pp. 61–73 (page 47).
179. Mathiasen, Lisa et al. (2016). ‘The Flexibility of a Homeodomain Transcription Factor Heterodimer and Its Allosteric Regulation by DNA Binding.’ In: *The FEBS Journal* 283.16, pp. 3134–3154 (page 47).
180. Huang, Shao-shan C. and Joseph R. Ecker (2018). ‘Piecing Together Cis-Regulatory Networks: Insights from Epigenomics Studies in Plants.’ In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 10.3, e1411 (page 47).

181. O'Malley, Ronan C. et al. (2016). 'Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape'. In: *Cell* 165.5, pp. 1280–1292 (pages [47](#), [48](#), [55](#), [69](#), [83](#), [96](#), [100](#), [105](#), [107](#), [148–150](#), [195](#), [204–209](#), [231](#), [264](#), [265](#)).
182. Mondal, Tanmoy et al. (2010). 'Characterization of the RNA Content of Chromatin'. In: *Genome Research* 20.7, pp. 899–907 (page [47](#)).
183. Richmond, Timothy J. and Curt A. Davey (2003). 'The Structure of DNA in the Nucleosome Core'. In: *Nature* 423.6936, p. 145 (page [47](#)).
184. Struhl, Kevin and Eran Segal (2013). 'Determinants of Nucleosome Positioning'. In: *Nature Structural & Molecular Biology* 20.3, pp. 267–273 (page [47](#)).
185. Zaret, Kenneth S. and Jason S. Carroll (2011). 'Pioneer Transcription Factors: Establishing Competence for Gene Expression'. In: *Genes & Development* 25.21, pp. 2227–2241 (page [48](#)).
186. Wu, Carl et al. (1979). 'The Chromatin Structure of Specific Genes: I. Evidence for Higher Order Domains of Defined DNA Sequence'. In: *Cell* 16.4, pp. 797–806 (page [49](#)).
187. Mather, E L and R P Perry (1983). 'Methylation Status and DNase I Sensitivity of Immunoglobulin Genes: Changes Associated with Rearrangement'. In: *Proceedings of the National Academy of Sciences of the United States of America* 80.15, pp. 4689–4693 (page [49](#)).
188. Bender, M. A et al. (2000). ' β -Globin Gene Switching and DNase I Sensitivity of the Endogenous β -Globin Locus in Mice Do Not Require the Locus Control Region'. In: *Molecular Cell* 5.2, pp. 387–393 (page [49](#)).
189. Wang, Xi and Robert T. Simpson (2001). 'Chromatin Structure Mapping in *Saccharomyces Cerevisiae* in Vivo with DNase I'. In: *Nucleic Acids Research* 29.9, pp. 1943–1950 (page [49](#)).
190. Cockerill, Peter N. (2000). 'Identification of DNaseI Hypersensitive Sites Within Nuclei'. In: *Transcription Factor Protocols*. Ed. by Martin J. Tymms. Methods in Molecular Biology™. Totowa, NJ: Humana Press, pp. 29–46 (page [49](#)).

191. Pfeifer, G P and A D Riggs (1991). 'Chromatin Differences between Active and Inactive X Chromosomes Revealed by Genomic Footprinting of Permeabilized Cells Using DNase I and Ligation-Mediated PCR.' In: *Genes & Development* 5.6, pp. 1102–1113 (page 49).
192. Feng, Junli and Bryant Villeponteau (1992). 'High-Resolution Analysis of c-Fos Chromatin Accessibility Using a Novel DNase I-PCR Assay'. In: *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* 1130.3, pp. 253–258 (page 49).
193. McArthur, Michael, Shawn Gerum and George Stamatoyannopoulos (2001). 'Quantification of DNaseI-sensitivity by Real-Time PCR: Quantitative Analysis of DNaseI-hypersensitivity of the Mouse β -Globin LCR11Edited by J. Karn'. In: *Journal of Molecular Biology* 313.1, pp. 27–34 (page 49).
194. Dorschner, Michael O. et al. (2004). 'High-Throughput Localization of Functional Elements by Quantitative Chromatin Profiling'. In: *Nature Methods* 1.3, pp. 219–225 (page 49).
195. Martins, Rui Pires, Adrian E. Platts and Stephen A. Krawetz (2007). 'Tracking Chromatin States Using Controlled DNase I Treatment and Real-Time PCR'. In: *Cellular & Molecular Biology Letters* 12.4, pp. 545–555 (page 49).
196. Crawford, Gregory E. et al. (2004). 'Identifying Gene Regulatory Elements by Genome-Wide Recovery of DNase Hypersensitive Sites'. In: *Proceedings of the National Academy of Sciences* 101.4, pp. 992–997 (page 49).
197. Sabo, Peter J. et al. (2004). 'Genome-Wide Identification of DNaseI Hypersensitive Sites Using Active Chromatin Sequence Libraries'. In: *Proceedings of the National Academy of Sciences* 101.13, pp. 4537–4542 (page 49).
198. Crawford, Gregory E. et al. (2006a). 'DNase-chip: A High-Resolution Method to Identify DNase I Hypersensitive Sites Using Tiled Microarrays'. In: *Nature Methods* 3.7, pp. 503–509 (page 49).
199. Sabo, Peter J. et al. (2006). 'Genome-Scale Mapping of DNase I Sensitivity *in Vivo* Using Tiling DNA Microarrays'. In: *Nature Methods* 3.7, pp. 511–518 (page 49).

200. Follows, George A. et al. (2006). 'Identifying Gene Regulatory Elements by Genomic Microarray Mapping of DNaseI Hypersensitive Sites'. In: *Genome Research* 16.10, pp. 1310–1319 (page 49).
201. Shu, Huan et al. (2012). 'Distinct Modes of DNA Accessibility in Plant Chromatin'. In: *Nature Communications* 3, p. 1281 (page 49).
202. Nagy, Peter L. et al. (2003). 'Genomewide Demarcation of RNA Polymerase II Transcription Units Revealed by Physical Fractionation of Chromatin'. In: *Proceedings of the National Academy of Sciences* 100.11, pp. 6364–6369 (page 49).
203. Boyle, Alan P. et al. (2008). 'High-Resolution Mapping and Characterization of Open Chromatin across the Genome'. In: *Cell* 132.2, pp. 311–322 (page 49).
204. Giresi, Paul G. et al. (2007). 'FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) Isolates Active Regulatory Elements from Human Chromatin'. In: *Genome Research* 17.6, pp. 877–885 (page 49).
205. Crawford, Gregory E. et al. (2006b). 'Genome-Wide Mapping of DNase Hypersensitive Sites Using Massively Parallel Signature Sequencing (MPSS)'. In: *Genome Research* 16.1, pp. 123–131 (page 49).
206. Song, Lingyun et al. (2011). 'Open Chromatin Defined by DNaseI and FAIRE Identifies Regulatory Elements That Shape Cell-Type Identity'. In: *Genome Research* 21.10, pp. 1757–1767 (page 49).
207. Jin, Wenfei et al. (2015). 'Genome-Wide Detection of DNase I Hypersensitive Sites in Single Cells and FFPE Tissue Samples'. In: *Nature* 528.7580, pp. 142–146 (page 49).
208. Buenrostro, Jason D. et al. (2013). 'Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-binding Proteins and Nucleosome Position'. In: *Nature Methods* 10.12, pp. 1213–1218 (page 49).
209. Gangadharan, Sunil et al. (2010). 'DNA Transposon Hermes Inserts into DNA in Nucleosome-Free Regions in Vivo'. In: *Proceedings of the National Academy of Sciences* 107.51, pp. 21966–21972 (page 49).

210. Cusanovich, Darren A. et al. (2015). ‘Multiplex Single-Cell Profiling of Chromatin Accessibility by Combinatorial Cellular Indexing’. In: *Science* 348.6237, pp. 910–914 (page 50).
211. Buenrostro, Jason D. et al. (2015). ‘Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation’. In: *Nature* 523.7561, pp. 486–490 (page 50).
212. Weber, Christopher M., Jorja G. Henikoff and Steven Henikoff (2010). ‘H2A.Z Nucleosomes Enriched over Active Genes Are Homotypic’. In: *Nature Structural & Molecular Biology* 17.12, pp. 1500–1507 (page 50).
213. Henikoff, Jorja G. et al. (2011). ‘Epigenome Characterization at Single Base-Pair Resolution’. In: *Proceedings of the National Academy of Sciences* 108.45, pp. 18318–18323 (page 50).
214. Pajoro, Alice et al. (2018). ‘Profiling Nucleosome Occupancy by MNase-seq: Experimental Protocol and Computational Analysis’. In: *Plant Chromatin Dynamics: Methods and Protocols*. Ed. by Marian Bemer and Célia Baroux. Methods in Molecular Biology. New York, NY: Springer New York, pp. 167–181 (page 50).
215. Kelly, Theresa K. et al. (2012). ‘Genome-Wide Mapping of Nucleosome Positioning and DNA Methylation within Individual DNA Molecules’. In: *Genome Research* 22.12, pp. 2497–2506 (page 50).
216. Li, J. and I Herskowitz (1993). ‘Isolation of ORC6, a Component of the Yeast Origin Recognition Complex by a One-Hybrid System’. In: *Science* 262.5141, pp. 1870–1874 (pages 50, 103).
217. Reece-Hoyes, John S. and A. J. Marian Walhout (2012). ‘Yeast One-Hybrid Assays: A Historical and Technical Perspective’. In: *Methods. Protein-Protein Interactions* 57.4, pp. 441–447 (pages 50, 104).
218. Reece-Hoyes, John S. et al. (2011). ‘Enhanced Yeast One-Hybrid Assays for High-Throughput Gene-Centered Regulatory Network Mapping’. In: *Nature Methods* 8.12, pp. 1059–1064 (page 50).

219. Reece-Hoyes, John S. and Albertha J. M. Walhout (2012). ‘Gene-Centered Yeast One-Hybrid Assays’. In: *Two Hybrid Technologies*. Ed. by Bernhard Suter and Erich E. Wanker. Vol. 812. Totowa, NJ: Humana Press, pp. 189–208 (page 51).
220. Bargmann, Bastiaan O. R. et al. (2013). ‘TARGET: A Transient Transformation System for Genome-Wide Transcription Factor Target Discovery’. In: *Molecular Plant* 6.3, pp. 978–980 (pages 51, 103, 106).
221. Bargmann, Bastiaan O. R. and Kenneth D. Birnbaum (2010). ‘Fluorescence Activated Cell Sorting of Plant Protoplasts’. In: *JoVE (Journal of Visualized Experiments)* 36, e1673 (page 51).
222. Dahl, John Arne and Philippe Collas (2008). ‘ μ ChIP—a Rapid Micro Chromatin Immunoprecipitation Assay for Small Cell Samples and Biopsies’. In: *Nucleic Acids Research* 36.3, e15–e15 (page 51).
223. Para, Alessia, Ying Li and Gloria M. Coruzzi (2018). ‘ μ ChIP-Seq for Genome-Wide Mapping of In Vivo TF-DNA Interactions in Arabidopsis Root Protoplasts’. In: *Root Development: Methods and Protocols*. Ed. by Daniela Ristova and Elke Barbez. Methods in Molecular Biology. New York, NY: Springer, pp. 249–261 (page 51).
224. Para, Alessia et al. (2014). ‘Hit-and-Run Transcriptional Control by bZIP1 Mediates Rapid Nutrient Signaling in Arabidopsis’. In: *Proceedings of the National Academy of Sciences* 111.28, pp. 10371–10376 (pages 51, 149).
225. Doidy, Joan et al. (2016). “Hit-and-Run” Transcription: De Novo Transcription Initiated by a Transient bZIP1 “Hit” Persists after the “Run”. In: *BMC Genomics* 17.1, p. 92 (page 51).
226. Galas, D J and A Schmitz (1978). ‘DNase Footprinting: A Simple Method for the Detection of Protein-DNA Binding Specificity’. In: *Nucleic Acids Research* 5.9, pp. 3157–3170 (page 52).
227. Sullivan, Alessandra M. et al. (2014). ‘Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. Thaliana’. In: *Cell Reports* 8.6, pp. 2015–2030 (page 52).

228. Zeng, Weihua and Ali Mortazavi (2012). ‘Technical Considerations for Functional Sequencing Assays’. In: *Nature Immunology* 13.9 (9), pp. 802–807 (page 53).
229. Garner, M M and A Revzin (1981). ‘A Gel Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia Coli Lactose Operon Regulatory System.’ In: *Nucleic Acids Research* 9.13, pp. 3047–3060 (page 53).
230. Tuerk, C. and L. Gold (1990). ‘Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase’. In: *Science* 249.4968, pp. 505–510 (page 53).
231. Solomon, Mark J., Pamela L. Larsen and Alexander Varshavsky (1988). ‘Mapping proteinDNA Interactions in Vivo with Formaldehyde: Evidence That Histone H4 Is Retained on a Highly Transcribed Gene’. In: *Cell* 53.6, pp. 937–947 (page 53).
232. Orlando, Valerio (2000). ‘Mapping Chromosomal Proteins in Vivo by Formaldehyde-Crosslinked-Chromatin Immunoprecipitation’. In: *Trends in Biochemical Sciences* 25.3, pp. 99–104 (page 53).
233. Hebbes, T R, A W Thorne and C Crane-Robinson (1988). ‘A Direct Link between Core Histone Acetylation and Transcriptionally Active Chromatin.’ In: *The EMBO Journal* 7.5, pp. 1395–1402 (page 53).
234. Das, Partha M. et al. (2004). ‘Chromatin Immunoprecipitation Assay’. In: *BioTechniques* 37.6, pp. 961–969 (page 54).
235. Weinmann, Amy S. et al. (2001). ‘Use of Chromatin Immunoprecipitation To Clone Novel E2F Target Promoters’. In: *Molecular and Cellular Biology* 21.20, pp. 6820–6832 (page 54).
236. Lee Kang, Sung-Hae, Karen Vieira and Jörg Bungert (2002). ‘Combining Chromatin Immunoprecipitation and DNA Footprinting: A Novel Method to Analyze Protein–DNA Interactions in Vivo’. In: *Nucleic Acids Research* 30.10, e44–e44 (page 54).
237. Weinmann, Amy S. et al. (2002). ‘Isolating Human Transcription Factor Targets by Coupling Chromatin Immunoprecipitation and CpG Island Microarray Analysis’. In: *Genes & Development* 16.2, pp. 235–244 (page 54).

238. Wells, Julie et al. (2003). ‘Identification of Novel pRb Binding Sites Using CpG Microarrays Suggests That E2F Recruits pRb to Specific Genomic Sites during S Phase’. In: *Oncogene* 22.10, pp. 1445–1460 (page 54).
239. Ren, Bing et al. (2000). ‘Genome-Wide Location and Function of DNA Binding Proteins’. In: *Science* 290.5500, pp. 2306–2309 (page 54).
240. Johnson, David S. et al. (2007). ‘Genome-Wide Mapping of in Vivo Protein-DNA Interactions’. In: *Science* 316.5830, pp. 1497–1502 (page 54).
241. Kidder, Benjamin L., Gangqing Hu and Keji Zhao (2011). ‘ChIP-Seq: Technical Considerations for Obtaining High-Quality Data’. In: *Nature Immunology* 12, pp. 918–922 (page 54).
242. Steensel, Bas van and Steven Henikoff (2000). ‘Identification of in Vivo DNA Targets of Chromatin Proteins Using Tethered Dam Methyltransferase’. In: *Nature Biotechnology* 18.4 (4), pp. 424–428 (page 54).
243. Sánchez-Romero, María A, Ignacio Cota and Josep Casadesús (2015). ‘DNA Methylation in Bacteria: From the Methyl Group to the Methylome’. In: *Current Opinion in Microbiology*. Environmental Microbiology • Extremophiles 25, pp. 9–16 (page 54).
244. Aughey, Gabriel N. and Tony D. Southall (2016). ‘Dam It’s Good! DamID Profiling of Protein-DNA Interactions’. In: *WIREs Developmental Biology* 5.1, pp. 25–37 (page 54).
245. Franco-Zorrilla, José M. et al. (2014). ‘DNA-binding Specificities of Plant Transcription Factors and Their Potential to Define Target Genes’. In: *Proceedings of the National Academy of Sciences* 111.6, pp. 2367–2372 (page 54).
246. Bartlett, Anna et al. (2017). ‘Mapping Genome-Wide Transcription-Factor Binding Sites Using DAP-seq’. In: *Nature Protocols* 12.8 (8), pp. 1659–1672 (page 55).
247. Bailey, T. L. and C. Elkan (1994). ‘Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers’. In: *Proceedings. International Conference on Intelligent Systems for Molecular Biology* 2, pp. 28–36 (page 56).
248. Reid, John E. and Lorenz Wernisch (2011). ‘STEME: Efficient EM to Find Motifs in Large Data Sets’. In: *Nucleic Acids Research* 39.18, e126–e126 (page 56).

249. Zambelli, Federico, Graziano Pesole and Giulio Pavese (2013). ‘Motif Discovery and Transcription Factor Binding Sites before and after the Next-Generation Sequencing Era’. In: *Briefings in Bioinformatics* 14.2, pp. 225–237 (page 56).
250. Ikebata, Hisaki and Ryo Yoshida (2015). ‘Repulsive Parallel MCMC Algorithm for Discovering Diverse Motifs from Large Sequence Sets’. In: *Bioinformatics* 31.10, pp. 1561–1568 (page 56).
251. Bailey, Timothy L. (2011). ‘DREME: Motif Discovery in Transcription Factor ChIP-seq Data’. In: *Bioinformatics* 27.12, pp. 1653–1659 (page 56).
252. Alipanahi, Babak et al. (2015). ‘Predicting the Sequence Specificities of DNA- and RNA-binding Proteins by Deep Learning’. In: *Nature Biotechnology* 33.8, pp. 831–838 (page 56).
253. Mutalik, Vivek K. et al. (2013). ‘Precise and Reliable Gene Expression via Standard Transcription and Translation Initiation Elements’. In: *Nature Methods* 10.4 (4), pp. 354–360 (pages 56, 223).
254. Bashor, Caleb J. and James J. Collins (2018). ‘Understanding Biological Regulation Through Synthetic Biology’. In: *Annual Review of Biophysics* 47.1, pp. 399–423 (page 56).
255. Freeman, Ronit, Tali Finder and Itamar Willner (2009). ‘Multiplexed Analysis of Hg²⁺ and Ag⁺ Ions by Nucleic Acid Functionalized CdSe/ZnS Quantum Dots and Their Use for Logic Gate Operations’. In: *Angewandte Chemie* 121.42, pp. 7958–7961 (pages 56, 57).
256. Gander, Miles W. et al. (2017). ‘Digital Logic Circuits in Yeast with CRISPR-dCas9 NOR Gates’. In: *Nature Communications* 8.1 (1), p. 15459 (page 57).
257. Xie, Zhen et al. (2011). ‘Multi-Input RNAi-Based Logic Circuit for Identification of Specific Cancer Cells’. In: *Science* 333.6047, pp. 1307–1311 (page 57).
258. Mendes, T. et al. (2017). ‘Systems and Synthetic Biology Applied to Health’. In: *Current Developments in Biotechnology and Bioengineering*. Elsevier, pp. 183–213 (page 57).

259. Brophy, Jennifer A. N. et al. (2022). ‘Synthetic Genetic Circuits as a Means of Reprogramming Plant Roots’. In: *Science* 377.6607, pp. 747–751 (pages [57](#), [60](#), [62](#), [63](#), [224](#), [248](#), [260](#)).
260. Bhaya, Devaki, Michelle Davison and Rodolphe Barrangou (2011). ‘CRISPR-Cas Systems in Bacteria and Archaea: Versatile Small RNAs for Adaptive Defense and Regulation’. In: *Annual Review of Genetics* 45.1, pp. 273–297 (page [58](#)).
261. Koonin, Eugene V, Kira S Makarova and Feng Zhang (2017). ‘Diversity, Classification and Evolution of CRISPR-Cas Systems’. In: *Current Opinion in Microbiology. Environmental Microbiology* * CRISPRcas9 37, pp. 67–78 (page [58](#)).
262. Jinek, Martin et al. (2012). ‘A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity’. In: *Science* 337.6096, pp. 816–821 (page [58](#)).
263. Zhang, F. (2019). ‘Development of CRISPR-Cas Systems for Genome Editing and Beyond’. In: *Quarterly Reviews of Biophysics* 52, e6 (page [59](#)).
264. Somyajit, Kumar et al. (2021). ‘Homology-Directed Repair Protects the Replicating Genome from Metabolic Assaults’. In: *Developmental Cell* 56.4, 461–477.e7 (pages [59](#), [157](#)).
265. Jeong, Yeong Yeop et al. (2021). ‘Optimization of Protoplast Regeneration in the Model Plant *Arabidopsis Thaliana*’. In: *Plant Methods* 17.1, p. 21 (page [59](#)).
266. Metje-Sprink, Janina et al. (2019). ‘DNA-Free Genome Editing: Past, Present and Future’. In: *Frontiers in Plant Science* 9 (page [59](#)).
267. Alper, Hal et al. (2005). ‘Tuning Genetic Control through Promoter Engineering’. In: *Proceedings of the National Academy of Sciences* 102.36, pp. 12678–12683 (page [59](#)).
268. Jensen, Peter Ruhdal and Karin Hammer (1998). ‘The Sequence of Spacers between the Consensus Sequences Modulates the Strength of Prokaryotic Promoters’. In: *Appl. Environ. Microbiol.* 64.1, pp. 82–87 (page [59](#)).
269. Cox, Robert Sidney, Michael G Surette and Michael B Elowitz (2007). ‘Programming Gene Expression with Combinatorial Promoters’. In: *Molecular Systems Biology* 3 (page [59](#)).

270. Curran, Kathleen A. et al. (2014). ‘Design of Synthetic Yeast Promoters via Tuning of Nucleosome Architecture’. In: *Nature Communications* 5, p. 4002 (page 59).
271. Kotopka, Benjamin J. and Christina D. Smolke (2020). ‘Model-Driven Generation of Artificial Yeast Promoters’. In: *Nature Communications* 11.1 (1), pp. 1–13 (pages 59, 60, 232, 250).
272. Redden, Heidi and Hal S. Alper (2015). ‘The Development and Characterization of Synthetic Minimal Yeast Promoters’. In: *Nature Communications* 6, p. 7810 (page 60).
273. Boer, Carl G. de et al. (2019). ‘Deciphering Eukaryotic Gene-Regulatory Logic with 100 Million Random Promoters’. In: *Nature Biotechnology*, pp. 1–10 (page 60).
274. Mattioli, Kaia et al. (2019). ‘High-Throughput Functional Analysis of lncRNA Core Promoters Elucidates Rules Governing Tissue Specificity’. In: *Genome Research* (pages 60, 194).
275. Jores, Tobias et al. (2021). ‘Synthetic Promoter Designs Enabled by a Comprehensive Analysis of Plant Core Promoters’. In: *Nature Plants* 7.6 (6), pp. 842–855 (pages 60, 61, 216, 218, 219, 232).
276. Porto, Milena Silva et al. (2014b). ‘Plant Promoters: An Approach of Structure and Function’. In: *Molecular Biotechnology* 56.1, pp. 38–49 (page 60).
277. Liu, Wusheng and C Neal Stewart (2016). ‘Plant Synthetic Promoters and Transcription Factors’. In: *Current Opinion in Biotechnology. Food Biotechnology • Plant Biotechnology* 37, pp. 36–44 (page 60).
278. Cai, Yao-Min et al. (2020a). ‘Rational Design of Minimal Synthetic Promoters for Plants’. In: *Nucleic Acids Research* 48.21, pp. 11845–11856 (pages 60, 61, 69, 70, 114, 160, 224, 232, 233, 241, 242).
279. Wang, Rui et al. (2015a). ‘Novel Green Tissue-Specific Synthetic Promoters and *Cis*-Regulatory Elements in Rice’. In: *Scientific Reports* 5, p. 18256 (page 60).
280. Rushton, Paul J. et al. (2002). ‘Synthetic Plant Promoters Containing Defined Regulatory Elements Provide Novel Insights into Pathogen- and Wound-Induced Signaling’. In: *The Plant Cell* 14.4, pp. 749–762 (page 61).

281. Sawant, Samir V. et al. (2005). ‘A Variety of Synergistic and Antagonistic Interactions Mediated by Cis-Acting DNA Motifs Regulate Gene Expression in Plant Cells and Modulate Stability of the Transcription Complex Formed on a Basal Promoter’. In: *Journal of Experimental Botany* 56.419, pp. 2345–2353 (pages [61](#), [224](#), [249](#)).
282. Wu, Chuan-Yin et al. (1998). ‘The GCN4 Motif in a Rice Glutelin Gene Is Essential for Endosperm-Specific Gene Expression and Is Activated by Opaque-2 in Transgenic Rice Plants’. In: *The Plant Journal* 14.6, pp. 673–683 (page [61](#)).
283. Zrimec, Jan et al. (2022). ‘Controlling Gene Expression with Deep Generative Design of Regulatory DNA’. In: *Nature Communications* 13.1 (1), p. 5099 (page [61](#)).
284. Brückner, Kathleen et al. (2015). ‘A Library of Synthetic Transcription Activator-like Effector-Activated Promoters for Coordinated Orthogonal Gene Expression in Plants’. In: *The Plant Journal: For Cell and Molecular Biology* 82.4, pp. 707–716 (pages [61](#), [224](#)).
285. Zhang, Ning, Leah K. McHale and John J. Finer (2019). ‘Changes to the Core and Flanking Sequences of G-box Elements Lead to Increases and Decreases in Gene Expression in Both Native and Synthetic Soybean Promoters’. In: *Plant Biotechnology Journal* 17.4, pp. 724–735 (pages [61](#), [224](#)).
286. Freire-Rios, Alejandra et al. (2020). ‘Architecture of DNA Elements Mediating ARF Transcription Factor Binding and Auxin-Responsive Gene Expression in *Arabidopsis*’. In: *Proceedings of the National Academy of Sciences* 117.39, pp. 24557–24566 (pages [61](#), [108](#), [113](#), [224](#)).
287. Mehrotra, Rajesh and Sandhya Mehrotra (2010). ‘Promoter Activation by ACGT in Response to Salicylic and Abscisic Acids Is Differentially Regulated by the Spacing between Two Copies of the Motif’. In: *Journal of Plant Physiology* 167.14, pp. 1214–1218 (pages [61](#), [249](#)).
288. Doshi, Aarti et al. (2020). ‘Small-Molecule Inducible Transcriptional Control in Mammalian Cells’. In: *Critical Reviews in Biotechnology* 40.8, pp. 1131–1150 (page [61](#)).

289. Triezenberg, S. J., R. C. Kingsbury and S. L. McKnight (1988). ‘Functional Dissection of VP16, the Trans-Activator of Herpes Simplex Virus Immediate Early Gene Expression.’ In: *Genes & Development* 2.6, pp. 718–729 (page 61).
290. Hirai, Hiroyuki, Tetsuya Tani and Nobuaki Kikyo (2010). ‘Structure and Functions of Powerful Transactivators: VP16, MyoD and FoxA.’ In: *The International Journal of Developmental Biology* 54.11-12, pp. 1589–1596 (page 62).
291. Garcia-Perez, Elena et al. (2022). ‘A Copper Switch for Inducing CRISPR/Cas9-based Transcriptional Activation Tightly Regulates Gene Expression in *Nicotiana Benthamiana*.’ In: *BMC Biotechnology* 22.1, p. 12 (pages 62, 224).
292. Bernabé-Orts, Joan Miquel et al. (2020). ‘A Memory Switch for Plant Synthetic Biology Based on the Phage ϕ C31 Integration System.’ In: *Nucleic Acids Research* 48.6, pp. 3379–3394 (page 62).
293. Hiratsu, Keiichiro et al. (2003). ‘Dominant Repression of Target Genes by Chimeric Repressors That Include the EAR Motif, a Repression Domain, in *Arabidopsis*.’ In: *The Plant Journal* 34.5, pp. 733–739 (pages 62, 248).
294. Tang, Xu et al. (2017). ‘A CRISPR–Cpf1 System for Efficient Genome Editing and Transcriptional Repression in Plants.’ In: *Nature Plants* 3.3 (3), pp. 1–5 (pages 62, 71, 72, 226, 227, 247).
295. Mahfouz, Magdy M. et al. (2012). ‘Targeted Transcriptional Repression Using a Chimeric TALE-SRDX Repressor Protein.’ In: *Plant Molecular Biology* 78.3, pp. 311–321 (pages 62, 248).
296. Dossani, Zain Y. et al. (2018). ‘A Combinatorial Approach to Synthetic Transcription Factor-Promoter Combinations for Yeast Strain Engineering.’ In: *Yeast* 35.3, pp. 273–280 (page 62).
297. Danila, Florence et al. (2022). ‘A Single Promoter-TALE System for Tissue-Specific and Tuneable Expression of Multiple Genes in Rice.’ In: *Plant Biotechnology Journal* 20.9, pp. 1786–1806 (pages 62, 224).

298. Piatek, Agnieszka et al. (2015). ‘RNA-guided Transcriptional Regulation in Planta via Synthetic dCas9-based Transcription Factors’. In: *Plant Biotechnology Journal* 13.4, pp. 578–589 (page [62](#)).
299. Papikian, Ashot et al. (2019). ‘Site-Specific Manipulation of Arabidopsis Loci Using CRISPR-Cas9 SunTag Systems’. In: *Nature Communications* 10.1 (1), p. 729 (pages [62](#), [63](#), [71](#), [72](#), [226](#), [229](#)).
300. Gilbert, Luke A. et al. (2014). ‘Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation’. In: *Cell* 159.3, pp. 647–661 (pages [62](#), [226](#)).
301. Pan, Changtian et al. (2021). ‘CRISPR–Act3.0 for Highly Efficient Multiplexed Gene Activation in Plants’. In: *Nature Plants* 7.7 (7), pp. 942–953 (pages [62](#), [63](#), [71](#), [226](#), [227](#), [229](#), [247](#), [256](#)).
302. Zhu, Haocheng, Chao Li and Caixia Gao (2020). ‘Applications of CRISPR–Cas in Agriculture and Plant Biotechnology’. In: *Nature Reviews Molecular Cell Biology* 21.11 (11), pp. 661–677 (page [63](#)).
303. Liu, Qier et al. (2021a). ‘Application of CRISPR/Cas9 in Crop Quality Improvement’. In: *International Journal of Molecular Sciences* 22.8 (8), p. 4206 (page [63](#)).
304. Wada, Naoki et al. (2020). ‘Precision Genome Editing in Plants: State-of-the-Art in CRISPR/Cas9-based Genome Engineering’. In: *BMC Plant Biology* 20.1, p. 234 (page [63](#)).
305. Liu, Lei et al. (2021b). ‘Enhancing Grain-Yield-Related Traits by CRISPR–Cas9 Promoter Editing of Maize CLE Genes’. In: *Nature Plants* 7.3 (3), pp. 287–294 (page [63](#)).
306. Rodríguez-Leal, Daniel et al. (2017). ‘Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing’. In: *Cell* 171.2, 470–480.e8 (pages [63](#), [157](#), [158](#), [259](#)).
307. Song, Xiaoguang et al. (2022). ‘Targeting a Gene Regulatory Element Enhances Rice Grain Yield by Decoupling Panicle Number and Size’. In: *Nature Biotechnology* 40.9 (9), pp. 1403–1411 (pages [63](#), [158](#), [259](#)).

308. Li, Zhenxiang et al. (2017b). ‘A Potent Cas9-derived Gene Activator for Plant and Mammalian Cells’. In: *Nature Plants* 3.12 (12), pp. 930–936 (page 63).
309. Patron, Nicola J. et al. (2015). ‘Standards for Plant Synthetic Biology: A Common Syntax for Exchange of DNA Parts’. In: *The New Phytologist* 208.1, pp. 13–19 (pages 63, 68, 69, 114).
310. Leydon, Alexander R. et al. (2020). ‘Engineering Synthetic Signaling in Plants’. In: *Annual Review of Plant Biology* 71.1, pp. 767–788 (page 63).
311. Foo, Mathias, Ozgur E. Akman and Declan G. Bates (2022). ‘Restoring Circadian Gene Profiles in Clock Networks Using Synthetic Feedback Control’. In: *npj Systems Biology and Applications* 8.1 (1), pp. 1–11 (page 63).
312. Del Vecchio, Domitilla, Aaron J. Dy and Yili Qian (2016). ‘Control Theory Meets Synthetic Biology’. In: *Journal of The Royal Society Interface* 13.120, p. 20160380 (page 63).
313. Del Vecchio, Domitilla et al. (2018). ‘Future Systems and Control Research in Synthetic Biology’. In: *Annual Reviews in Control* 45, pp. 5–17 (page 63).
314. Hsiao, Victoria, Anandh Swaminathan and Richard M. Murray (2018). ‘Control Theory for Synthetic Biology: Recent Advances in System Characterization, Control Design, and Controller Implementation for Synthetic Biology’. In: *IEEE Control Systems Magazine* 38.3, pp. 32–62 (page 63).
315. Foo, Mathias et al. (2018). ‘A Framework for Engineering Stress Resilient Plants Using Genetic Feedback Control and Regulatory Network Rewiring’. In: *ACS Synthetic Biology* 7.6, pp. 1553–1564 (page 63).
316. Wei, Shaobo et al. (2022). ‘A Transcriptional Regulator That Boosts Grain Yields and Shortens the Growth Duration of Rice’. In: *Science* 377.6604, eabi8455 (page 63).
317. Aoki, Stephanie K. et al. (2019). ‘A Universal Biomolecular Integral Feedback Controller for Robust Perfect Adaptation’. In: *Nature* 570.7762 (7762), pp. 533–537 (page 64).

318. Brooks, Matthew D. et al. (2019). ‘Network Walking Charts Transcriptional Dynamics of Nitrogen Signaling by Integrating Validated and Predicted Genome-Wide Interactions’. In: *Nature Communications* 10.1, pp. 1–13 (pages [65](#), [103](#), [133](#), [135](#), [143](#), [151–154](#), [224](#), [231](#), [248](#), [249](#), [251](#), [255](#)).
319. Baig, Hasan et al. (2021). ‘Synthetic Biology Open Language Visual (SBOL Visual) Version 3.0’. In: *Journal of Integrative Bioinformatics* 18.3 (page [67](#)).
320. Pollak, Bernardo et al. (2018). ‘Loop Assembly: A Simple and Open System for Recursive Fabrication of DNA Circuits’. In: *bioRxiv*, p. 247593 (pages [67](#), [68](#), [161](#)).
321. Sarrion-Perdigones, Alejandro et al. (2013). ‘GoldenBraid 2.0: A Comprehensive DNA Assembly Framework for Plant Synthetic Biology’. In: *Plant Physiology* 162.3, pp. 1618–1631 (pages [68](#), [69](#), [114](#)).
322. Sauret-Güeto, Susanna et al. (2020). ‘Systematic Tools for Reprogramming Plant Gene Expression in a Simple Model, *Marchantia Polymorpha*’. In: *ACS Synthetic Biology* 9.4, pp. 864–882 (pages [68–71](#), [73](#), [114](#)).
323. Cai, Yao-Min, Jose A. Carrasco Lopez and Nicola J. Patron (2020b). ‘Phytobricks: Manual and Automated Assembly of Constructs for Engineering Plants’. In: *DNA Cloning and Assembly: Methods and Protocols*. Ed. by Sunil Chandran and Kevin W. George. Methods in Molecular Biology. New York, NY: Springer US, pp. 179–199 (pages [69](#), [114](#)).
324. Marillonnet, Sylvestre and Ramona Grützner (2020). ‘Synthetic DNA Assembly Using Golden Gate Cloning and the Hierarchical Modular Cloning Pipeline’. In: *Current Protocols in Molecular Biology* 130.1 (pages [69](#), [71](#)).
325. Engler, Carola et al. (2014). ‘A Golden Gate Modular Cloning Toolbox for Plants’. In: *ACS Synthetic Biology* 3.11, pp. 839–843 (pages [70](#), [71](#), [73](#), [77](#), [114](#), [227](#), [228](#), [232](#), [243](#), [247](#), [252](#)).
326. Wang, Rongchen et al. (2009). ‘A Genetic Screen for Nitrate Regulatory Mutants Captures the Nitrate Transporter Gene *NRT1.1*’. In: *Plant Physiology* 151.1, pp. 472–478 (pages [70–72](#)).

327. Castel, Baptiste et al. (2019). ‘Optimization of T-DNA Architecture for Cas9-mediated Mutagenesis in Arabidopsis’. In: *PLOS ONE* 14.1, e0204778 (pages [70](#), [90](#)).
328. Shimada, Takashi L., Tomoo Shimada and Ikuko Hara-Nishimura (2010). ‘A Rapid and Non-Destructive Screenable Marker, FAST, for Identifying Transformed Seeds of Arabidopsis Thaliana’. In: *The Plant Journal* 61.3, pp. 519–528 (pages [70](#), [71](#), [77](#), [227](#), [228](#), [243](#), [247](#), [252](#)).
329. Pédelacq, Jean-Denis et al. (2006). ‘Engineering and Characterization of a Superfolder Green Fluorescent Protein’. In: *Nature Biotechnology* 24.1 (1), pp. 79–88 (page [71](#)).
330. Hily, Jean-Michel et al. (2009). ‘A Transformation Booster Sequence (TBS) from *Petunia Hybrida* Functions as an Enhancer-Blocking Insulator in Arabidopsis Thaliana’. In: *Plant Cell Reports* 28.7, pp. 1095–1104 (pages [71](#), [227](#)).
331. Chen, Baohui et al. (2013). ‘Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System’. In: *Cell* 155.7, pp. 1479–1491 (pages [71](#), [89](#), [161](#), [226](#)).
332. Lawrenson, Tom et al. (2015). ‘Induction of Targeted, Heritable Mutations in Barley and Brassica Oleracea Using RNA-guided Cas9 Nuclease’. In: *Genome Biology* 16.1, p. 258 (pages [71](#), [89](#)).
333. Yu, Xiao-Hong and Chang-Jun Liu (2006). ‘Development of an Analytical Method for Genome-Wide Functional Identification of Plant Acyl-Coenzyme A-dependent Acyltransferases’. In: *Analytical Biochemistry* 358.1, pp. 146–148 (page [72](#)).
334. Gan, Han Ming, Melvin V. L. Lee and Michael A. Savka (2019). ‘Improved Genome of Agrobacterium Radiobacter Type Strain Provides New Taxonomic Insight into Agrobacterium Genomospecies 4’. In: *PeerJ* 7, e6366 (page [76](#)).
335. Tartoff, K. D. and C. A. Hobbs (n.d.). ‘Improved Media for Growing Plasmid and Cosmid Clones’. In: *Bethesda Research Laboratories Focus* 9.12 () (page [77](#)).
336. Yoo, Sang-Dong, Young-Hee Cho and Jen Sheen (2007). ‘*Arabidopsis* Mesophyll Protoplasts: A Versatile Cell System for Transient Gene Expression Analysis’. In: *Nature Protocols* 2.7, pp. 1565–1572 (page [78](#)).

337. Granger, Brian E. and Fernando Pérez (2021). ‘Jupyter: Thinking and Storytelling With Code and Data’. In: *Computing in Science & Engineering* 23.2, pp. 7–14 (pages [79](#), [88](#), [89](#), [93](#), [100](#)).
338. Lobet, Guillaume, Loïc Pagès and Xavier Draye (2011). ‘A Novel Image-Analysis Toolbox Enabling Quantitative Analysis of Root System Architecture’. In: *Plant Physiology* 157.1, pp. 29–39 (pages [79](#), [169](#)).
339. Schindelin, Johannes et al. (2012). ‘Fiji: An Open-Source Platform for Biological-Image Analysis’. In: *Nature Methods* 9.7 (7), pp. 676–682 (pages [79](#), [169](#)).
340. McKinney, Wes (2010). ‘Data Structures for Statistical Computing in Python’. In: *Proceedings of the 9th Python in Science Conference*, pp. 56–61 (pages [79](#), [88](#)).
341. Seabold, Skipper and Josef Perktold (2010). ‘Statsmodels: Econometric and Statistical Modeling with Python’. In: *Proceedings of the 9th Python in Science Conference*, pp. 92–96 (page [79](#)).
342. Vallat, Raphael (2018). ‘Pingouin: Statistics in Python’. In: *Journal of Open Source Software* 3.31, p. 1026 (pages [79](#), [95](#)).
343. Bedre, Renesh (2021). *Reneshbedre/Bioinfokit: Bioinformatics Data Analysis and Visualization Toolkit*. Version 1.0.5. Zenodo (page [80](#)).
344. Tukey, John W. (1949). ‘Comparing Individual Means in the Analysis of Variance’. In: *Biometrics* 5.2, p. 99 (pages [80](#), [138–141](#), [172](#), [176](#), [178](#), [181](#), [184](#)).
345. Waskom, Michael L. (2021). ‘Seaborn: Statistical Data Visualization’. In: *Journal of Open Source Software* 6.60, p. 3021 (pages [80](#), [89](#), [95](#)).
346. Maussion, Gilles et al. (2021). ‘Auto-qPCR; a Python-Based Web App for Automated and Reproducible Analysis of qPCR Data’. In: *Scientific Reports* 11.1 (1), p. 21293 (page [88](#)).
347. Shapiro, S. S. and M. B. Wilk (1965). ‘An Analysis of Variance Test for Normality (Complete Samples)’. In: *Biometrika* 52.3-4, pp. 591–611 (pages [89](#), [95](#)).
348. Levene, Howard (1960). ‘Robust Tests for Equality of Variances’. In: *Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling*. Ed. by Ingram Olkin et al. Vol. 2. Stanford University Press, pp. 278–292 (pages [89](#), [95](#)).

349. Welch, B. L. (1947). ‘The Generalization of ‘Student’s’ Problem When Several Different Population Variances Are Involved’. In: *Biometrika* 34.1-2, pp. 28–35 (pages [89](#), [116–118](#), [120](#), [122](#), [124](#), [126](#), [128](#), [130](#), [132](#), [134](#), [136](#), [172](#), [176](#), [178](#), [181](#), [184](#), [229](#), [230](#), [234–242](#), [245](#)).
350. Potter, Kevin C. et al. (2018). ‘Cytokinin Modulates Context-Dependent Chromatin Accessibility through the Type-B Response Regulators’. In: *Nature Plants* 4.12 (12), pp. 1102–1111 (pages [89](#), [98](#), [100](#), [109](#), [111](#), [120](#), [122](#), [124](#), [126](#), [128](#), [130](#), [132](#), [138–141](#), [160](#), [199](#), [202](#), [226](#), [268](#)).
351. Concordet, Jean-Paul and Maximilian Haeussler (2018). ‘CRISPOR: Intuitive Guide Selection for CRISPR/Cas9 Genome Editing Experiments and Screens’. In: *Nucleic Acids Research* 46.W1, W242–W245 (pages [89](#), [160](#), [185](#), [226](#), [227](#)).
352. Howe, Kevin L et al. (2020). ‘Ensembl Genomes 2020—Enabling Non-Vertebrate Genomic Research’. In: *Nucleic Acids Research* 48.D1, pp. D689–D695 (pages [89](#), [96](#)).
353. Team, Python Core (2020). *Python: A Dynamic, Open Source Programming Language*. Version 3. Python Software Foundation. (pages [89](#), [95](#)).
354. Doench, John G. et al. (2016). ‘Optimized sgRNA Design to Maximize Activity and Minimize Off-Target Effects of CRISPR-Cas9’. In: *Nature Biotechnology* 34.2 (2), pp. 184–191 (pages [89](#), [158](#), [160](#)).
355. Pereira, Filipa et al. (2015). ‘Pydna: A Simulation and Documentation Tool for DNA Assembly Strategies Using Python’. In: *BMC Bioinformatics* 16.1, p. 142 (page [89](#)).
356. Ewing, Brent and Phil Green (1998). ‘Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities’. In: *Genome Research* 8.3, pp. 186–194 (page [92](#)).
357. Slater, Guy St C. and Ewan Birney (2005). ‘Automated Generation of Heuristics for Biological Sequence Comparison’. In: *BMC Bioinformatics* 6.1, p. 31 (page [92](#)).
358. Martin, Marcel (2011). ‘Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads’. In: *EMBnet.journal* 17.1 (1), pp. 10–12 (page [92](#)).

359. Cock, Peter J. A. et al. (2009). ‘Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics’. In: *Bioinformatics* 25.11, pp. 1422–1423 (pages [93](#), [100](#)).
360. Clement, Kendell et al. (2019). ‘CRISPResso2 Provides Accurate and Rapid Genome Editing Sequence Analysis’. In: *Nature Biotechnology* 37.3 (3), pp. 224–226 (page [93](#)).
361. Langmead, Ben and Steven L. Salzberg (2012). ‘Fast Gapped-Read Alignment with Bowtie 2’. In: *Nature Methods* 9.4 (4), pp. 357–359 (page [93](#)).
362. Quinlan, Aaron R. and Ira M. Hall (2010). ‘BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features’. In: *Bioinformatics* 26.6, pp. 841–842 (pages [93](#), [96](#), [98](#)).
363. Cheng, Chia-Yi et al. (2017). ‘Araport11: A Complete Reannotation of the *Arabidopsis Thaliana* Reference Genome’. In: *The Plant Journal* 89.4, pp. 789–804 (pages [94](#), [166](#), [196](#)).
364. Grant, Charles E., Timothy L. Bailey and William Stafford Noble (2011). ‘FIMO: Scanning for Occurrences of a given Motif’. In: *Bioinformatics* 27.7, pp. 1017–1018 (pages [94](#), [96](#), [99](#), [102](#), [107](#), [109](#), [150](#), [264](#), [265](#)).
365. Kruskal, William H. and W. Allen Wallis (1952). ‘Use of Ranks in One-Criterion Variance Analysis’. In: *Journal of the American Statistical Association* 47.260, pp. 583–621 (page [95](#)).
366. Terpilowski, Maksim A. (2019). ‘Scikit-Posthocs: Pairwise Multiple Comparison Tests in Python’. In: *Journal of Open Source Software* 4.36, p. 1169 (page [95](#)).
367. Dunn, Olive Jean (1964). ‘Multiple Comparisons Using Rank Sums’. In: *Technometrics* 6.3, p. 241 (page [95](#)).
368. Dale, Ryan K., Brent S. Pedersen and Aaron R. Quinlan (2011). ‘Pybedtools: A Flexible Python Library for Manipulating Genomic Datasets and Annotations’. In: *Bioinformatics* 27.24, pp. 3423–3424 (page [96](#)).
369. Neph, Shane et al. (2012). ‘BEDOPS: High-Performance Genomic Feature Operations’. In: *Bioinformatics* 28.14, pp. 1919–1920 (page [96](#)).

370. Bailey, Timothy L. et al. (2009). ‘MEME Suite: Tools for Motif Discovery and Searching’. In: *Nucleic Acids Research* 37 (suppl_2), W202–W208 (pages 96, 100).
371. Benjamini, Yoav and Yosef Hochberg (1995). ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57.1, pp. 289–300 (pages 96, 100, 199).
372. Czechowski, Tomasz et al. (2005). ‘Genome-Wide Identification and Testing of Superior Reference Genes for Transcript Normalization in Arabidopsis’. In: *PLANT PHYSIOLOGY* 139.1, pp. 5–17 (pages 96, 97, 195–197, 217, 218).
373. Yanai, Itai et al. (2005). ‘Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification’. In: *Bioinformatics* 21.5, pp. 650–659 (pages 96, 97).
374. Kryuchkova-Mostacci, Nadezda and Marc Robinson-Rechavi (2017). ‘A Benchmark of Gene Expression Tissue-Specificity Metrics’. In: *Briefings in Bioinformatics* 18.2, pp. 205–214 (page 97).
375. Rousseeuw, Peter J. (1987). ‘Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis’. In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65 (pages 99, 212).
376. Dreos, René et al. (2017). ‘The Eukaryotic Promoter Database in Its 30th Year: Focus on Non-Vertebrate Organisms’. In: *Nucleic Acids Research* 45.D1, pp. D51–D55 (page 99).
377. Heger, A. et al. (2013). ‘GAT: A Simulation Framework for Testing the Association of Genomic Intervals’. In: *Bioinformatics* 29.16, pp. 2046–2048 (pages 99, 214, 278).
378. Klopfenstein, D. V. et al. (2018). ‘GOATOOLS: A Python Library for Gene Ontology Analyses’. In: *Scientific Reports* 8.1 (page 99).
379. Fisher, R. A. (1922). ‘On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P’. In: *Journal of the Royal Statistical Society* 85.1, p. 87 (pages 100, 199).

380. Yu, Guangchuang et al. (2012). ‘clusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters’. In: *OMICS: A Journal of Integrative Biology* 16.5, pp. 284–287 (page 100).
381. {R Core Team} (2022). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing (page 100).
382. Kanehisa, M. (2000). ‘KEGG: Kyoto Encyclopedia of Genes and Genomes’. In: *Nucleic Acids Research* 28.1, pp. 27–30 (page 100).
383. Zulkower, Valentin and Susan Rosser (2020). ‘DNA Features Viewer: A Sequence Annotation Formatting and Plotting Library for Python’. In: *Bioinformatics* 36.15. Ed. by Anthony Mathelier, pp. 4350–4352 (page 100).
384. Lopez-Delisle, Lucille et al. (2021). ‘pyGenomeTracks: Reproducible Plots for Multivariate Genomic Datasets’. In: *Bioinformatics* 37.3, pp. 422–423 (page 100).
385. Hunter, John D. (2007). ‘Matplotlib: A 2D Graphics Environment’. In: *Computing in Science & Engineering* 9.3, pp. 90–95 (page 100).
386. Kunzmann, Patrick and Kay Hamacher (2018). ‘Biotite: A Unifying Open Source Computational Biology Framework in Python’. In: *BMC Bioinformatics* 19.1, p. 346 (page 100).
387. Chang, Katherine Noelani et al. (2013). ‘Temporal Transcriptional Response to Ethylene Gas Drives Growth Hormone Cross-Regulation in Arabidopsis’. In: *eLife* 2, e00675 (page 102).
388. Martin, O.C., A. Krzywicki and M. Zagorski (2016). ‘Drivers of Structural Features in Gene Regulatory Networks: From Biophysical Constraints to Biological Function’. In: *Physics of Life Reviews* 17, pp. 124–158 (page 102).
389. Ambrosini, Giovanna et al. (2020). ‘Insights Gained from a Comprehensive All-against-All Transcription Factor Binding Motif Benchmarking Study’. In: *Genome Biology* 21.1, p. 114 (pages 102, 186).
390. Jayaram, Narayan, Daniel Usvyat and Andrew C. R. Martin (2016). ‘Evaluating Tools for Transcription Factor Binding Site Prediction’. In: *BMC Bioinformatics* 17.1, p. 547 (page 102).

391. Hellman, Lance M. and Michael G. Fried (2007). ‘Electrophoretic Mobility Shift Assay (EMSA) for Detecting Protein–Nucleic Acid Interactions’. In: *Nature Protocols* 2.8 (8), pp. 1849–1861 (page [103](#)).
392. Van den Broeck, Lisa et al. (2017). ‘From Network to Phenotype: The Dynamic Wiring of an Arabidopsis Transcriptional Network Induced by Osmotic Stress’. In: *Molecular Systems Biology* 13.12, p. 961 (pages [103](#), [154](#)).
393. Konishi, Mineko and Shuichi Yanagisawa (2013). ‘Arabidopsis NIN-like Transcription Factors Have a Central Role in Nitrate Signalling’. In: *Nature Communications* 4.1, pp. 1–9 (pages [103](#), [142](#), [148](#), [231](#), [236](#), [249](#), [253](#)).
394. Varala, Kranthi et al. (2018). ‘Temporal Transcriptional Logic of Dynamic Regulatory Networks Underlying Nitrogen Signaling and Use in Plants’. In: *Proceedings of the National Academy of Sciences* 115.25, p. 6494 (page [103](#)).
395. Alvarez, José M. et al. (2019). ‘Local Changes in Chromatin Accessibility and Transcriptional Networks Underlying the Nitrate Response in Arabidopsis Roots’. In: *Molecular Plant* (page [103](#)).
396. Konishi, Mineko and Shuichi Yanagisawa (2010). ‘Identification of a Nitrate-Responsive Cis-Element in the Arabidopsis NIR1 Promoter Defines the Presence of Multiple Cis-Regulatory Elements for Nitrogen Response’. In: *The Plant Journal* 63.2, pp. 269–282 (pages [107](#), [139](#), [146](#), [152](#), [231](#), [243](#), [249](#), [253](#)).
397. Pierre-Jerome, Edith et al. (2016). ‘Functional Analysis of Molecular Interactions in Synthetic Auxin Response Circuits’. In: *Proceedings of the National Academy of Sciences* 113.40, pp. 11354–11359 (page [108](#)).
398. Cancé, Coralie et al. (2022). ‘Auxin Response Factors Are Keys to the Many Auxin Doors’. In: *New Phytologist* 235.2, pp. 402–419 (pages [108](#), [115](#), [148](#), [150](#)).
399. Luo, Jie, Jing-Jing Zhou and Jin-Zhi Zhang (2018). ‘Aux/IAA Gene Family in Plants: Molecular Structure, Regulation, and Function’. In: *International Journal of Molecular Sciences* 19.1 (1), p. 259 (pages [115](#), [154](#)).

400. Roosjen, Mark, Sébastien Paque and Dolf Weijers (2018). ‘Auxin Response Factors: Output Control in Auxin Biology’. In: *Journal of Experimental Botany* 69.2, pp. 179–188 (pages [115](#), [143](#), [146](#), [147](#)).
401. Lv, Bingsheng et al. (2020). ‘Non-canonical AUX / IAA Protein IAA 33 Competes with Canonical AUX / IAA Repressor IAA 5 to Negatively Regulate Auxin Signaling’. In: *The EMBO Journal* 39.1 (pages [115](#), [143](#), [154](#)).
402. R  th, Jochen, Rudolf J. Schweyen and Heribert Hirt (1994). ‘The Plant Transcription Factor TGA1 Stimulates Expression of the CaMV 35S Promoter in *Saccharomyces Cerevisiae*’. In: *Plant Molecular Biology* 25.2, pp. 323–328 (page [116](#)).
403. Lam, E et al. (1989). ‘Site-Specific Mutations Alter in Vitro Factor Binding and Change Promoter Expression Pattern in Transgenic Plants.’ In: *Proceedings of the National Academy of Sciences* 86.20, pp. 7890–7894 (page [116](#)).
404. Lam, Eric and Yolanda Kam-Po Lam (1995). ‘Binding Site Requirements and Differential Representation of TGA Factors in Nuclear ASF-1 Activity’. In: *Nucleic Acids Research* 23.18, pp. 3778–3785 (page [116](#)).
405. Yu, Lin-Hui et al. (2016a). ‘Overexpression of Arabidopsis NLP7 Improves Plant Growth under Both Nitrogen-Limiting and -Sufficient Conditions by Enhancing Nitrogen and Carbon Assimilation’. In: *Scientific Reports* 6.1 (1), p. 27795 (pages [142](#), [186](#), [187](#)).
406. Lin, Jia-Hui, Lin-Hui Yu and Cheng-Bin Xiang (2019). ‘ARABIDOPSIS NITRATE REGULATED 1 Acts as a Negative Modulator of Seed Germination by Activating ABI3 Expression’. In: *New Phytologist* 0 (ja) (page [142](#)).
407. Jia, Yuebin et al. (2020b). ‘PIFs Coordinate Shade Avoidance by Inhibiting Auxin Repressor ARF18 and Metabolic Regulator QQS’. In: *New Phytologist* 228.2, pp. 609–621 (pages [143](#), [146](#), [159](#), [230](#), [259](#)).
408. Maki, Hiromasa et al. (2019). ‘ANAC032 Regulates Root Growth through the MYB30 Gene Regulatory Network’. In: *Scientific Reports* 9.1, pp. 1–13 (page [146](#)).
409. ‘Positive Feedback, Bistability and Memory’ (2020). In: Uri Alon. *An Introduction to Systems Biology. Design Principles of Biological Circuits*. CRC Press (page [147](#)).

410. ‘Autoregulation’ (2020). In: Uri Alon. *An Introduction to Systems Biology. Design Principles of Biological Circuits*. CRC Press (page [147](#)).
411. Vernoux, Teva et al. (2011). ‘The Auxin Signalling Network Translates Dynamic Input into Robust Patterning at the Shoot Apex’. In: *Molecular Systems Biology* 7.1, p. 508 (page [147](#)).
412. Alexandre, Cristina M et al. (2018). ‘Complex Relationships between Chromatin Accessibility, Sequence Divergence, and Gene Expression in Arabidopsis Thaliana’. In: *Molecular Biology and Evolution* 35.4, pp. 837–854 (page [147](#)).
413. Boer, D. Roeland et al. (2014). ‘Structural Basis for DNA Binding Specificity by the Auxin-Dependent ARF Transcription Factors’. In: *Cell* 156.3, pp. 577–589 (page [148](#)).
414. Stigliani, Arnaud et al. (2019). ‘Capturing Auxin Response Factors Syntax Using DNA Binding Models’. In: *Molecular Plant* 12.6, pp. 822–832 (page [148](#)).
415. Li, Jian-Feng et al. (2011). ‘Large-Scale Protein-Protein Interaction Analysis in Arabidopsis Mesophyll Protoplasts by Split Firefly Luciferase Complementation’. In: *PLOS ONE* 6.11, e27364 (pages [148](#), [154](#)).
416. Jolma, Arttu et al. (2015). ‘DNA-dependent Formation of Transcription Factor Pairs Alters Their Binding Specificity’. In: *Nature* 527.7578, pp. 384–388 (page [150](#)).
417. Xu, Mengxue et al. (2021). ‘Stochastic Gene Expression Drives Mesophyll Protoplast Regeneration’. In: *Science Advances* 7.33, eabg8466 (pages [151](#), [248](#)).
418. Demirer, Gozde S. et al. (2019). ‘High Aspect Ratio Nanomaterials Enable Delivery of Functional Genetic Material without DNA Integration in Mature Plants’. In: *Nature Nanotechnology* 14.5, pp. 456–464 (page [151](#)).
419. Doyle, Cara et al. (2019). ‘A Simple Method for Spray-on Gene Editing in Planta’. In: *bioRxiv* (page [151](#)).
420. Hajiahmadi, Zahra et al. (2019). ‘Enhancement of Tomato Resistance to Tuta Absoluta Using a New Efficient Mesoporous Silica Nanoparticle-Mediated Plant Transient Gene Expression Approach’. In: *Scientia Horticulturae* 243, pp. 367–375 (page [151](#)).

421. Lenaerts, Bert, Bertrand C.Y. Collard and Matty Demont (2019). 'Review: Improving Global Food Security through Accelerated Plant Breeding'. In: *Plant Science* 287, p. 110207 (page 156).
422. Liseron-Monfils, Christophe and Doreen Ware (2015). 'Revealing Gene Regulation and Associations through Biological Networks'. In: *Current Plant Biology*. Transcriptome Networks (Including Y1H and Single-Cell or Cell Type-Specific Profiles) 3–4, pp. 30–39 (page 157).
423. Zhang, Yi et al. (2016). 'Efficient and Transgene-Free Genome Editing in Wheat through Transient Expression of CRISPR/Cas9 DNA or RNA'. In: *Nature Communications* 7.1 (1), p. 12617 (pages 157, 260).
424. Bandyopadhyay, Anindya et al. (2020). 'CRISPR-Cas12a (Cpf1): A Versatile Tool in the Plant Genome Editing Tool Box for Agricultural Advancement'. In: *Frontiers in Plant Science* 11 (pages 157, 260).
425. Ma, Xiaonan et al. (2020). 'Highly Efficient DNA-free Plant Genome Editing Using Virally Delivered CRISPR–Cas9'. In: *Nature Plants* 6.7 (7), pp. 773–779 (pages 157, 260).
426. Knott, Gavin J. and Jennifer A. Doudna (2018). 'CRISPR-Cas Guides the Future of Genetic Engineering'. In: *Science* 361.6405, pp. 866–869 (page 157).
427. Belhaj, Khaoula et al. (2015). 'Editing Plant Genomes with CRISPR/Cas9'. In: *Current Opinion in Biotechnology*. Food Biotechnology • Plant Biotechnology 32, pp. 76–84 (page 157).
428. Komor, Alexis C. et al. (2016). 'Programmable Editing of a Target Base in Genomic DNA without Double-Stranded DNA Cleavage'. In: *Nature* 533.7603 (7603), pp. 420–424 (page 157).
429. Gaudelli, Nicole M. et al. (2017). 'Programmable Base Editing of A • T to G • C in Genomic DNA without DNA Cleavage'. In: *Nature* 551.7681 (7681), pp. 464–471 (page 157).
430. Sretenovic, Simon and Yiping Qi (2022). 'Plant Prime Editing Goes Prime'. In: *Nature Plants* 8.1 (1), pp. 20–22 (page 157).

431. Ahmad, Niaz, Muhammad Jawad Akbar Awan and Shahid Mansoor (2022). 'Improving Editing Efficiency of Prime Editor in Plants'. In: *Trends in Plant Science* 0.0 (page 157).
432. Meyer, Rachel S. and Michael D. Purugganan (2013). 'Evolution of Crop Species: Genetics of Domestication and Diversification'. In: *Nature Reviews Genetics* 14.12 (12), pp. 840–852 (page 157).
433. Gaillochet, Christophe, Ward Develtere and Thomas B. Jacobs (2020). 'CRISPR Screens in Plants: Approaches, Guidelines, and Future Prospects'. In: *The Plant Cell* (page 158).
434. Morgens, David W. et al. (2017). 'Genome-Scale Measurement of off-Target Activity Using Cas9 Toxicity in High-Throughput Screens'. In: *Nature Communications* 8.1 (1), p. 15178 (page 158).
435. Viswanatha, Raghuvir et al. (2018). 'Pooled Genome-Wide CRISPR Screening for Basal and Context-Specific Fitness Gene Essentiality in Drosophila Cells'. In: *eLife* 7. Ed. by Michael Boutros and Patricia J Wittkopp, e36333 (page 158).
436. Wei, Jin et al. (2021). 'Genome-Wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection'. In: *Cell* 184.1, 76–91.e13 (page 158).
437. Wang, Tim et al. (2015b). 'Identification and Characterization of Essential Genes in the Human Genome'. In: *Science* 350.6264, pp. 1096–1101 (page 158).
438. Gonçalves, Emanuel et al. (2021). 'Minimal Genome-Wide Human CRISPR-Cas9 Library'. In: *Genome Biology* 22.1, p. 40 (page 158).
439. Yu, Chuanzhao et al. (2022). 'Establishment of a Pig CRISPR/Cas9 Knockout Library for Functional Gene Screening in Pig Cells'. In: *Biotechnology Journal* 17.7, p. 2100408 (page 158).
440. Meng, Xiangbing et al. (2017). 'Construction of a Genome-Wide Mutant Library in Rice Using CRISPR/Cas9'. In: *Molecular Plant* 10.9, pp. 1238–1241 (page 158).
441. Lu, Yuming et al. (2017). 'Genome-Wide Targeted Mutagenesis in Rice Using the CRISPR/Cas9 System'. In: *Molecular Plant* 10.9, pp. 1242–1245 (page 158).

442. Liu, Hai-Jun et al. (2020). ‘High-Throughput CRISPR/Cas9 Mutagenesis Streamlines Trait Gene Identification in Maize[OPEN]’. In: *The Plant Cell* 32.5, pp. 1397–1413 (page 158).
443. Jacobs, Thomas B. et al. (2017). ‘Generation of a Collection of Mutant Tomato Lines Using Pooled CRISPR Libraries’. In: *Plant Physiology* 174.4, pp. 2023–2037 (page 158).
444. Bai, Mengyan et al. (2020). ‘Generation of a Multiplex Mutagenesis Population via Pooled CRISPR-Cas9 in Soya Bean’. In: *Plant Biotechnology Journal* 18.3, pp. 721–731 (page 158).
445. Butt, Haroon et al. (2019). ‘CRISPR Directed Evolution of the Spliceosome for Resistance to Splicing Inhibitors’. In: *Genome Biology* 20.1, p. 73 (page 158).
446. Kuang, Yongjie et al. (2020). ‘Base-Editing-Mediated Artificial Evolution of OsALS1 In Planta to Develop Novel Herbicide-Tolerant Rice Germplasms’. In: *Molecular Plant* 13.4, pp. 565–572 (page 158).
447. Li, Chao et al. (2020a). ‘Targeted, Random Mutagenesis of Plant Genes with Dual Cytosine and Adenine Base Editors’. In: *Nature Biotechnology* 38.7 (7), pp. 875–882 (page 158).
448. Somssich, Marc et al. (2016). ‘CLAVATA-WUSCHEL Signaling in the Shoot Meristem’. In: *Development* 143.18, pp. 3238–3248 (page 158).
449. Wang, Wei et al. (2022b). ‘Multiplexed Promoter and Gene Editing in Wheat Using a Virus-based Guide RNA Delivery System’. In: *Plant Biotechnology Journal*, pbi.13910 (pages 158, 259).
450. Iida, Midori et al. (2020). ‘A Simple and Practical Workflow for Genotyping of CRISPR–Cas9-based Knockout Phenotypes Using Multiplexed Amplicon Sequencing’. In: *Genes to Cells* 25.7, pp. 498–509 (pages 158, 162).
451. Yu, Chun-Ping, Jinn-Jy Lin and Wen-Hsiung Li (2016b). ‘Positional Distribution of Transcription Factor Binding Sites in *Arabidopsis Thaliana*’. In: *Scientific Reports* 6, p. 25164 (page 160).

452. Pauwels, Laurens et al. (2018). ‘A Dual sgRNA Approach for Functional Genomics in *Arabidopsis Thaliana*’. In: *G3 Genes/Genomes/Genetics* 8.8, pp. 2603–2615 (pages [160](#), [185](#)).
453. Durr, Julius et al. (2018). ‘Highly Efficient Heritable Targeted Deletions of Gene Clusters and Non-Coding Regulatory Regions in *Arabidopsis* Using CRISPR/Cas9’. In: *Scientific Reports* 8.1 (1), p. 4443 (page [160](#)).
454. Ordon, Jana et al. (2017). ‘Generation of Chromosomal Deletions in Dicotyledonous Plants Employing a User-Friendly Genome Editing Toolkit’. In: *The Plant Journal* 89.1, pp. 155–168 (page [160](#)).
455. Feng, Zhengyan et al. (2014). ‘Multigeneration Analysis Reveals the Inheritance, Specificity, and Patterns of CRISPR/Cas-induced Gene Modifications in *Arabidopsis*’. In: *Proceedings of the National Academy of Sciences* 111.12, pp. 4632–4637 (page [163](#)).
456. Okushima, Yoko et al. (2005). ‘Functional Genomic Analysis of the *AUXIN RESPONSE FACTOR* Gene Family Members in *Arabidopsis Thaliana* : Unique and Overlapping Functions of *ARF7* and *ARF19*’. In: *The Plant Cell* 17.2, pp. 444–463 (pages [179](#), [190](#)).
457. Safari, Fatemeh et al. (2019). ‘CRISPR Cpf1 Proteins: Structure, Function and Implications for Genome Editing’. In: *Cell & Bioscience* 9.1, p. 36 (page [185](#)).
458. Wolabu, Tezera W. et al. (2020). ‘Improving the Genome Editing Efficiency of CRISPR/Cas9 in *Arabidopsis* and *Medicago Truncatula*’. In: *Planta* 252.2, p. 15 (page [185](#)).
459. Pan, Changtian et al. (2016). ‘CRISPR/Cas9-mediated Efficient and Heritable Targeted Mutagenesis in Tomato Plants in the First and Later Generations’. In: *Scientific Reports* 6.1 (1), p. 24765 (page [185](#)).
460. Zhang, Ning et al. (2020b). ‘Generation and Molecular Characterization of CRISPR/Cas9-Induced Mutations in 63 Immunity-Associated Genes in Tomato Reveals Specificity and a Range of Gene Modifications’. In: *Frontiers in Plant Science* 11 (page [185](#)).

461. Knoll, Alexander, Friedrich Fauser and Holger Puchta (2014). 'DNA Recombination in Somatic Plant Cells: Mechanisms and Evolutionary Consequences'. In: *Chromosome Research* 22.2, pp. 191–201 (page 185).
462. Khodaverdian, Varandt Y. et al. (2017). 'Secondary Structure Forming Sequences Drive SD-MMEJ Repair of DNA Double-Strand Breaks'. In: *Nucleic Acids Research* 45.22, pp. 12848–12861 (page 185).
463. Baxter, Laura et al. (2012). 'Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants'. In: *The Plant Cell* 24.10, pp. 3949–3965 (page 185).
464. Joly-Lopez, Zoé et al. (2017). 'Abiotic Stress Phenotypes Are Associated with Conserved Genes Derived from Transposable Elements'. In: *Frontiers in Plant Science* 8 (page 186).
465. Monfared, Mona M. et al. (2011). 'Overlapping and Antagonistic Activities of BASIC PENTACYSTEINE Genes Affect a Range of Developmental Processes in Arabidopsis'. In: *The Plant Journal* 66.6, pp. 1020–1031 (pages 186, 188).
466. Mathieu, Olivier et al. (2003). 'Identification and Characterization of Transcription Factor IIIA and Ribosomal Protein L5 from Arabidopsis Thaliana'. In: *Nucleic Acids Research* 31.9, pp. 2424–2433 (page 186).
467. Khan, Nafees A (2014). 'The Ethylene: From Senescence Hormone to Key Player in Plant Metabolism'. In: *Journal of Plant Biochemistry & Physiology* 2.2 (pages 187, 188).
468. Levy, Yaron Y. et al. (2002). 'Multiple Roles of Arabidopsis VRN1 in Vernalization and Flowering Time Control'. In: *Science* 297.5579, pp. 243–246 (page 187).
469. Heo, Jae Bok, Sibum Sung and Sarah M. Assmann (2012). 'Ca²⁺-Dependent GTPase, Extra-large G Protein 2 (XLG2), Promotes Activation of DNA-binding Protein Related to Vernalization 1 (RTV1), Leading to Activation of Floral Integrator Genes and Early Flowering in Arabidopsis*'. In: *Journal of Biological Chemistry* 287.11, pp. 8242–8253 (page 187).

470. Miyashima, Shunsuke et al. (2019). ‘Mobile PEAR Transcription Factors Integrate Positional Cues to Prime Cambial Growth’. In: *Nature* 565.7740 (7740), pp. 490–494 (page 187).
471. Krishnaswamy, Sowmya et al. (2011). ‘Functional Characterization of Four APETALA2-family Genes (RAP2.6, RAP2.6L, DREB19 and DREB26) in Arabidopsis’. In: *Plant Molecular Biology* 75.1, pp. 107–127 (page 188).
472. Mu, Xiaohuan and Jie Luo (2019). ‘Evolutionary Analyses of NIN-like Proteins in Plants and Their Roles in Nitrate Signaling’. In: *Cellular and Molecular Life Sciences* 76.19, pp. 3753–3764 (page 188).
473. Zhang, Liqing and Wen-Hsiung Li (2004). ‘Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes’. In: *Molecular Biology and Evolution* 21.2, pp. 236–239 (page 192).
474. Butte, Atul J., Victor J. Dzau and Susan B. Glueck (2001). ‘Further Defining Housekeeping, or “Maintenance,” Genes Focus on “A Compendium of Gene Expression in Normal Human Tissues”’. In: *Physiological Genomics* 7.2, pp. 95–96 (page 192).
475. Schug, Jonathan et al. (2005). ‘Promoter Features Related to Tissue Specificity as Measured by Shannon Entropy’. In: *Genome Biology* 6.4, R33 (page 192).
476. Araújo, Ilka Schultheiß et al. (2017). ‘Stochastic Gene Expression in Arabidopsis Thaliana’. In: *Nature Communications* 8.1 (1), p. 2132 (pages 192, 219).
477. Zhang, Xiaoyu et al. (2006). ‘Genome-Wide High-Resolution Mapping and Functional Analysis of DNA Methylation in Arabidopsis’. In: *Cell* 126.6, pp. 1189–1201 (page 193).
478. Takuno, Shohei and Brandon S. Gaut (2012). ‘Body-Methylated Genes in Arabidopsis Thaliana Are Functionally Important and Evolve Slowly’. In: *Molecular Biology and Evolution* 29.1, pp. 219–227 (page 193).
479. Aceituno, Felipe F. et al. (2008). ‘The Rules of Gene Expression in Plants: Organ Identity and Gene Body Methylation Are Key Factors for Regulation of Gene Expression in Arabidopsis Thaliana’. In: *BMC Genomics* 9.1, p. 438 (page 193).

480. Zuo, Yong-Chun and Qian-Zhong Li (2011). ‘Identification of TATA and TATA-less Promoters in Plant Genomes by Integrating Diversity Measure, GC-Skew and DNA Geometric Flexibility’. In: *Genomics* 97.2, pp. 112–120 (page 193).
481. Yamamoto, Yoshiharu Y. et al. (2009). ‘Heterogeneity of Arabidopsis Core Promoters Revealed by High-Density TSS Analysis’. In: *The Plant Journal: For Cell and Molecular Biology* 60.2, pp. 350–362 (page 193).
482. Yamamoto, Yoshiharu Y. et al. (2011). ‘Characteristics of Core Promoter Types with Respect to Gene Structure and Expression in Arabidopsis Thaliana’. In: *DNA Research* 18.5, pp. 333–342 (page 193).
483. Einarsson, Hjorleifur et al. (2022). ‘Promoter Sequence and Architecture Determine Expression Variability and Confer Robustness to Genetic Variants’. In: *eLife* 11. Ed. by Eduardo Eyras, e80943 (pages 193, 194, 215, 218).
484. Jones, Peter A. (2012). ‘Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond’. In: *Nature Reviews Genetics* 13.7 (7), pp. 484–492 (page 193).
485. Vinogradov, A. E. and O. V. Anatskaya (2017). ‘DNA Helix: The Importance of Being AT-rich’. In: *Mammalian Genome* 28.9, pp. 455–464 (pages 193, 216).
486. Wei, Kai, Lei Ma and Tingting Zhang (2019). ‘Characterization of Gene Promoters in Pig: Conservative Elements, Regulatory Motifs and Evolutionary Trend’. In: *PeerJ* 7, e7204 (pages 193, 194, 216).
487. Vinogradov, Alexander E. (2001). ‘Bendable Genes of Warm-blooded Vertebrates’. In: *Molecular Biology and Evolution* 18.12, pp. 2195–2200 (page 193).
488. — (2003). ‘DNA Helix: The Importance of Being GC-rich’. In: *Nucleic Acids Research* 31.7, pp. 1838–1844 (page 193).
489. — (2005). ‘Noncoding DNA, Isochores and Gene Expression: Nucleosome Formation Potential’. In: *Nucleic Acids Research* 33.2, pp. 559–563 (page 193).
490. Liu, Ming-Jung et al. (2015). ‘Determinants of Nucleosome Positioning and Their Influence on Plant Gene Expression’. In: *Genome Research* 25.8, pp. 1182–1195 (page 193).

491. Mukhopadhyay, Pamela, Surajit Basak and Tapash Chandra Ghosh (2008). ‘Differential Selective Constraints Shaping Codon Usage Pattern of Housekeeping and Tissue-specific Homologous Genes of Rice and Arabidopsis’. In: *DNA Research* 15.6, pp. 347–356 (page [193](#)).
492. Armisen, David, Alain Lecharny and Sébastien Aubourg (2008). ‘Unique Genes in Plants: Specificities and Conserved Features throughout Evolution’. In: *BMC Evolutionary Biology* 8.1, p. 280 (page [193](#)).
493. Walther, Dirk, Roman Brunnemann and Joachim Selbig (2007). ‘The Regulatory Code for Transcriptional Response Diversity and Its Relation to Genome Structural Properties in *A. Thaliana*’. In: *PLOS Genetics* 3.2, e11 (pages [194](#), [218](#), [250](#)).
494. Das, Sanjukta and Manju Bansal (2019). ‘Variation of Gene Expression in Plants Is Influenced by Gene Architecture and Structural Properties of Promoters’. In: *PLOS ONE* 14.3, e0212678 (pages [194](#), [215](#), [216](#), [218](#), [219](#)).
495. Thomas, David et al. (2015). ‘DNA Entropy Reveals a Significant Difference in Complexity between Housekeeping and Tissue Specific Gene Promoters’. In: *Computational Biology and Chemistry* 58, pp. 19–24 (page [194](#)).
496. Cortijo, Sandra et al. (2019). ‘Widespread Inter-Individual Gene Expression Variability in *Arabidopsis Thaliana*’. In: *Molecular Systems Biology* 15.1, e8591 (pages [194](#), [216](#), [217](#)).
497. Sonawane, Abhijeet Rajendra et al. (2017). ‘Understanding Tissue-Specific Gene Regulation’. In: *Cell Reports* 21.4, pp. 1077–1088 (page [195](#)).
498. Dreos, René, Giovanna Ambrosini and Philipp Bucher (2016). ‘Influence of Rotational Nucleosome Positioning on Transcription Start Site Selection in Animal Promoters’. In: *PLOS Computational Biology* 12.10, e1005144 (pages [214](#), [278](#)).
499. Di Matteo, Adele et al. (2005). ‘Structural Basis for the Interaction between Pectin Methyltransferase and a Specific Inhibitor Protein’. In: *The Plant Cell* 17.3, pp. 849–858 (page [215](#)).

500. Golkaram, Mahdi et al. (2017). ‘The Role of Chromatin Density in Cell Population Heterogeneity during Stem Cell Differentiation’. In: *Scientific Reports* 7.1 (1), p. 13307 (pages [216](#), [217](#)).
501. Kaplan, Noam et al. (2009). ‘The DNA-encoded Nucleosome Organization of a Eukaryotic Genome’. In: *Nature* 458.7236 (7236), pp. 362–366 (page [216](#)).
502. Albert, Istvan et al. (2007). ‘Translational and Rotational Settings of H2A.Z Nucleosomes across the *Saccharomyces Cerevisiae* Genome’. In: *Nature* 446.7135 (7135), pp. 572–576 (page [216](#)).
503. Mavrich, Travis N. et al. (2008). ‘Nucleosome Organization in the *Drosophila* Genome’. In: *Nature* 453.7193 (7193), pp. 358–362 (page [216](#)).
504. Zhang, Lingang et al. (2004). ‘GC/AT-content Spikes as Genomic Punctuation Marks’. In: *Proceedings of the National Academy of Sciences* 101.48, pp. 16855–16860 (page [217](#)).
505. GHOSH, SOURAV et al. (2015). ‘Distinct Patterns of Epigenetic Marks and Transcription Factor Binding Sites across Promoters of Sense-Intronic Long Noncoding RNAs’. In: *Journal of Genetics* 94.1, pp. 17–25 (page [217](#)).
506. Dorrity, Michael W. et al. (2021). ‘The Regulatory Landscape of *Arabidopsis Thaliana* Roots at Single-Cell Resolution’. In: *Nature Communications* 12.1 (1), p. 3334 (page [218](#)).
507. Engström, Pär G. et al. (2007). ‘Genomic Regulatory Blocks Underlie Extensive Microsynteny Conservation in Insects’. In: *Genome Research* 17.12, pp. 1898–1908 (page [218](#)).
508. Carninci, Piero et al. (2006). ‘Genome-Wide Analysis of Mammalian Promoter Architecture and Evolution’. In: *Nature Genetics* 38.6, pp. 626–635 (page [218](#)).
509. Bae, Sang-Hun, Hyun Wook Han and Jisook Moon (2015). ‘Functional Analysis of the Molecular Interactions of TATA Box-Containing Genes and Essential Genes’. In: *PLOS ONE* 10.3, e0120848 (page [218](#)).

510. Fonseca, João Pedro et al. (2019). ‘A Toolkit for Rapid Modular Construction of Biological Circuits in Mammalian Cells’. In: *ACS Synthetic Biology* 8.11, pp. 2593–2606 (page 223).
511. Riglar, David T. and Pamela A. Silver (2018). ‘Engineering Bacteria for Diagnostic and Therapeutic Applications’. In: *Nature Reviews Microbiology* 16.4 (4), pp. 214–225 (page 223).
512. Brophy, Jennifer A. N. and Christopher A. Voigt (2014). ‘Principles of Genetic Circuit Design’. In: *Nature Methods* 11.5 (5), pp. 508–520 (page 223).
513. Ouwerkerk, Pieter B. et al. (2001). ‘Glucocorticoid-Inducible Gene Expression in Rice’. In: *Planta* 213.3, pp. 370–378 (page 223).
514. Zuo, Jianru, Qi-Wen Niu and Nam-Hai Chua (2000). ‘An Estrogen Receptor-Based Transactivator XVE Mediates Highly Inducible Gene Expression in Transgenic Plants’. In: *The Plant Journal* 24.2, pp. 265–273 (page 223).
515. Selma, Sara et al. (2019). ‘Strong Gene Activation in Plants with Genome-Wide Specificity Using a New Orthogonal CRISPR/Cas9-based Programmable Transcriptional Activator’. In: *Plant Biotechnology Journal* 17.9, pp. 1703–1705 (page 224).
516. Selma, Sara et al. (2022). ‘Custom-Made Design of Metabolite Composition in *N. Benthamiana* Leaves Using CRISPR Activators’. In: *Plant Biotechnology Journal* 20.8, pp. 1578–1590 (page 224).
517. Li, Zhenxiang, Xiangyu Xiong and Jian-Feng Li (2020b). ‘The Working Dead: Repurposing Inactive CRISPR-associated Nucleases as Programmable Transcriptional Regulators in Plants’. In: *aBIOTECH* 1.1, pp. 32–40 (pages 226, 227).
518. Konermann, Silvana et al. (2015). ‘Genome-Scale Transcriptional Activation by an Engineered CRISPR-Cas9 Complex’. In: *Nature* 517.7536, pp. 583–588 (page 226).
519. Maeda, Yoshie et al. (2018). ‘A NIGT1-centred Transcriptional Cascade Regulates Nitrate Signalling and Incorporates Phosphorus Starvation Signals in Arabidopsis’. In: *Nature Communications* 9.1, pp. 1–14 (pages 231, 251).

520. Schindler, U, H Beckmann and A R Cashmore (1992). ‘TGA1 and G-box Binding Factors: Two Distinct Classes of Arabidopsis Leucine Zipper Proteins Compete for the G-box-like Element TGACGTGG.’ In: *The Plant Cell* 4.10, pp. 1309–1319 (page 249).
521. Khan, Zaiba H et al. (2022). ‘Genome-Wide Analysis of AAAG and ACGT Cis-Elements in Arabidopsis Thaliana Reveals Their Involvement with Genes Downregulated under Jasmonic Acid Response in an Orientation Independent Manner’. In: *G3 Genes/Genomes/Genetics* 12.5, jkac057 (page 249).
522. Amack, Stephanie C., Savio S. Ferreira and Mauricio S. Antunes (2022). ‘Tuning the Transcriptional Activity of the CaMV 35S Promoter in Plants by Single-Nucleotide Changes in the TATA Box’. In: *ACS Synthetic Biology* (page 251).
523. Betts, Scott D. et al. (2019). ‘Uniform Expression and Relatively Small Position Effects Characterize Sister Transformants in Maize and Soybean’. In: *Frontiers in Plant Science* 10 (page 252).
524. Patterson, Kurt et al. (2010). ‘Distinct Signalling Pathways and Transcriptome Response Signatures Differentiate Ammonium- and Nitrate-Supplied Plants’. In: *Plant, Cell & Environment* 33.9, pp. 1486–1501 (pages 254, 255).
525. Mergner, Julia et al. (2020). ‘Mass-Spectrometry-Based Draft of the Arabidopsis Proteome’. In: *Nature* 579.7799 (7799), pp. 409–414 (page 254).
526. Hai, T and T Curran (1991). ‘Cross-Family Dimerization of Transcription Factors Fos/Jun and ATF/CREB Alters DNA Binding Specificity.’ In: *Proceedings of the National Academy of Sciences* 88.9, pp. 3720–3724 (page 254).
527. Xu, Lei et al. (2023). ‘The Evaluation of Active Transcriptional Repressor Domain for CRISPRi in Plants’. In: *Gene* 851, p. 146967 (page 256).
528. Singh, Purnima et al. (2019). ‘Analysis of Rice Proteins with DLN Repressor Motif/S’. In: *International Journal of Molecular Sciences* 20.7 (7), p. 1600 (page 256).
529. Sekine, Daisuke et al. (2021). ‘Improving Quantitative Traits in Self-Pollinated Crops Using Simulation-Based Selection With Minimal Crossing’. In: *Frontiers in Plant Science* 12 (page 260).

530. Bian, C et al. (2023). *Conservation and Divergence of Regulatory Architecture in Nitrate-Responsive Plant Gene Circuits*. preprint. *Plant Biology* (page [261](#)).