
Basal promoter landscape in
Eucalyptus grandis:
Annotation of distal transcription
start sites and core promoter usage

By

IDA CECILIA VAN JAARSVELD

Submitted in partial fulfilment of the degree

MAGISTER SCIENTIAE BIOINFORMATICS

in the Faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria

11 MAY 2014

Supervisor: Prof Alexander A. Myburg
Co-supervisors: Dr Eshchar Mizrachi, Prof Fourie Joubert

SUBMISSION DECLARATION

I, Ida Cecilia van Jaarsveld, declare that the thesis, which I hereby submit for the degree *Magister Scientiae* Bioinformatics in the Department of Biochemistry, at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signature:



Date:

11 May 2014

ACKNOWLEDGEMENTS

I would like to acknowledge my supervisors, Prof Zander Myburg, Prof Fourie Joubert and Dr Eshchar Mizrahi. Thank you for enabling me to grow in the ways that I needed to, allowing me to explore and critically interrogate computational methods, to take steps back so that I may take surer steps forward. I am grateful to Dr Fatima Parker, Mark Tyrrell and Roger Elliot for facilitating completion. I would also like to acknowledge Prof. Yves Van de Peer and Thomas Van Parys for hosting me at the Bioinformatics and Systems Biology Unit, University of Ghent, Belgium. Thank you to my fellow students, Dr. Oliver Bezuidt and Pooja Singh for their constant friendship and support. I must thank my darlings, Daena and Colin van Jaarsveld, for unwittingly being the brightest rays of sunshine. I also thank my dearest cousins Michael (Ekim) and Jonathan (Onoj) McInerney, for their ample and genuine love and support. A tremendous thank you to Anastasios Michael Constantaras, who has shared this journey most intimately, I thank him for his patience, support and creative analogies; for keeping the faith and fostering my own. And lastly, thank you to Kobie van Jaarsveld, who remembered and reignited a strength in me which I had long forgotten, and showed me the all-important alternative to the hockey stick.

I would like to acknowledge Sappi, Mondi, the National Research Foundation (NRF), the South African Department of Science and Technology (DST) and the University of Pretoria Postgraduate Travel and Academic Achievement Awards for enabling me to explore and foster my career as a researcher.

For Team Paradiso waiting at the line, and my big brother, Kobie,
I could not have finished this without you.

SUMMARY

Basal promoter landscape in *Eucalyptus grandis*: Annotation of distal transcription start sites and core promoter usage

Ida Cecilia van Jaarsveld

Supervised by Prof Alexander A. Myburg

Co-supervised by Dr Eshchar Mizrahi and Prof Fourie Joubert

Submitted in partial fulfilment of the degree Magister Scientiae Bioinformatics

Department of Biochemistry

University of Pretoria

Transcription is a complex biological phenomenon, whereby RNA is transcribed from single stranded template DNA by assembling targeted regulatory inputs at the promoter region. Transcription is regulated through many hierarchically organised mechanisms, including chromosome positioning and organisation, the binding of transcription factors, and DNA's secondary and tertiary structures at the region of transcription initiation. The core promoter is the distinct functional unit of DNA overlapping the transcription start site, which possesses linear regulator capacity and renders DNA permissive to transcription. In plants, core promoter and enhancer studies are of particularly high impact for those traits which under strong transcriptional control. Cellulose biosynthesis in immature xylem, the tissue which forms wood, is one such trait, and is studied extensively in the herbaceous model plant organism, *Arabidopsis thaliana*, and the economically important woody perennial, *Eucalyptus grandis*. The release of the *E. grandis* genome sequence has provided a much-needed reference to study transcriptional control, not only for those traits that make it a dominant fibre crop, but genome-wide. We aimed to use empirical transcript evidence to perform a high-throughput genome-wide curation of the 5' UTR annotations and empirically infer transcription start sites (TSSs) of the nascent *E. grandis* genome annotation. We then aimed to use the curated TSSs to define core promoter classes based on their sequence

composition and to determine the putative expression profiles and functional associations of each.

We used deep *E. grandis* mRNA sequencing data across seven diverse tissues and PASA assembled *E. grandis* ESTs to empirically curate 5' UTR annotations. We improved 17,085 annotations, added 7,596 for which there was no previous annotation and retained 3,675 that possessed only a predicted TSS without empirical evidence. These complementary data were used to define distal transcription start sites (dTSS) by a novel, prioritising, computational rule-based method. From these dTSS annotations, we extracted the core promoters (from -100 to +50) and described the core promoter landscape by hexamer positional over-representation analysis. We found three types of hexamer over-representation in the core promoter, that being *broad*, *spiked* and *low*. *Broad* hexamers were classified into 5 distinct core promoter classes, including TA, CT, GA, W and S. These were further assessed for putative expression profiles (specificity and level) and functional associations. TA resembles the conserved TATA-box core promoter, although displays a bimodal distribution, low expression levels and the greatest tissue specificity. CT and GA are over-represented both up and downstream of the dTSS and show narrow windows of greater enrichment with phasic constraint. W and S occur in close proximity to the dTSS, with S displaying the most constitutive and highest expression profile. *Spiked* hexamers occur in close proximity to the dTSS and *low* hexamers are enriched for those pyrimidine-rich hexamers found in *Arabidopsis thaliana* and *Oryza sativa* core promoters as the Y Patch. We found that *E. grandis* core promoters include those such as the TATA-box class which is conserved across kingdoms, the CT and GA classes, which are conserved in *Arabidopsis*, and a number of classes which, thus far, appear unique to *Eucalyptus*. We postulate possible underlying mechanisms of each core promoter class based on their sequence composition and suggest regulation by TBP binding (TA), nucleosome positioning (W), DNA stability (S), and non-B-DNA conformation (CT and GA). This research provides a basal understanding of *cis*-transcriptional regulation at the core promoter in this economically important woody plant species and provides insight into the mechanisms of permissive transcription across plant species.

TABLE OF CONTENTS

Submission declaration	i
Acknowledgements.....	ii
Summary.....	iv
List of Tables	ix
List of Figures.....	x
List of Supplementary Tables.....	xi
List of Supplementary Figures.....	xi
List of Supplementary Notes	xi
List of Abbreviations	xii
Preface.....	xiv
CHAPTER 1: LITERATURE REVIEW	
TRANSCRIPTIONAL REGULATION AND THE CORE PROMOTER LANDSCAPE	1
1.1 Summary.....	2
1.2 Introduction.....	2
1.3 The complex cascade of transcription.....	3
1.4 Core promoter characterisation in plants	9
1.5 <i>In silico</i> methods to describe core promoter features	10
1.6 <i>Eucalyptus grandis</i>	11
1.7 Acknowledgements.....	12
1.8 References.....	12
1.9 Tables.....	21
CHAPTER 2	
EMPIRICAL CURATION OF 5' UTRS IN THE <i>EUCALYPTUS GRANDIS</i> GENOME	24
2.1 Summary.....	25
2.2 Introduction.....	25
2.3 Materials and Methods.....	28
2.3.1. Genomic and RNA transcription data acquisition	28
2.3.2 Annotation of 5' UTRs using mRNA-seq data.....	28
2.3.3 Length comparison of predicted and empirically derived 5' UTRs	29
2.4 Results.....	30

2.4.1 mRNA-seq transcript evidence extends upstream of predicted 5' UTRs	30
2.4.2 Empirical curation using expressed transcript data	31
2.5 Discussion	31
2.6 Conclusion	34
2.7 References	35
2.8 Tables and Figures	38
2.9 Supplementary material	49
2.10 Additional files	56
CHAPTER 3	
FIVE CORE PROMOTER CLASSES DRIVE TRANSCRIPTION IN <i>EUCALYPTUS GRANDIS</i>	57
3.1 Summary	58
3.2 Introduction	58
3.3 Materials and Methods	62
3.3.1 Hexamer over-representation	62
3.3.2 Hexamer clustering	63
3.3.3 Identifying gene groups	64
3.3.3 Expression analysis	64
3.3.4 Gene Ontology enrichment analysis	65
3.4 Results	65
3.4.1 Three distribution types of hexamer enrichment in <i>E. grandis</i> core promoters	65
3.4.2 Five core promoter types of <i>E. grandis</i> can be defined by simple repeat sequences	65
3.4.3 Core promoter classes are associated with different expression profiles	67
3.4.4 Core promoter classes drive functionally distinct gene categories	67
3.4.5 <i>Spiked</i> hexamers occur preferentially at the dTSS	68
3.5 Discussion	68
3.6 Acknowledgments	72
3.7 References	73
3.8 Tables and figures	77
3.9 Supplementary Material	88
3.10 Additional files	96
CHAPTER 4 : CONCLUDING REMARKS	
4.1 A genome-wide study of transcription in <i>Eucalyptus grandis</i>	99

4.2 Future perspectives	101
4.3 Acknowledgments.....	103
4.4 References.....	103
ADDITIONAL FILES DVD.....	106

LIST OF TABLES

Table 1.1. IUPAC nucleotide codes.....	21
Table 1.2. List of applicable publications in which core promoter elements are described. ...	22
Table 2.1. Two-way comparison of 5' UTR lengths between <i>FGH</i> , <i>PASA</i> and <i>NGS</i> sources	38
Table 2.2. Number of <i>NGS</i> annotations omitted by exclusionary criteria.....	39
Table 3.1. Core promoter classes detected and defined in <i>E. grandis</i>	77
Table 3.2. FPKM distribution for each promoter class.....	78

LIST OF FIGURES

Figure 2.1. Use of data sources for the curation of <i>E. grandis</i> 5' UTRs.	40
Figure 2.2. Length distributions of <i>FGH</i> , <i>PASA</i> and <i>NGS</i> determined 5' UTRs.....	41
Figure 2.3. Final contributions of each source to the prioritised collection of <i>E. grandis</i> 5' UTR annotations	42
Figure 2.4. Percentage and absolute frequency of prioritised 5' UTR sources per scaffold ...	43
Figure 2.5. Distribution of prioritised <i>E. grandis</i> 5' UTR lengths	44
Figure 2.6. Representation of <i>E. grandis</i> 5' UTR gff file	45
Figure 2.7. <i>NGS</i> data continues upstream of <i>PASA</i> and <i>FGH</i> annotations.....	45
Figure 2.8. Close proximity head-to-head <i>E. grandis</i> genes are annotated with <i>NGS</i> 5' UTRs	47
Figure 2.9. Filtering criteria remove spurious upstream splicing events from <i>NGS</i> 5' UTR annotations	48
Figure 3.1. Overview of methodology.....	79
Figure 3.2. Odds Ratio distribution of over-represented hexamers determined by Fishers Exact Test.....	80
Figure 3.3. Bootstrapped hierarchical clustering showing hexamer co-occurrence across promoters	81
Figure 3.4. Venn diagram of core promoter class co-occurrence per promoter.	82
Figure 3.5. Expression profiles of <i>E. grandis</i> core promoter classes.	83
Figure 3.6. Expression specificity of <i>E. grandis</i> core promoter classes.....	84
Figure 3.7. Expression specificity of core promoter classes.....	85
Figure 3.8. REVIGO TreeMap display of GO terms enrichment for two promoter classes. ...	86
Figure 3.9. Summary of core promoter classes defined in <i>E. grandis</i>	87

LIST OF SUPPLEMENTARY TABLES

Supplementary Table 3.1. Formulaic contingency table for Fisher's Exact Test of hexamer over-representation	88
Supplementary Table 3.2. Text contingency table for Fisher's Exact Test of hexamer over-representation.....	89

LIST OF SUPPLEMENTARY FIGURES

Supplementary Figure 2.1. Use of NGS data to curate gene annotations.....	49
Supplementary Figure 2.2. Distribution of <i>A. thaliana</i> 5' UTR lengths.....	50
Supplementary Figure 2.3. Illustrations of predicted and empirical 5' UTR discordance	51
Supplementary Figure 3.1. Hexamer over-representation for types <i>spiked</i> and <i>low</i>	90
Supplementary Figure 3.2. Enrichment of <i>spiked</i> hexamers at the dTSS.....	91

LIST OF SUPPLEMENTARY NOTES

Supplementary Note 2.1. Pseudocode to delimit 5' UTRs from aligned mRNA-seq data.....	52
Supplementary Note 2.2. Pseudocode to extract and modify 5' UTR annotations from Phytozome gene model annotations.....	53
Supplementary Note 2.3. Pseudocode to modify EST tabular data and extract 5' UTRs	54
Supplementary Note 2.4. Script with example Linux commands to separate the merged gff3 file.	55
Supplementary Note 3.1. Pseudocode to determine over-represented hexamers	92
Supplementary Note 3.2. Procedure to cluster over-represented hexamers.	93
Supplementary Note 3.3. Procedure to determine core promoter class enrichment regions and gene lists.....	94
Supplementary Note 3.4. Procedure to determine expression level and specificity	95

LIST OF ABBREVIATIONS

BREd	TFIIB Recognition Element downstream
BREu	TFIIB Recognition Element upstream
CesA	Cellulose Synthase
CPE	Core promoter element
CRM	<i>Cis</i> -regulatory module
DoE	Department of Energy
DPE	Downstream promoter element
dsDNA	Double stranded DNA
dTSS	Distal transcription start site
EST	Expressed sequence tag
FL	Flower
G4	G-quadruplex
GBP	GAGA-binding proteins
GO	Gene Ontology
GTF	General transcription factor
Inr	Initiator
IX	Immature xylem
JGI	Joint Genome Institute
ML	Mature leaf
MM	Markov model
MTE	Motif Ten Element
NDR	Nucleosome depleted region
NGS	Next Generation Sequencing
NTP	Nucleoside triphosphate
PH	Phloem
PIC	Pre-initiation complex
PSI	Percent-spliced-in
RSAT	Regulatory Sequence Analysis Tools
RT	Root
SCW	Secondary cell wall

ssDNA	Single stranded DNA
ST	Shoot tip
TAPs	Transcription associated proteins
TBP	TATA-binding protein
TFBS	Transcription factor binding site
TSS	Transcription start site
TSSD	Transcription start site distribution
UTR	Untranslated region
WMW	Wilcoxon-Mann-Whitney
YL	Young leaf

PREFACE

The early 21st century is renowned for the near-exponential growth in technological and digital applications. This is no different for genetic research, which has seen an explosion of high-throughput experimentation and subsequent analysis of digital genetic data. Enter bioinformatics: the research discipline synergising genetics, statistics and computer science. The amalgamation of these fields allows the conversion of sequencing and other high-throughput data into significant, biologically meaningful results and hypotheses. Two fundamental research activities, the sequencing of DNA and RNA, have provided extensive clues of biological phenomena, answering pertinent questions, whilst asking many more. Genome-wide research is thus often descriptive, and employs exploratory data analysis techniques for hypothesis generation. These methods have been explored *in plantae*. Although the majority of work has been pursued on the model herbaceous annual, *Arabidopsis thaliana*, many of the findings are transferable to other organisms and provide broad spectrum physiological insight. These inferences, together with high-throughput genomics and transcriptomics projects in other plant organisms, have provided insights into plant evolution, disease susceptibility and resistance, pathogen defense and other agriculturally and economically pertinent traits.

The *Eucalyptus* genus, endemic to Australia, is grown globally, and is an efficient carbon-sequestrator and subsequent biomass producer. Apart from more traditional uses as pulp, paper and other chemical cellulose products, the potential as a fast-growing, high yield, non-food-source biofuel feedstock renders members of this genus of considerable economic value in the agricultural sector. The economic value, and the yield increase derived from research activities, has motivated the sequencing and annotation of the *E. grandis* genome. Coupling this, the sequencing of mRNA, describing expression profiles, alternative transcript isoforms and characterising variance in diverse tissues, provides detail on organ and population diversity. Arguably the most economically vital process of *E. grandis* is that of xylogenesis, the biosynthesis of wood. The complex feat of secondary cell wall formation, constituting the deposition of cellulose, lignin, and hemicelluloses, is regulated by transcriptional mechanisms. DNA provides insight of these transcriptional mechanisms by allowing the study of regulatory regions, whilst quantified mRNA abundance provides insight by correlating expression profiles to these regulatory DNA features.

This study uses the v1.1 *E. grandis* genome assembly and annotation (Phytozome V1.0), and differential expression mRNA-seq data generated by the Forest Molecular Genetics Programme from diverse *E. grandis* tissues and organs. These data were manipulated beyond their typical use case for the high-throughput curation of 5' UTRs to infer distal transcription start sites (Chapter 2), and subsequently characterise core promoters (Chapter 3). The two aspects of algorithm design explored in this research were data reduction and pattern mining, both of which required extensive defensive programming, including both object-oriented and scripting design. The size of the data sets required comprehensive data management practices, including data provenance, and performance aware algorithm design. A plethora of statistical procedures were applied to the data to derive appropriate and statistically validated results. This thesis is the fruition of developing and using these skills, the intense interrogation of data generated at the forefront of genetic research technology, and the appropriate and semantically-aware interpretation of these results. This thesis presents these results in journal format, and so provides key insights into the basal regulatory mechanisms of *E. grandis*.

The research underlying this thesis has been reported in the following publications and conference presentations:

PUBLICATIONS IN ISI RATED JOURNALS

Myburg, AA, Grattapaglia, D, Tuskan, G, et al. Genome sequence of *Eucalyptus grandis*: A global tree crop for fiber and energy. (In press – *Nature*).

PRESENTATIONS IN NATIONAL AND INTERNATIONAL CONFERENCES

van Jaarsveld IC, Mizrachi E, Joubert F, Myburg AA. 2012. Annotating transcription start sites and core promoter usage in *Eucalyptus grandis*. Joint South African Genetics Society and South African Society of Bioinformatics and Computational Biology Conference, Stellenbosch, 10-12 September 2012 (oral presentation).

van Jaarsveld IC, Mizrachi E, Joubert F, Myburg AA. 2012. Ensemble optimisation of *cis*-regulatory element discovery: *in planta* benchmark and discovery in *Eucalyptus*. South African Association of Botanists Conference, Pretoria, 16-18 January 2012 (oral presentation).

van Jaarsveld IC, Mizrachi E, Joubert F, Myburg AA. 2011. Describing promoter landscapes in *Eucalyptus grandis*. Nucleotides to Networks Symposium, Ghent, Belgium, 21 May 2011 (poster presentation).

van Jaarsveld IC, Mizrachi E, Joubert F, Myburg AA. 2011. *De novo* discovery of *cis*-regulatory motifs and modules implicated in the transcriptional regulation of genes associated with secondary cell wall formation and cellulose biosynthesis in *Eucalyptus*. Joint International Society for Computational Biology and African Society for Bioinformatics and Computational Biology Conference, 9-11 March, 2011 (poster presentation).

CHAPTER 1: LITERATURE REVIEW

Transcriptional regulation and the core promoter landscape

1.1 SUMMARY

The initiation and successful elongation of transcription is a complex biological phenomenon. A step-wise cascade of events renders core promoters permissive to recognition by a host of general transcription factors. The recruitment of these factors ultimately facilitates the binding of RNA Polymerase II, which in an open complex, transcribes a molecule of RNA. This process is highly regulated, with co-operative, combinatorial and competitive regulatory inputs. Core promoters are thus key functional units of the genome, yet lack comprehensive description and functional characterisation. *Eucalyptus grandis* is an economically important fibre crop, and crucial to its yield of fibre and pulp, is the understanding of transcriptional mechanisms which drive xylogenesis and the enrichment of cellulose in secondary cell walls of immature xylem.

1.2 INTRODUCTION

The DNA of an organism, near identical in each cell, comprises a gene repertoire capable of achieving complex and highly variable cellular phenotypes. This is not achieved by the full repertoire, but rather by the specific expression of a subset of genes to developmental, homeostatic and environmental stimuli. Thus protein-coding and functional non-coding RNA genes are temporospatially transcribed from template DNA (Irimia *et al.*, 2013). Although crucial to organism success, this phenomenon is not yet fully understood. There are two conflicting schools of thought regarding transcriptional regulation: that which describes the process as precisely regulated (Lionnet & Singer, 2012), and that which describes the process as a stochastic (Kaern *et al.*, 2005) phenomenon with high rates of failed or abortive transcription (Yuzenkova *et al.*, 2011). Both schools agree that transcription is a semi-hierarchically networked biological process with intricate feedback mechanisms, with many inputs required at the target DNA to co-operatively drive successful transcription initiation and elongation. The core promoter binds the basal transcription machinery, and its sequence composition drives the physiochemical properties which permit this affinity. This formidable regulatory task utilises multi-tiered composite mechanisms to direct regulatory input assembly (Prohaska *et al.*, 2010). These tiers include nuclear, chromatin and linear levels of organisation, and it is becoming clear that these tiers should not be considered in isolation. This review explores the non-random organisation of the millions of base pairs of DNA

within the nucleus, and the resultant regulatory consequences. The linear regulatory capacity of DNA is discussed in greater detail, followed by a review on the experimental and computational methods by which these may be characterised. Lastly, the applications of such technologies on the economically important global fibre crop, *Eucalyptus grandis*, are explored.

1.3 THE COMPLEX CASCADE OF TRANSCRIPTION

Transcription is a consequence of the interaction of three primary molecules, namely RNA Polymerase II (RNA Pol II), DNA and nucleoside triphosphates (NTPs) (Borukhov & Nudler, 2008). After RNA Pol II covalently crosslinks to DNA and induces melting of double stranded DNA (dsDNA), it reads template single stranded DNA (ssDNA) and catalyses unidirectional phosphodiester bonds between complementarily sequestered NTPs, thus transcribing a molecule of RNA. The functional units of the genome which lie upstream of genes and permit RNA Pol II crosslinking are termed core promoters. RNA Pol II binding requires core promoter DNA to be in a permissive state, the result of complex cascade of molecular interactions and conformation changes which amount to regulatory inputs at the promoter. These regulatory inputs originate in both *cis* (Thompson *et al.*, 2004; Geisler *et al.*, 2006; Walther *et al.*, 2007; Wray, 2007; Vandepoele *et al.*, 2009; Priest *et al.*, 2009) and *trans* (Pai & Engelke, 2010; Cremer & Cremer, 2010), with the transcriptional outputs being described in both quantitative (Lionnet & Singer, 2012) and qualitative terms (Wray, 2007).

In Eukaryotes, RNA Pol II is unable to independently recognize promoters. Its binding, and subsequent persistence (Dundr *et al.*, 2002), are facilitated by the presence of several general transcription factors (GTFs) TFIIA, TFIIB, TFIID, TFIIE, TFIIIF and TFIIH and their various co-factors, which assemble as the pre-initiation complex (PIC). GTFs show high conformational flexibility (Cianfrocco & Nogales, 2013) and render core promoter DNA permissive to transcription, recruit RNA Pol II, stabilize the PIC, support successful elongation by interaction with the nascent RNA (Bushnell *et al.*, 2004), and provide redundancy in cases of abolished or diminished binding by promoter DNA mutation (Cianfrocco & Nogales, 2013). The step-wise (Dundr *et al.*, 2002) assembly of the PIC is initiated by DNA sequence recognition of TFIID's subunit, TATA-binding protein (TBP), at the TATA box (TATAWA; see Table 1.1 for IUPAC codes) core promoter element (CPE). This CPE element is topologically conserved and resides ~35 bp upstream from the dominant

transcription start site (TSS) in both metazoans and plants. The binding of TBP and associated factors of TFIID allows for the recruitment of other GTFs and ultimately the successful recruitment of RNA Pol II.

The core promoter has been extensively studied in metazoans (Valen *et al.*, 2009; Hoskins *et al.*, 2010; Zou *et al.*, 2011; Lenhard *et al.*, 2012; Kwak *et al.*, 2013), describing other prominent CPEs and the GTFs with which they interact. TFIIB Regulatory Element Upstream (BREu; SSRCGCC) and TFIIB Binding Regulatory Element Downstream (BREd; RTDKKKK) flank the TATA box and facilitate binding of TFIIB. Downstream Promoter Element (DPE; RGWYVT) and Motif Ten Element (MTE; CSARCSSAAC) are downstream of the TSS and interact with TFIID. The initiator element (Inr; YYANWYY) underlined at the TSS, drives TSS selection and specificity. Interestingly, the established TATA box consensus is present in less than 3% of functional human TBP binding sites (Venters & Pugh, 2013). Molina & Grotewold (2005) similarly describe the TATA box to be present within the *Arabidopsis* TBP functional window far less frequently than originally expected. TATA box is, nonetheless, the most widely known CPE, which is result of it having the strongest consensus aiding detection, rather than its ubiquity. Most metazoan promoters lack a TATA box. The BREu, BREd, DPE and MTE elements are similarly rare and are not required to co-occur (Kadonaga, 2012), although their combined presence and the adherence to consensus increases the affinity for GTF binding and PIC stability (Cianfrocco & Nogales, 2013). In metazoan promoters devoid of a TATA box, CpG islands are prevalent (Choy *et al.*, 2010). In plants, which lack the methylation patterns observed of CpG islands, an enrichment of simple repeat sequences is observed (Yamamoto *et al.*, 2007a; Maruyama *et al.*, 2012). Although the TATA box is taxonomically prolific, occurring in bacteria, plants, yeast and other more complex metazoans, the composition of TATA-less promoters differs significantly (Yamamoto *et al.*, 2007a), and thus the mechanisms driving transcription at these loci remain unclear. Apart from CPE consensus adherence in the correct topological window, accessibility is a major determinant of GTF affinity, a feature which is controlled by epigenetic mechanisms such as nucleosome positioning and histone modifications.

Approximately 147 bp of DNA wrap around eight histone subunits to form a nucleosome (Klug & Lutter, 1981), joined to the adjacent nucleosome by linker DNA of variable lengths. DNA organised into nucleosomes thus resembles a 10 nm beads-on-a-string fibre. Nucleosome occupation is in general inhibitory of transcription by competitive binding or

occlusion at the CPEs (Brown *et al.*, 2013). However, nucleosome occupation can be altered at promoter sites by a cascade of protein-protein interactions and conformational changes. The histone subunits H3 and H4 are amenable to covalent modifications such as acetylation and methylation, epigenetic marks of chromatin state and regulated transcription (Du *et al.*, 2013). Histone activators are gene-specific recruiters of acetyltransferases and methyltransferases which modify histone lysine and arginine tail residues. Open chromatin and thus permissive PIC assembly is associated with H3K4me3 (tri-methylated lysine residue at position 4 of histone H3) and H3K4me2 as they recruit ATP-dependent chromatin remodelers. These remodelers either evict the nucleosome or reposition the DNA with respect to histone subunits making CPE available for GTF binding. Histone depletion in yeast proximal promoters is associated with constitutive (Narlikar *et al.*, 2007) and conserved (Rosin *et al.*, 2012) expression, while precise nucleosome positioning is associated with regulated transcription. The utility of histone modifications extends beyond that of transcription initiation, with H3K36me3 of the +1 nucleosome associated with successful transcription elongation (Guenther *et al.*, 2007). Dong *et al.* (2012) support this by finding correlation with H3K36me3 marks and RNA-Seq gene expression (successful full-length transcription and Poly-A+ post-transcriptional modification) while H3K4me3 marks are correlated with CAGE transcripts (5' cap modification occurs during elongation).

Apart from nucleosome occlusion of CPEs as a negative regulator of transcription (Dong *et al.*, 2011), the presence or absence and the positioning of nucleosomes induces DNA supercoiling (Kouzine & Levens, 2007). Supercoiling, both negative and positive, alter the periodicity (Brick *et al.*, 2008; Zhai *et al.*, 2011; Kravatskaya *et al.*, 2013) and torsion of linear DNA. These changes in DNA conformation affect the shape and polarity of DNA grooves, which in turn alter the binding affinity of TFs and other DNA binding proteins (Zhitnikova *et al.*, 2013). The looping of DNA is also required for the combinatorial interaction of several *cis*-regulatory sites (Kouzine & Levens, 2007; Hakim *et al.*, 2010; Geggier & Vologodskii, 2010; Shandilya & Roberts, 2012). B-DNA, the standard DNA conformation, is, however, an intrinsically rigid molecule. Contrarily, gene-looping, protein-DNA interactions and successful transcription initiation require varying degrees of DNA flexibility. Every third atom along the DNA backbone is a hydrophilic polar charged phosphate, rendering the backbones, and the double helix hydrophilic. This promotes interaction with proteins such as histones and TFs, as well as RNA molecules. Thus it is the binding of TFs that induce topologically specific DNA bending and flexibility. The resultant

change in curvature and coiling of the core promoter fulfills a structural, organizational or regulatory function (Papantonis & Cook, 2010; Hakim *et al.*, 2010). The DNA itself is also capable of different morphologies, such as Z-DNA, a left-handed double helix, H-DNA, a triplex structure, and G-quadruplex DNA. H-DNA forms in CT·GA dinucleotide repeat regions and results in both single and triple-stranded DNA, resulting in an open chromatin conformation permissive to the binding of TFs. The formation of G-quadruplex DNA (G4) on the leading strand, leaves the reverse or template strand open in a loop conformation, a structure in which transcription is highly permissive (Sawaya *et al.*, 2013; Maizels & Gray, 2013). Non-B-DNA conformations arise as a result of supercoiling (Kouzine & Levens, 2007), and thus describes the manner in which protein-DNA interactions which induce supercoiling can introduce exaggerated conformational changes.

The DNA structure is constantly morphing, responding to factors and internal physiochemical interactions. It is thus not surprising, but no less remarkable, that the highly specific and complex phenotypes derived from DNA expression, show a high level of expression stochasticity. Lionnet & Singer (2012) discuss advances made by single-cell expression analyses. These suggest that transcription occurs in bursts, where genes randomly switch “on” and “off” and show considerable variation in a cell population is dependent on cell-cycle (Zopf *et al.*, 2013). Seila *et al.* (2009) also show that despite strand specificity of core promoter elements, transcription is likely to occur bi-directionally at most transcriptionally active promoters. Only a small proportion of initiated transcripts successfully elongate into full length RNA molecules. Transcript abortion occurs prolifically while transcripts are shorter than 5 bp. Once this length is reached, the nascent transcript is able to interact with TFIIA and the catalytic center of the open RNA Pol II complex which stabilizes transcription (Shandilya & Roberts, 2012). Further elongation stabilizers include histone modifications of the +1 nucleosome. RNA Pol II pausing downstream of the TSS is an abundant phenomenon (Mokry *et al.*, 2012), and has been associated with sequence composition downstream of the TSS (Kwak *et al.*, 2013). RNA Pol II pausing is associated with nucleosome depletion and the accessibility of CPEs and enhancers by their respective factors (Adelman & Lis, 2012). In addition to the stochasticity of successful transcription elongation and promoter escape, there is also a high level of variability in TSS selection.

Numerous studies quantifying the 5' termini of transcripts found, rather than a discrete TSS, transcription initiating at different abundances over variably broad ranges (Carninci *et al.*,

2006; Yamamoto *et al.*, 2009; van Heeringen *et al.*, 2011; Zhao *et al.*, 2011). This evidence of TSS dispersion refutes the qualitative assignment of a single position as the transcription start site (TSS). Rather, a quantitative descriptor, such as a transcription start site distribution (TSSD; Zhao *et al.*, 2011) better describes the transcription initiation pattern of the majority of genes across kingdoms (Hoskins *et al.*, 2010). Comprehensive capturing of capped transcripts is employed in many techniques such as CAGE (de Hoon & Hayashizaki, 2008), deepCAGE (Valen *et al.*, 2009), nanoCAGE (Salimullah *et al.*, 2011), CT-MPSS (Yamamoto *et al.*, 2009) and TSS-seq (Kanai *et al.*, 2011). This is the most accurate way to define TSSDs. In human, broad, sharp, broad with peak and multimodal TSSDs have been described (Balwierz *et al.*, 2009) Zhao *et al.* (2011) describe the distributions as scattered, dense and ultra-dense. For *Arabidopsis*, the TSSDs have been described as either broad or convergent, with measures regarding the kurtosis and tail-length of the distribution (Yamamoto *et al.*, 2009). For each of these distributions, core promoter sequence biases have been detected, with TATA promoters associated with sharp/dense/convergent TSSDs, and TATA-less with broad/scattered. Robb *et al.* (2013) posit a model for alternative TSS selection based on the length and width of the transcription bubble that is formed in promoter melting. Transcription bubble expansion and the associated “scrunching” (Kapanidis *et al.*, 2006) of adjacent DNA induces the use of downstream TSSs, whereas transcription bubble contraction and the associated DNA “unscrunching” (Kapanidis *et al.*, 2006) result in the preferential use of upstream TSSs. Such alternative TSS use changes the periphery of the 5' UTR sequence, and may alter translation efficiency, further propagating the stochasticity from gene to protein, and thus functional, expression.

The linear organization of DNA upstream to the 5' periphery of a gene is largely responsible for the *cis* regulation of gene expression and is crucial for the co-operative binding of TFs and assembly of core regulatory inputs. Core promoter classes have been described according to their nucleotide composition, with alternative compositions possessing alternative properties (Lenhard *et al.*, 2012). The linear *cis*-regulatory capacity of DNA is most often described in terms of the potential to bind factors by cognate consensus patterns (Wingender *et al.*, 1996, 2000; Lescot *et al.*, 2002). However, there are many physiochemical properties which contribute to regulatory inputs and are independent of TF affinity for a DNA consensus (Hoskins *et al.*, 2010). These physiochemical properties are inherently derived from the sequence composition of the core promoter. DNA melting, curvature, rigidity and torsion are all properties dependent on sequence composition (Geggier & Vologodskii, 2010).

Nucleosome occupancy is also a product of sequence composition (Kaplan *et al.*, 2009). The looping of DNA into concisely structured foci of transcriptional activity is dependent of composition specific properties (Boedicker *et al.*, 2013). TSS selection is also guided by general sequence composition, as the size of the transcription-bubble is a product of melting potential and thermodynamic stability (Robb *et al.*, 2013), and the resultant scrunching is a determinant of transcription initiation positioning (Kapanidis *et al.*, 2006; Chinnaraj *et al.*, 2013). Thus enriched core promoter sequence may fulfill two co-dependent roles, physiochemical structure as a direct or indirect component of a regulatory input, and for the consensus driven recognition by TFs.

The distinction between the core and proximal and distal promoters is in their regulatory capacity. Core promoters may bind GTFs or possess particular physiochemical properties to regulate transcription, but the PIC binds at the core promoter and this region acts as a discrete regulator in that a gene is either switched “on” or “off”. Enhancers are more abundant in the proximal promoter than the distal, but both bind TFs which regulate the temporospatial transcription of a gene and are regarded as modulators. As TFs commonly bind as a complex, there are typically a number of enhancers or binding sites in any one promoter (Wagner, 1999), which may be homotypic or heterotypic. A cluster of functional TFBSs is a *cis*-regulatory module (CRM) and are often only functional if their constituent TFBS are within a particular order, range from one another, strand orientation and distance from the transcription start site (TSS) (Aerts *et al.*, 2003; Zeigler *et al.*, 2007; Priest *et al.*, 2009; Su *et al.*, 2010). CRMs may also bind at transcription factories, assemblies of transcriptional machinery, the interacting TFs and co-factors which are perpetually anchored at discrete sites within the nucleus (Sutherland & Bickmore, 2009). This opposes the original ideology of a TF diffusing through the nucleus and happening upon a region where it may bind, to the more orchestrated movement of a locus to a semi-permanent transcription factory (Pai & Engelke, 2010). Although this phenomenon is near fully accepted, the extent to which this occurs is still unknown. It is most likely that both methods of TF-DNA interaction occur. Supporting the hypothesis of DNA requiring long range mobility for successful transcription, are chromosomal territories and their inter-chromosomal regions. A cell’s chromosomes, during interphase, are organised into distinct chromosome territories (Fraser & Bickmore, 2007; Pai & Engelke, 2010; Cremer & Cremer, 2010; Sáez-Vásquez & Gadál, 2010; Gagnic & Ionescu-Tirgoviste, 2013). Chromosomes positioned towards the centre of the nucleus have lower transcriptional activity than those at the periphery. Further, those loci which are located

at the periphery of each chromosome's territory, called the inter-chromosomal region, display greater transcriptional activity than those loci which are positioned at the chromosome territory foci (Hakim *et al.*, 2010). The DNA does not form such structures alone. It requires a repertoire of proteins which behave as a scaffold for such DNA organisation (Wasson & Hartemink, 2009). Proteins not only fix chromosomes in a particular morphology, but also fix the structure to a point in the nucleus (Zhao *et al.*, 2009). The dynamic movement of chromosomes and loci according to their levels of transcriptional activity have also been documented (Fraser & Bickmore, 2007; Deng *et al.*, 2012). This indicates that a locus may be preferentially expressed if located peripherally (Cerná *et al.*, 2004).

1.4 CORE PROMOTER CHARACTERISATION IN PLANTS

Plant core promoters are described by the elements i) TATA-box, ii) Y Patch, iii) GA elements, iv) CA elements and v) coreless (Hieno *et al.*, 2014). The core promoter composition in plants is enriched with W residues, which are the least stable in DNA complementarity (Maruyama *et al.*, 2012), as well as dinucleotide repeat sequences. WR repeats are underrepresented, excepting TATATA and TATAAA, with enrichment for AT possibly driving promoter context (Yamamoto *et al.*, 2007b; Mogno *et al.*, 2010). YR residues are over-represented upstream of the TSS (Maruyama *et al.*, 2012) and also constitute the plant Inr element (Yamamoto *et al.*, 2007b). Y repeat sequences are over-represented proximal to the TSS, termed the Y-patch, which has been modeled to co-occur with TATA box and the Inr (YR). The initial study defining the plant specific Y Patch indicates a peak position at -13, comprising of sequences TCTCTC, CCTCTC, CTTCTC, CTCCTC, CTCTTC, CTCTCC, TCCCTC, TTCTTC and TTCTCT (Yamamoto *et al.*, 2007b). In later work (Yamamoto *et al.*, 2009), the analysis region was extended to include 200 bp downstream and the GA repeat sequence was characterized, found to occur in TATA-less promoters, and associated with dispersed TSSDs. Both CT and GA repeats have been shown to bind GAGA binding proteins BASIC PENTACYSTEINE (BPC) and BARLEY B RECOMBINANT (BBR) in plants and to have both activating and inhibitory expression responses (Berger & Dubreucq, 2012). There is at this point little empirical evidence for the functional mechanisms of sequence composition enriched in plant core promoters. Apart from TBP recognising the canonical TATA box and possibly a poly-T tract (Ahn *et al.*, 2012) and putative binding of GBP at GA repeat elements, little else is known about the sequence specific structural conformations or the recruitment of GTFs in plants. Perhaps the most

poorly described aspect of plant promoters, is the interaction of TFs and their cognate enhancers with the core transcriptional machinery, and how such interactions drive combinatorial control of gene expression.

1.5 *IN SILICO* METHODS TO DESCRIBE CORE PROMOTER FEATURES

As gene regulation is a multi-faceted process, it is unsurprising that the field of regulatory genomics is similarly broad. Specific areas of investigation include: transcriptional networks (Mutwil *et al.*, 2009; Street *et al.*, 2011); whole-genome promoter region detection (Abeel *et al.*, 2009); identification of regulators (TFs and co-factors) (Persson *et al.*, 2005); *cis*-regulation (Vandepoele *et al.*, 2009); nucleosome occupancy (Lee *et al.*, 2004); 3D chromatin organisation (Hakim *et al.*, 2010); methylation states (Aceituno *et al.*, 2008); and nuclear organisation of TFs (Xu & Cook, 2008), amongst others. For the detection of CPEs, while technological advances have allowed the nucleotide resolution determination of TFBSs (Zhong *et al.*, 2010) and characterization of DNA accessibility in proximity to the promoter (Guenther *et al.*, 2007; Swamy *et al.*, 2011), the majority of studies rely heavily on *in silico* pattern mining of DNA sequence. This approach relies primarily on the over-representation of patterns in a target dataset when compared to a control dataset. A plethora of statistical tests can be used to determine over-represented “words” or motifs, but the accuracy of these results are heavily reliant on the correct acquisition of the target promoter dataset, and the appropriate selection of a control dataset for maximum discriminatory power. Several *in silico* methods to determine functional core promoter sequences are summarized in Table 1.2. Typically, the positional conservation or topological constraint on the occurrence of a word is an indicator of conservation, which is in turn an indicator of putative functional relevance. In each of these studies, the correct annotation of the TSS, or TSSD is crucial, as the topological restraint is with respect to these features. Incorrect annotations decrease the signal to noise ratio, and hamper the identification of true signals. To limit the search space and enrich for true regulatory elements, expression data are often used to select a number of candidate loci which are putatively co-regulated (Vandepoele *et al.*, 2009; Mutwil *et al.*, 2011). There is a slight disconnect in that steady-state RNA abundances are measured while inferences are made about the transcription initiation rate or persistence, where the RNA abundances may be affected by other factors such as mRNA degradation or ribosomal stalling. Nevertheless, this method is popular and has provided valuable insights into transcriptional regulation.

1.6 *EUCALYPTUS GRANDIS*

The *Eucalyptus* genus, endemic to Australia, has over 700 species divided into 13 subgenera. Because of the genus' remarkably broad adaptability, it now has members grown in many parts of the world as economically important fibre crop species (Taylor et al. 2009). The superior growth and multipurpose wood properties have resulted in the research of a number of these species, most typically in the properties of their wood and xylogenesis (Ranik & Myburg, 2006; Myburg *et al.*, 2008; Grattapaglia & Kirst, 2008; Rengel *et al.*, 2009; Mizrachi *et al.*, 2010; Creux *et al.*, 2011, 2013; Juanita & Mun, 2011; Rencoret *et al.*, 2011; Hussey *et al.*, 2013). The largest of the subgenera, Symphyomyrtus, comprises a number of economically important eucalypt species, including *E. grandis*, *E. urophylla*, *E. globulus*, *E. calmaldulensis*, *E. gunnii* and *E. nitens*. Marker assisted breeding has allowed for hybrid plants, such as that of *E. grandis* x *E. urophylla* and *E. nitens* x *E. globulus*. These hybrids possess optimised growth and wood density for maximal economic yield (Taylor et al. 2009). Of pivotal import is the process of cellulose biosynthesis in the formation of the secondary cell wall (SCW). Cellulose is the desired molecule for downstream applications in pulp, paper, chemical cellulose and biofuel feedstock technologies and industries (Mizrachi *et al.*, 2012). Cellulose thus requires isolation from hemicelluloses and lignin, which are likewise deposited in the SCW during xylogenesis. Cellulose biosynthesis has been shown to be under strong transcriptional control (Hussey *et al.*, 2013). The promoters of cellulose synthase (*CesA*), the membrane bound catalytic unit which synthesises cellulose at the cell wall, have been described showing several conserved *cis*-elements, yet also to lack a TATA-box for TBP binding (Creux *et al.*, 2011, 2013). In order to fully understand and characterise transcription, it is necessary to determine which CPEs or core structural conformations render the DNA permissive to transcription, so that enhancer elements may spatiotemporally modulate gene expression.

The International *Eucalyptus* Genome Network (EUCAGEN; www.eucagen.org) released the *E. grandis* genome sequence in January 2011 (Myburg *et al.*, in press) The 8X Sanger coverage genome was sequenced and assembled by the Joint Genome Institute (JGI). The full genome sequence, in combination with mRNA-Seq expression data for seven diverse tissues (Hefer *et al.*, in preparation), provide the necessary data to curate TSS annotations and perform a core promoter analysis. We aim to empirically curate 5' UTR annotations using 454 EST and mRNA-seq data and identify core promoter classes and their specific

expression and functional associations in *E. grandis*. Empirically substantiated TSS and core promoter annotations will complement the release of the *E. grandis* genome and allow for the generation of testable working hypotheses of promoter regulatory capacity in this economically important fibre crop species.

1.7 ACKNOWLEDGEMENTS

I would like to acknowledge all members of the Forest Molecular Genetics group at the University of Pretoria for their critical discussion of xylogenesis, *Eucalyptus* biology, high-throughput experimentation of nucleic acids and the relevant downstream bioinformatics. I would also like to acknowledge Alexander A. Myburg, Eshchar Mizrahi and Fourie Joubert for their review of this manuscript.

1.8 REFERENCES

- Abeel T, Van de Peer Y, Saeys Y. 2009.** Toward a gold standard for promoter prediction evaluation. *Bioinformatics* **25**: i313–i320.
- Aceituno FF, Moseyko N, Rhee SY, Gutiérrez R a. 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*. *BMC Genomics* **9**: 438.
- Adelman K, Lis JT. 2012.** Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics* **13**: 720–731.
- Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. 2003.** Computational detection of *cis*-regulatory modules. *Bioinformatics* **19**: ii5–ii14.
- Ahn S, Huang C-L, Ozkumur E, Zhang X, Chinnala J, Yalcin A, Bandyopadhyay S, Russek S, Unlü MS, DeLisi C, et al. 2012.** TATA binding proteins can recognize nontraditional DNA sequences. *Biophysical Journal* **103**: 1510–1517.
- Bailey TL, Williams N, Misleh C, Li WW. 2006.** MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research* **34**: W369–373.
- Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E. 2009.** Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biology* **10**: R79.
- Berger N, Dubreucq B. 2012.** Evolution goes GAGA: GAGA binding proteins across kingdoms. *Biochimica et Biophysica Acta* **1819**: 863–868.

- Bernard V, Brunaud V, Lecharny A. 2010.** TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics* **11**: 166–180.
- Boedicker JQ, Garcia HG, Phillips R. 2013.** Theoretical and experimental dissection of DNA loop-mediated repression. *Physical Review Letters* **110**: 018101.
- Borukhov S, Nudler E. 2008.** RNA polymerase: the vehicle of transcription. *Trends in Microbiology* **16**: 126–134.
- Brick K, Watanabe J, Pizzi E. 2008.** Core promoters are predicted by their distinct physicochemical properties in the genome of *Plasmodium falciparum*. *Genome Biology* **9**: R178.
- Brown CR, Mao C, Falkovskaia E, Jurica MS, Boeger H. 2013.** Linking stochastic fluctuations in chromatin structure and gene expression. *PLoS Biology* **11**: e1001621.
- Bushnell DA, Westover KD, Davis RE, Kornberg RD. 2004.** Structural basis of transcription: an RNA polymerase II-TFIIB cocystal at 4.5 Angstroms. *Science* **303**: 983–988.
- Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple C a M, Taylor MS, Engström PG, Frith MC, et al. 2006.** Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics* **38**: 626–635.
- Cerná A, Cuadrado A, Jouve N, Díaz De La Espina S, De La Torre C. 2004.** Z-DNA, a new in situ marker for transcription. *European Journal of Histochemistry* **48**: 49–56.
- Chinnaraj M, Strick TR, Ebright RH. 2013.** Flexibility in transcription start-site selection by RNA Polymerase involves transcription-bubble expansion (“scrunching”) or contraction (“unscrunching”). *Biophysical Journal* **104**: 585a–586a.
- Choy M-K, Movassagh M, Goh H-G, Bennett MR, Down TA, Foo RSY. 2010.** Genome-wide conserved consensus transcription factor binding motifs are hyper-methylated. *BMC Genomics* **11**: 519.
- Cianfrocco MA, Nogales E. 2013.** Regulatory interplay between TFIID’s conformational transitions and its modular interaction with core promoter DNA. *Transcription* **4**: 1–7.
- Cremer T, Cremer M. 2010.** Chromosome territories. *Cold Spring Harbor Perspectives in Biology* **2**: a003889.
- Creux NM, De Castro MH, Ranik M, Mathabatha MF, Myburg AA. 2013.** Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in *Eucalyptus*. *Tree Genetics & Genomes* **9**: 989–1004.
- Creux N, De Castro M, Ranik M, Spokevicius A, Bossinger G, Maritz-Olivier C, Myburg Z. 2011.** In silico and functional characterization of the promoter of a *Eucalyptus* secondary cell wall associated cellulose synthase gene (*EgCesA1*). *BMC Proceedings* **5**: P107.

- Deng B, Melnik S, Cook PR. 2013.** Transcription factories, chromatin loops, and the dysregulation of gene expression in malignancy. *Seminars in Cancer Biology* **23**: 65-71.
- Dong X, Greven MC, Kundaje A, Djebali S, Brown JB, Cheng C, Gingeras TR, Gerstein M, Guigó R, Birney E, et al. 2012.** Modeling gene expression using chromatin features in various cellular contexts. *Genome Biology* **13**: R53.
- Dong D, Shao X, Zhang Z. 2011.** Differential effects of chromatin regulators and transcription factors on gene regulation: a nucleosomal perspective. *Bioinformatics* **27**: 147–152.
- Du Z, Li H, Wei Q, Zhao X, Wang C, Zhu Q, Yi X, Xu W, Liu XS, Jin W, et al. 2013.** Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa* L. japonica. *Molecular Plant* **6**: 1463–1472.
- Dundr M, Hoffmann-Rohrer U, Hu Q, Grummt I, Rothblum LI, Phair RD, Misteli T. 2002.** A kinetic framework for a mammalian RNA polymerase in vivo. *Science* **298**: 1623–1626.
- Fraser P, Bickmore W. 2007.** Nuclear organization of the genome and the potential for gene regulation. *Nature* **447**: 413–417.
- Gagniuc P, Ionescu-Tirgoviste C. 2013.** Gene promoters show chromosome-specificity and reveal chromosome territories in humans. *BMC genomics* **14**: 278.
- Geggier S, Vologodskii A. 2010.** Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 15421–15426.
- Geisler M, Kleczkowski L a, Karpinski S. 2006.** A universal algorithm for genome-wide in silico identification of biologically significant gene promoter putative *cis*-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in *Arabidopsis*. *The Plant Journal* **45**: 384–398.
- Grattapaglia D, Kirst M. 2008.** *Eucalyptus* applied genomics: from gene sequences to breeding tools. *The New Phytologist* **179**: 911–929.
- Guenther MG, Levine SS, Boyer L a, Jaenisch R, Young R a. 2007.** A chromatin landmark and transcription initiation at most promoters in human cells. *Cell* **130**: 77–88.
- Hakim O, Sung M-H, Hager GL. 2010.** 3D shortcuts to gene regulation. *Current Opinion in Cell Biology* **22**: 305-313.
- Van Heeringen SJ, Akhtar W, Jacobi UG, Akkers RC, Suzuki Y, Veenstra GJC. 2011.** Nucleotide composition-linked divergence of vertebrate core promoter architecture. *Genome Research* **21**: 410–421.
- Hieno A, Naznin HA, Hyakumachi M, Sakurai T, Tokizawa M, Koyama H, Sato N, Nishiyama T, Hasebe M, Zimmer AD, et al. 2014.** ppdb: plant promoter database version 3.0. *Nucleic Acids Research* **42**: D1188–92.

- De Hoon M, Hayashizaki Y. 2008.** Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *BioTechniques* **44**: 627–632.
- Hoskins R a, Landolin JM, Brown JB, Sandler JE, Takahashi H, Lassmann T, Yu C, Booth BW, Zhang D, Wan KH, et al. 2010.** Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Research* **21**: 182–192.
- Hughes JD, Estep PW, Tavazoie S, Church GM. 2000.** Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* **296**: 1205–1214.
- Hussey SG, Mizrachi E, Creux NM, Myburg AA. 2013.** Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 325.
- Irimia M, Maeso I, Roy SW, Fraser HB. 2013.** Ancient *cis*-regulatory constraints and the evolution of genome architecture. *Trends in Genetics* **29**: 521–528.
- Juanita B, Mun C. 2011.** Bioethanol production from tension and opposite wood of *Eucalyptus globulus* using organosolv pretreatment and simultaneous saccharification and fermentation. *Journal Of Industrial Microbiology* **38**: 1861–1866.
- Kadonaga JT. 2012.** Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**: 40–51.
- Kaern M, Elston TC, Blake WJ, Collins JJ. 2005.** Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**: 451–464.
- Kanai A, Suzuki K, Tanimoto K, Mizushima-Sugano J, Suzuki Y, Sugano S. 2011.** Characterization of STAT6 target genes in human B cells and lung epithelial cells. *DNA Research* **18**: 379–392.
- Kanhere A, Bansal M. 2005.** Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Research* **33**: 3165–3175.
- Kapanidis AN, Margeat E, Ho SO, Kortkhonjia E, Weiss S, Ebright RH. 2006.** Initial transcription by RNA polymerase proceeds through a DNA scrunching mechanism. *Science* **314**: 1144–1147.
- Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009.** The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362–366.
- Klug A, Lutter LC. 1981.** The helical periodicity of DNA on the nucleosome. *Nucleic Acids Research* **9**: 4267–4284.
- Kouzine F, Levens D. 2007.** Supercoil-driven DNA structures regulate genetic transactions. *Frontiers in Bioscience* **12**: 4409–4423.

- Kravatskaya GI, Chechetkin VR, Kravatsky Y V, Tumanyan VG. 2013.** Structural attributes of nucleotide sequences in promoter regions of supercoiling-sensitive genes: How to relate microarray expression data with genomic sequences. *Genomics* **101**: 1–11.
- Kwak H, Fuda NJ, Core LJ, Lis JT. 2013.** Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339**: 950–953.
- Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004.** Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics* **36**: 900–905.
- Lenhard B, Sandelin A, Carninci P. 2012.** Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* **13**: 233–245.
- Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van de Peer Y, Rouzé P, Rombauts S. 2002.** PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Research* **30**: 325–327.
- Lionnet T, Singer RH. 2012.** Transcription goes digital. *EMBO Reports* **13**: 313–321.
- Liu X, Brutlag D, Liu J. 2002.** An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology* **20**: 835–839.
- Maizels N, Gray LT. 2013.** The G4 genome. *PLoS Genetics* **9**: e1003468.
- Maruyama K, Todaka D, Mizoi J, Yoshida T, Kidokoro S, Matsukura S, Takasaki H, Sakurai T, Yamamoto YY, Yoshiwara K, et al. 2012.** Identification of *cis*-acting promoter elements in cold- and dehydration-induced transcriptional pathways in *Arabidopsis*, rice, and soybean. *DNA Research* **19**: 37–49.
- Mizrachi E, Hefer C a, Ranik M, Joubert F, Myburg A a. 2010.** *De novo* assembled expressed gene catalog of a fast-growing *Eucalyptus* tree produced by Illumina mRNA-Seq. *BMC Genomics* **11**: 681.
- Mizrachi E, Mansfield SD, Myburg AA. 2012.** Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54–62.
- Mogno I, Vallania F, Mitra RD, Cohen BA. 2010.** TATA is a modular component of synthetic promoters. *Genome Research* **20**: 1391–1397.
- Mokry M, Hatzis P, Schuijers J, Lansu N, Ruzius F-P, Clevers H, Cuppen E. 2012.** Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes. *Nucleic Acids Research* **40**: 148–158.
- Molina C, Grotewold E. 2005.** Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* **6**: 1471–2164.

Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, Fernie AR, Usadel B, Nikoloski Z, Persson S. 2011. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. *The Plant Cell* **23**: 895-910.

Mutwil M, Ruprecht C, Giorgi FM, Bringmann M, Usadel B, Persson S. 2009. Transcriptional wiring of cell wall-related genes in *Arabidopsis*. *Molecular Plant* **2**: 1015–1024.

Myburg A, Bradfield J, Cowley E, Creux N, De Castro M, Hatherell T-L, Mphahlele M, O'Neill M, Ranik M, Solomon L, et al. 2008. Forest and fibre genomics: biotechnology tools for applied tree improvement. *Southern Forests* **70**: 59–68.

Narlikar L, Gordân R, Hartemink AJ. 2007. A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Computational Biology* **3**: e215.

Pai DA, Engelke DR. 2010. Spatial organization of genes as a component of regulated expression. *Chromosoma* **119**: 13–25.

Papantonis A, Cook PR. 2010. Genome architecture and the role of transcription. *Current Opinion in Cell Biology* **22**: 271–276.

Pavesi G, Mereghetti P, Mauri G, Pesole G. 2004. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* **32**: W199–203.

Persson S, Wei H, Milne J, Page GP, Somerville CR. 2005. Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 8633–8638.

Priest HD, Filichkin SA, Mockler TC. 2009. *Cis*-regulatory elements in plant cell signaling. *Current Opinion in Plant Biology* **12**: 643–649.

Prohaska SJ, Stadler PF, Krakauer DC. 2010. Innovation in gene regulation: the case of chromatin computation. *Journal of Theoretical Biology* **265**: 27–44.

Ranik M, Myburg AA. 2006. Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiology* **26**: 545–556.

Rencoret J, Gutiérrez A, Nieto L, Jiménez-Barbero J, Faulds CB, Kim H, Ralph J, Martínez AT, Del Río JC. 2011. Lignin composition and structure in young versus adult *Eucalyptus globulus* plants. *Plant Physiology* **155**: 667–682.

Rengel D, San Clemente H, Servant F, Ladouce N, Paux E, Wincker P, Couloux A, Sivadon P, Grima-Pettenati J. 2009. A new genomic resource dedicated to wood formation in *Eucalyptus*. *BMC Plant Biology* **9**: 36.

Robb NC, Cordes T, Hwang LC, Gryte K, Duchi D, Craggs TD, Santoso Y, Weiss S, Ebricht RH, Kapanidis AN. 2013. The transcription bubble of the RNA polymerase-promoter open complex exhibits conformational heterogeneity and millisecond-scale

dynamics: implications for transcription start-site selection. *Journal of Molecular Biology* **425**: 875–885.

Rosin D, Hornung G, Tirosh I, Gispan A, Barkai N. 2012. Promoter nucleosome organization shapes the evolution of gene expression. *PLoS Genetics* **8**: e1002579.

Sáez-Vásquez J, Gadal O. 2010. Genome organization and function: a view from Yeast and *Arabidopsis*. *Molecular Plant* **3**: 678–690.

Salimullah M, Mizuho S, Plessy C, Carninci P. 2011. NanoCAGE: A high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harbor Protocols* **2011**: 5559.

Sawaya S, Bagshaw A, Buschiazzi E, Kumar P, Chowdhury S, Black M a, Gemmell N. 2013. Microsatellite tandem repeats are abundant in human promoters and are associated with regulatory elements. *PloS One* **8**: e54710.

Schmid CD, Praz V, Delorenzi M, Perier R, Bucher P. 2004. The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Research* **32**: D82–D85.

Seila AC, Core LJ, Lis JT, Sharp P a. 2009. Divergent transcription: a new feature of active promoters. *Cell Cycle* **8**: 2557–2564.

Shandilya J, Roberts SGE. 2012. The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochimica et Biophysica Acta* **1819**: 391–400.

Street NR, Jansson S, Hvidsten TR. 2011. A systems biology model of the regulatory network in *Populus* leaves reveals interacting regulators and conserved regulation. *BMC Plant Biology* **11**: 13.

Su J, Teichmann S a., Down T a. 2010. Assessing computational methods of *cis*-regulatory module prediction. *PLoS Computational Biology* **6**: e1001020.

Sutherland H, Bickmore WA. 2009. Transcription factories: gene expression in unions? *Nature Reviews Genetics* **10**: 457–466.

Swamy KBS, Chu W-Y, Wang C-Y, Tsai H-K, Wang D. 2011. Evidence of association between nucleosome occupancy and the evolution of transcription factor binding sites in yeast. *BMC Evolutionary Biology* **11**: 150.

Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y. 2001. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* **17**: 1113–1122.

Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE. 2004. Decoding human regulatory circuits. *Genome Research* **14**: 1967–1974.

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, et al. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Research* **19**: 255–265.

Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. 2009. Unraveling transcriptional control in *Arabidopsis* using *cis*-regulatory elements and coexpression networks. *Plant Physiology* **150**: 535–546.

Venters BJ, Pugh BF. 2013. Genomic organization of human transcription initiation complexes. *Nature* **502**: 53–58.

Wagner A. 1999. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15**: 776.

Walther D, Brunnemann R, Selbig J. 2007. The regulatory code for transcriptional response diversity and its relation to genome structural properties in *A. thaliana*. *PLoS Genetics* **3**: e11.

Wasson T, Hartemink AJ. 2009. An ensemble model of competitive multi-factor binding of the genome. *Genome Research* **19**: 2101–2012.

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Prüss M, Reuter I, Schacherer F. 2000. TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Research* **28**: 316–319.

Wingender E, Dietze P, Karas H, Knüppel R. 1996. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Research* **24**: 238–241.

Wray GA. 2007. The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics* **8**: 206–216.

Xu M, Cook PR. 2008. Similar active genes cluster in specialized transcription factories. *The Journal of Cell Biology* **181**: 615–623.

Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J. 2007a. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Research* **35**: 6219–6226.

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T. 2007b. Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**: 67.

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. 2009. Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *The Plant Journal* **60**: 350–362.

Yuzenkova Y, Tadigotla VR, Severinov K, Zenkin N. 2011. A new basal promoter element recognized by RNA polymerase core enzyme. *The EMBO Journal* **30**: 3766–3775.

- Zeigler RD, Gertz J, Cohen BA. 2007.** A *cis*-regulatory logic simulator. *BMC Bioinformatics* **8**: 272.
- Zhai HL, Wang XH, Huang XY, Shan ZJ. 2011.** The analysis of core promoter sequences based on their chemical features. *Chemometrics and Intelligent Laboratory Systems* **107**: 245–250.
- Zhao R, Bodnar MS, Spector DL. 2009.** Nuclear neighborhoods and gene expression. *Current Opinion in Genetics & Development* **19**: 172–179.
- Zhao X, Valen E, Parker BJ, Sandelin A. 2011.** Systematic clustering of transcription start site landscapes. *PloS One* **6**: e23409.
- Zhitnikova MY, Boryskina OP, Shestopalova A V. 2013.** Sequence-specific transitions of the torsion angle gamma change the polar-hydrophobic profile of the DNA grooves: implication for indirect protein-DNA recognition. *Journal of Biomolecular Structure & Dynamics* **1**: 37–41.
- Zhong R, Lee C, Ye Z-H. 2010.** Global analysis of direct targets of secondary wall NAC master switches in *Arabidopsis*. *Molecular Plant* **3**: 1087–1103.
- Zopf CJ, Quinn K, Zeidman J, Maheshri N. 2013.** Cell-cycle dependence of transcription dominates noise in gene expression. *PLoS Computational Biology* **9**: e1003161.
- Zou Y, Huang W, Gu Z, Gu X. 2011.** Predominant gain of promoter TATA Box after gene duplication associated with stress responses. *Molecular Biology and Evolution* **28**: 2893–2904.

1.9 TABLES

Table 1.1. IUPAC nucleotide codes. Table showing the structures for each of the four nucleotides present in DNA, as well as the IUPAC codes for nucleotide groups.

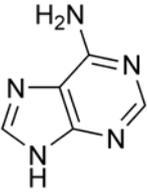
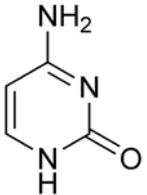
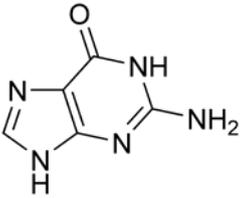
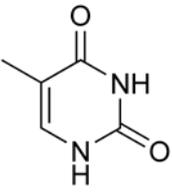
Iupac Code	Biochemical Properties/Similarity	Nucleotides
A	 https://commons.wikimedia.org/wiki/File:Adenine_chemical_structure.png	Adenine
C	 https://commons.wikimedia.org/wiki/File:Cytosine_chemical_structure.png	Cytosine
G	 https://commons.wikimedia.org/wiki/File:Guanine_chemical_structure.png	Guanine
T	 https://commons.wikimedia.org/wiki/File:Thymine_chemical_structure.png	Thymine
M	Amino (C-N-H ₂) groups can form imino tautomers (rare)	A; C
W	Form weak hydrogen bonds (2H)	A; T
R	Purine (2 heterocyclic rings)	A; G
Y	Pyrimidine (1 heterocyclic ring)	C; T
S	Form strong hydrogen bonds (3H)	C; G
K	Keto (C=O) groups can form enol tautomers (rare)	G; T
B	not A (B comes after A)	C; G; T
D	not C (D comes after C)	A; G; T
H	not G (H comes after G)	A; C; T
V	not T (V comes after T and U)	A; C; G
N	Any	A; C; G; T

Table 1.2. Applicable research publications in which core promoter elements are described. Advancement of core promoter research using various experimental procedures and bioinformatics techniques.

	Author	Organism(s)	Method	Background	Main findings
a)	Kanhere & Bansal, 2005	2540 vertebrate and 198 plant promoters from Eukaryotic Promoter Database (Schmid <i>et al.</i> , 2004); <i>Escherichia coli</i> ; <i>Bacillus subtilis</i>	Predicted stability curvature and bendability of -1000 to -1 and +1 to +500 promoter regions.	[-1000, -1] and [+1, +500] shuffled to retain mononucleotide frequency.	Promoter regions are less stable and less bendable, but contain elements to enhance topologically constrained bendability
b)	Molina & Grotewold, 2005	<i>Arabidopsis thaliana</i>	Motifs predicted by MEME (Bailey <i>et al.</i> , 2006) and AlignAce (Hughes <i>et al.</i> , 2000) in [-50, -1] and [+1,+50] of 12,749 promoters. +1 determined by fl-cDNA clones.	50bp shuffling of mononucleotide frequencies ~65% A/T to ~35% C/G and ~61% A/T to ~39% C/G of [-50, -1] and [+1,+50] respectively.	TATA-Box is less prevalent than previous estimates for plants. Describe 20 over-represented motifs including the R and Y repeat sequences.
c)	Yamamoto <i>et al.</i> , 2007b	<i>Arabidopsis thaliana</i> ; <i>Oryza sativa</i>	Smoothing window of 15 and 21 bps for hexamers and octamers respectively. Peak height and area above base line used to determine significance.	Mean and standard deviation of octamer occurrence calculated from 1000 random promoter samples	YR rule at the TSS and the Y repeat patch occurring at -13. TATA box is conserved at -35 and ~300 peaks in [-200,-51] are putative TFBSs for modulators.
d)	Brick <i>et al.</i> , 2008	<i>Plasmodium falciparum</i>	Measured 59 physicochemical properties of DNA are assessed in -100 to +50 promoter regions.	Exonic and intergenic.	Physicochemical properties are predictive of core promoters. TA at TSS, following YR (PyPu) rule.
e)	Yamamoto <i>et al.</i> , 2009	<i>Arabidopsis thaliana</i> ;	[-1000, +200] to analyse promoters derived from publicly available TSSs by LDSS (Yamamoto <i>et al.</i> , 2007a). CT-MPSS to identify TSSDs.	Mean and standard deviation of kmers calculated from 1000 random promoter samples	GAGAGA and CACACA repeat elements downstream of TSS. Confirmation of the Y-patch. Identification of plant Inr-like and Kozak sequence.

Continues on next page...

Table 1.2 continued. List of applicable publications in which core promoter elements are described. Advancement of core promoter research using various experimental procedures and bioinformatics techniques.

	Author	Organism(s)	Method	Background	Main findings
f)	Bernard <i>et al.</i> , 2010	<i>Arabidopsis thaliana</i>	99% confidence interval determined by linear regression of background occurrence of hexamers. Searched for non-even distribution above confidence interval in [-300,+500]	[-1000,-300] of 14927 <i>A. thaliana</i> promoters	TC element enriched at [-39,-26] in TATA-less promoters of <i>A. thaliana</i> , which are not present in metazoans and are not conserved in orthologous gene pairs in higher plants (<i>O. sativa</i>)
g)	Zhao <i>et al.</i> , 2011	<i>Homo sapiens</i> ; <i>Mus musculus</i>	Searched for TATA position weight matrix in [-100,100] and CpG repeats in [-300,300] with 0 the dominant peak of the TSSD.	NA	CpG islands are associated with broad TSSDs, whereas TATA-Box, Inr, DPEu/d are associated with sharp, narrow TSSDs.
h)	van Heeringen <i>et al.</i> , 2011	<i>Homo sapiens</i> ; <i>Xenopus tropicalis</i>	MEME (Bailey <i>et al.</i> , 2006) Motif-Sampler (Thijs <i>et al.</i> , 2001), Weeder (Pavesi <i>et al.</i> , 2004) and MDmodule (Liu <i>et al.</i> , 2002) occurrence of $>=2$ sd above background with bin size of 20. Promoter set from [-400,+100]	Random sequence generated by first order markov model derived from promoter set.	Used TSS-seq and CAGE to identify dominant TSSs. Identified 24 enriched motifs in <i>X. tropicalis</i> , some of which are conserved in vertebrates.
i)	Maruyama <i>et al.</i> , 2012	<i>Arabidopsis thaliana</i> ; <i>Oryza sativa</i> ; <i>Glycine max</i>	Z-test comparison of cold and dehydration response gene's promoters (microarray determined) with the expected frequency in randomly sampled promoter in 50bp windows [-100,-51], [-50,-1] and [+1,+50]	1000 random samplings of promoter per species (n=100)	W residues are over-represented in promoter sequences, rice and soybean also contain over-represented S residues. WR hexamers are under-represented expect for TATATA and TATAAA. YR hexamers are overrepresented in [-50,-1]
j)	Venters & Pugh, 2013	<i>Homo sapiens</i>	Determined TBP and TFIID binding by ChIP-exo. Searched for TATA consensus in 6,511 promoters, with up to 3 mismatches within 80 bp of sense-strand TFIIB binding signal midpoint. BREu, BREd and Inr were searched for within -40, +40 and +60 from 5,546 TATA box hits.	NA	Estimate 500,000 promoter initiator complexes occur in human genome. ~85% of tested promoters contain 0-3 mismatch of TATA box, while only 3% contain canonical TATA box. Confirmed co-occurrence with BREu/d and Inr.

CHAPTER 2 : Empirical curation of 5' UTRs in the *Eucalyptus grandis* genome

Ida C van Jaarsveld^{1,2}, Eshchar Mizrachi², Fourie Joubert¹ & Alexander A Myburg²

1 Bioinformatics and Computational Biology Unit, Department of Biochemistry, Genomics Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa

2 Department of Genetics, Genomics Research Institute (GRI), Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa

This chapter is formatted as a research article for a peer-reviewed journal (New Phytologist). Much of this content has been accepted by Nature as supplementary material in “Genome sequence of *Eucalyptus grandis*: A global tree crop for fiber and energy” (Myburg *et al.*, in press). AA Myburg conceptualized the study, E Mizrachi facilitated the study design, IC van Jaarsveld designed the study and methodology, wrote all custom scripts and workflows, implemented the design, interpreted the results and wrote the manuscript. AA Myburg, E Mizrachi and F Joubert approved and supervised the study and reviewed the manuscript.

2.1 SUMMARY

- The first release of the *Eucalyptus grandis* genome sequence and annotation motivates the need for gene model curation. The 5' untranslated regions (UTRs) lack intrinsic DNA signatures for annotation, and are more divergent than their protein-coding counterparts. Only 57% of the Phytozome V1.0 *E. grandis* gene models are equipped with an annotated 5' UTR and we expect that a significant proportion of these are truncated. We hypothesise that empirical transcript evidence can improve the number and quality of 5' UTR annotations.
- ~2.9 million *E. grandis* ESTs and ~700 million paired-end mRNA-seq reads from seven diverse *E. grandis* tissues and organs were used to empirically curate a maximum number of *E. grandis* 5' UTRs. A prioritized composite set of annotations comprises the longest empirical 5' UTR, and in cases of no empirical evidence and a *FGH* prediction, the predicted 5' UTR.
- RNA-seq annotated 5' UTRs were annotated for 22,983 genes, whereas only 8,040 were annotated by ESTs. Using high throughput empirical curation from these two sources, it was possible to substantiate or improve 17,085 5' UTR annotations, and add 7,596 which had no prior annotation.
- The curated 5' UTRs allow the inference of high confidence distal transcription start sites, which can be used in the downstream assessment of promoters and transcriptional regulatory features of this economically important fibre crop species.

2.2 INTRODUCTION

Confidence in an emerging genome annotation is instilled through curation of functionally relevant features by expert investigation and empirical evidence. The Online Resource for Community Annotation of Eukaryotes (ORCAE; Sterck *et al.*, 2012) provides a curation platform for the *Eucalyptus grandis* genome annotation. ORCAE allows for the browsing, interrogation and correction of annotated functional units at nucleotide resolution. The functional unit of interrogation in this chapter is the 5' untranslated region (UTR). The *E. grandis* 5' UTRs (Phytozome V1.0) were predicted simultaneously with protein-coding loci by FGenesH (Solovyev *et al.*, 2006). Briefly, this process isolates genomic regions spanning

1kb either side of aligned ESTs, and together with angiosperm protein sequences, are submitted to both intrinsic and extrinsic methods of *in silico* gene prediction. *In silico* 5' UTR annotation uses *Arabidopsis*-derived genomic features such as the TATA-Box and oligomer frequencies to computationally assign putative transcription start sites (TSSs) and thus delimit the 5' UTR (Shahmuradov *et al.*, 2005; Solovyev *et al.*, 2006) within the 1kb extended genomic regions of EST alignment. False positive TSS annotation estimates for this approach range between 1 per 700 to 1000 bp, and there is a bias toward TATA promoter sensitivity (Shahmuradov *et al.*, 2005).

In an attempt to circumvent poor performance, particularly in cases of multiple possible TSSs per locus, the most 3' annotation is preferred, possibly truncating the 5' UTR model. Despite the high false positive rate, gains in sensitivity are not adequately achieved, as is evident with only ~57% of the current *E. grandis* gene models assigned a 5' UTR. Sensitivity is impeded as the TATA-box, a key signature in TSS prediction, occurs less frequently than previously thought, and is absent in the majority of both metazoan and plant genes (Lenhard *et al.*, 2012). The poor sensitivity is thus a result of the subtle intrinsic DNA signatures, which results in proximal promoters not being well described, particularly in non-model organisms. TSSs are however empirically, and thus more reliably, distinguishable by transcript support such as ESTs (Kan *et al.*, 2000; Haas, 2003), full length cDNA (Molina & Grotewold, 2005; Tanaka *et al.*, 2009; Soderlund *et al.*, 2009) and next-generation sequencing (NGS) RNA-seq (Zhang *et al.*, 2010) data.

The expectation of promoter sequence complexity has grown proportionately with the advancements of quantitative and descriptive transcriptomics and the number of sequenced genomes (Zuo & Li, 2011; Lenhard *et al.*, 2012; Liu *et al.*, 2013). Transcription is seldom initiated at a single discrete coordinate, but rather over a genomic region, of which the range and distribution differ significantly between genes and their tissue, developmental and response specific expression (Balwierz *et al.*, 2009; Valen *et al.*, 2009; Yamamoto *et al.*, 2009; Ni *et al.*, 2010; Zhao *et al.*, 2011). This is compounded by a high level of stochasticity influencing the positioning and success of transcription initiation (Kaern *et al.*, 2005; Shandilya & Roberts, 2012). High confidence TSS distribution (TSSD) (Zhao *et al.*, 2011) annotation is still rare in plants and requires a specific transcript-preparation protocol such as DeepCAGE (Valen *et al.*, 2009). A useful co-ordinate in the TSSD is the most 5' position, as it represents the first position at which DNA transcription is permissive at a particular gene.

Studying these promoter and distal transcription start site (dTSS) features can provide valuable insight as to the underlying mechanisms of transcription initiation. The most 5' evidence of continuous high-quality, full-length expressed transcripts empirically identifies (Haas, 2003; Molina & Grotewold, 2005) the distal co-ordinates of TSSDs, or the “dTSS”.

Genome annotation is a continuous process, with manual curation crucial to the confidence in feature description and respective functional inference. ORCAE (Sterck *et al.*, 2012) demonstrates discordance between current 5' UTR annotations and empirical transcript evidence. Where the 5' UTR annotations are unsubstantiated by empirical data, they may be absent, truncated or extended. Typically, 5' UTRs are challenging to annotate *in silico* as they lack intrinsic annotation signatures of other feature-rich gene components such as exons, and are extrinsically more divergent from homologous gene models than their protein-coding counterparts. They are thus more prone to misannotation. However, there is extensive *E. grandis* empirical transcript evidence (full-length mRNA-seq) that has been generated for the expression analysis of diverse developing tree tissues (Hefer *et al.*, in preparation). This illustrates not only the necessity, but the opportunity to validate or correct current 5' UTR annotations by use of available transcript evidence.

A prominent focus of *E. grandis* research, particularly that of the Forest Molecular Genetics Programme at the University of Pretoria, is the experimentation and detailed description of transcriptional regulation relating to wood formation in trees (Creux *et al.*, 2011, 2013; Mizrachi *et al.*, 2012). Promoters are fundamental cis-regulatory loci at which transcriptional inputs are assembled, ultimately driving gene expression. The *E. grandis* v1.1 gene models are used to intricately assess the promoters of genes which yield key wood property, growth and resistance traits. The inference of these promoter regions, and the extraction of their respective DNA sequence, hinges on the TSS annotations. Thus the integrity of promoter analysis results is highly dependent on accurate TSS annotations. To extract maximum value from downstream analyses, it is essential that TSSs are curated and that the annotations used are empirically substantiated.

In the absence of 5' captured full-length mRNA or the specific isolation of 5' tags, we hypothesise that mRNA-seq data is adequate for the curation, encompassing both validation and correction, of *in silico* annotated *E. grandis* 5' UTRs. We expect that the most 5' evidence of transcription infers a distal transcription start site and a maximally inclusive 5'

UTR annotation. We have developed a method for the systematic high-throughput curation of dTSSs, which are central to promoters, core regulatory loci. The prioritised inclusion of mRNA-seq and EST data improves both the number and quality of 5' UTR annotations, providing the genomic co-ordinates for future analyses of regulatory loci in the *E. grandis* genome, including Chapter 3.

2.3 MATERIALS AND METHODS

2.3.1. Genomic and RNA transcription data acquisition

The Phytozome (Goodstein *et al.*, 2012) v8.0 (“Egrandis_201”) unmasked genome assembly and gene-exon gff3 file (V1.0) were downloaded from the Phytozome *E. grandis* ftp repository (ftp://ftp.jgi-psf.org/pub/JGI_data/phytozome/v8.0/Egrandis/). Approximately 2.9 million *E. grandis* ESTs (454 reads), assembled by PASA (Haas, 2003) into 152,670 locus-associated EST assemblies, were downloaded per request from the ORCAE curation platform (Sterck *et al.*, 2012). EST alignment was performed using GenomeThreader (Gremme *et al.*, 2005) with 95% similarity and 90% identity thresholds (Sterck *et al.*, 2012). These EST annotations were preferred over those from Phytozome’s GBrowse (Stein *et al.*, 2002) instance as i) the *E. grandis* and sister *Eucalyptus*’ ESTs are assembled separately, thus reducing noise in the dataset (Supplementary Figure 2.3); ii) the ESTs are associated with *E. grandis* loci *a priori*, thus reducing the amount of required parsing and processing; and iii) the EST splice boundary coordinates are available as opposed to just the start and end. 5' UTRs predicted by FGenesH (Solovyev *et al.*, 2006) and defined by the Phytozome (Goodstein *et al.*, 2012) gene models are referred to as “FGH” 5' UTRs in the remainder of the text; those inferred by the ORCAE (Sterck *et al.*, 2012) PASA (Haas, 2003) assembled *E. grandis* EST data are referred to as “PASA” 5' UTRs; those inferred by mRNA-seq, as “NGS” 5' UTRs; and the most 3' initiation codon per locus is referred to as “INIT”.

2.3.2 Annotation of 5' UTRs using mRNA-seq data

Paired-end Illumina mRNA-seq data for shoot tip, young leaf, mature leaf, flower, root, phloem and immature xylem tissues (Hefer *et al.* in preparation) were mapped individually to the genome. Approximately 750 million, paired-end, 75 nt reads were aligned to the *E. grandis* genome using TopHat (Trapnell *et al.*, 2009) with default parameters. The process to delimit 5' UTRs from mRNA-seq data is described in Supplementary Note 2.1. Briefly, the

resulting SAM alignment files were merged into a bulk alignment file using `SAMtools` (Li *et al.*, 2009), which was subsequently filtered for a mapping quality of at least two (three or less possible hits in the genome) and for introns less than 6500 bp, as extracted by the extended CIGAR (Li *et al.*, 2009) string. The remaining reads were reconverted to BAM format. `Igvttools` (Thorvaldsdóttir *et al.*, 2012) was then used to convert the BAM to the per-base coverage wiggle format. All possible splicing events were extracted from the CIGAR strings, which were submitted, in unison with the bulk BAM file, to `bam2ssj` (Pervouchine *et al.*, 2012) for splice junction quantification. Those splice-events with a splicing index and percent-spliced-in (PSI) value of at least 0.4, and raw splice-junction coverage of at least 5, were retained. The wiggle representing coverage and high confidence splice junctions was then used to delimit the 5' UTRs. The following exclusionary criteria were applied to NGS coverage data: i) overlapping gene models; ii) continuous coverage spanning more than 9 kb; iii) continuous 5' coverage into an upstream gene model; iv) coverage of less than 5 at *INIT*; and v) a 5' UTR splice junction spanning to an adjacent gene model. If the above criteria at a given locus were permissive, the NGS 5' UTR regions were defined by: $NGS \in \{kc_1, kc_2, kc_3 \dots kc_n\}$ where k is the normalisation constant defined by $k = 1/r$; r is the raw coverage at *INIT*; and c is the raw depth of coverage at positions $\{1, 2, 3 \dots n\}$ upstream of *INIT*, so that $c \geq 1$. Thus, a normalised array was built from *INIT* upstream until a discontinuation in coverage. Areas of zero coverage were permitted if they were bounded by a significantly supported splice junction unique to that respective gene model.

2.3.3 Length comparison of predicted and empirically derived 5' UTRs

The *FGH* 5' UTR regions were extracted from the `gff3` as: the region from the most 5' exon position to and including *INIT* (Supplementary Note 2.2). The *PASA* 5' UTRs were extracted from the tabular *PASA* file (Sterck *et al.*, 2012) if a given locus-associated EST assembly overlapped the most downstream initiation codon of the respective locus (Supplementary Note 2.3). Again, the 5' UTR is defined from the most 5' position of the EST assembly, to and excluding *INIT*. Pairwise comparisons of 5' UTR length distributions and per-gene 5' UTR length correlation were tested using Kolmogorov-Smirnoff and Pearson correlation respectively.

2.4 RESULTS

2.4.1 mRNA-seq transcript evidence extends upstream of predicted 5' UTRs

At each locus, the *FGH*, *PASA* and *NGS* 5' UTR lengths were compared. From these lengths, 5' UTR annotation sources were prioritized per locus (n). An empirical annotation (*PASAn*; *NGSn*) is prioritized over *FGHn*. The longest empirical transcript was preferred, i.e. *PASAn*>*NGSn* or *NGSn*>*PASAn* (Figure 2.1a). Those loci which have a 5' UTR reported by only *FGH* retain their *FGH* annotation as the best current annotation. Distal transcription start sites (dTSSs) may be inferred by the first position of the prioritised 5' UTRs, with greater confidence associated with those that are empirically substantiated, and, if by *NGS*, those with higher coverage at *INIT*, the scores of which are detailed in gff3 format (Supplementary Note 2.1, line 30).

Of the 20,760 *E. grandis* loci annotated with an *in silico* predicted 5' UTR, 35% (7,306) of loci were also represented by *PASA* data, whereas 75% (15,569) were also represented by *NGS*. The co-occurrence of 5' UTR annotation per locus is depicted in Figure 2.1b, showing 5,790 loci are represented by all three sources. Despite the co-evidence at loci, the respective annotations differ significantly in length (Kolmogorov-Smirnoff test, $p < 0.05$). The discrepancy between coordinates is calculable by the resultant 5' UTR lengths. Only 519 loci had annotations of a similar length (standard deviation < 5 bp). The pairwise comparison of 5' UTR lengths (Table 2.1) indicates that the majority (63%) of *in silico* predictions were unsubstantiated by the empirical evidence used to derive them. However, near half (56%) of all loci that were annotated by both *FGH* and *PASA* were of equal length. This is not surprising due to the prediction algorithm (FGenesH) using the *PASA* evidence, in combination with intrinsic gene composition signals, to predict and define the gene models and 5' UTRs. Comparing the two empirical sources, *NGS* supported an additional 16,641, and the majority (86%) of those loci co-annotated by *NGS* and *PASA* were extended by *NGS*. In comparing *NGS* to *FGH*, although the majority (86%) of the *in silico* annotations were co-annotated and extended by *NGS*, there is considerable value in retaining the *FGH* annotations, as 5,191 had no *NGS* complement. In *Arabidopsis thaliana*, in which only ~61% of gene annotations contain a 5' UTR (Lamesch *et al.*, 2012), the mean and median 5' UTR lengths are 299 and 118 bp respectively (Supplementary Figure 2.2). These are shorter than the lengths reported for *E. grandis* *FGH* 5' UTRs (347; 154), as well *NGS* (539; 283) and

PASA (533; 151). The Pearson Correlation Coefficient between *FGH* and *PASA* 5' UTR lengths is 0.55, 0.28 between *NGS* and *FGH*, and only 0.20 between *PASA* and *NGS*, the length distributions of which are depicted in Figure 2.2. The low r^2 values (0.3, 0.08 and 0.04 respectively), representing per-gene length similarity, indicate the extent of discordance between 5' UTR annotations between the three data sources. As hypothesized, the *FGH* models are significantly shorter (Kolmogorov-Smirnov test, $p < 0.05$), indicating possible truncation of *in silico* models. This was expected as FGenesH (Solovyev *et al.*, 2006) predictions favour those most proximal to the translation start site. mRNA-seq data is suited to splice junction detection and identified 2,095 high confidence 5' UTR introns. This exemplifies the ability for *NGS* data to detect longer transcript isoforms, including those which may be in minor abundance (see Additional file 2.1).

2.4.2 Empirical curation using expressed transcript data

Using high throughput empirical curation, it was possible to substantiate or improve the annotation of 17,085 5' UTRs, and annotate an additional 7,596 5' UTRs for genes that had no prior 5' UTR annotation. The final prioritized composite set of annotations comprises the longest empirical 5' UTR, and in cases of no empirical evidence and a *FGH* prediction, the predicted 5' UTR, proportions of which are detailed in Figure 2.3. We have augmented the total percentage of annotated 5' UTRs from 57% of all gene models to 78%, achieving the desired increase in both quality and number of 5' UTR annotations. Empirically annotated 5' UTRs allow greater confidence in the inference of dTSSs, yet *in silico* annotations remain invaluable *in lieu* of transcriptome data. The per scaffold and total contributions to the composite prioritised 5' UTR annotations are detailed in Figure 2.4 and Figure 2.5 respectively. There are relatively fewer 5' UTR annotations on the minor scaffolds when compared to major scaffolds (Figure 2.4), resulting from the high occurrence of gene annotations starting at the first position of minor scaffolds.

2.5 DISCUSSION

Eucalyptus grandis gene models possess *in silico* predicted 5' UTR annotations, of which 4,081 (~19%) are substantiated by empirical EST evidence. A method to delimit 5' UTRs by mRNA-seq and EST transcripts has been successfully designed and implemented, extending 13,431 5' UTR annotations, and identifying 7,596 for which there was no prior annotation.

The inclusion of mRNA-seq and EST empirical evidence has thus improved both the number and quality of 5' UTR annotations reported by Phytozome. These empirical annotations serve as a form of high-throughput curation, an essential quality-improvement process in genome annotation. These curated models have been made available to the *Eucalyptus* community by inclusion in the official *Eucalyptus grandis* genome publication (Myburg, *et al.*, in press). This is in the form of one amalgamated gff3 file (Additional file 2.2), describing 5' UTR annotations from each source (See Figure 2.6). Each annotation is supplied with an “extended” or “prioritised” tag, to allow unambiguous parsing and extraction (Supplementary Note 2.1).

There are three classes of expected NGS 5' UTR misannotation, these being omission, truncation, and extension. Omission affects approximately 37% of the loci and is the result of a locus either not being expressed or failing the filtering criteria (See Table 2.2 and Supplementary Note 2.1). Expression is absent at loci for several possible reasons. The first is that these loci are only expressed under specific growth or response conditions and are not represented in the current mRNA-seq data for sampled tissues. It is also possible, and expected at this primary stage of genome annotation, that several gene model predictions are incorrect and thus are not transcribed. However, the 22,983 genes which do possess NGS 5' UTRs are transcribed, indicating more prominent biological utility and thus more likely to be targets of downstream meta-analyses.

Omission is also possible in several scenarios where there is significant gene expression. Adjacent gene models in close proximity may have continuous read alignment spanning the intergenic region. In such cases, read assignment is ambiguous, and delimiting the termination of the transcript is not possible. Close proximity adjacency is not however an indicator for omission, as the method is able to discern such 5' UTRs (Figure 2.8). Genes in close proximity are not restricted to transcription initiation occurring in the intergenic region, a property inadvertently biased towards by this method, as promoter elements and enhancers may reside within an upstream gene, as is the case with *Populus trichocarpa Cesa7* (Bartel, personal communication). The distinctively high number of *E. grandis* gene models in tandem duplicate (35%, Myburg *et al.*, in press) encourages spliced mapping of mRNA-seq reads between adjacent and tandem gene models, resulting in NGS annotation omission by failing this filtering criterion. Gene model errors can also result in NGS annotation omission. In cases where the first exon is erroneously predicted to start further upstream of the true

initiation site, the genes are transcribed, but do not pass the depth of coverage filtering criterion at the false translation start site.

Truncation occurs when continuous mRNA-seq read alignment does not proceed as far as the true start of transcription. This can result at the molecular level as an artefact from random priming of poly(A)+ captured RNA, which has the propensity to under-represent 5' termini (Khrameeva & Gelfand, 2012). Given that the full transcript is indeed sequenced, filtering of redundant reads and low-confidence splice junctions can truncate models. Low complexity regions are common in promoters, and are intrinsic to their functionality, thus making it likely that these reads map to several regions in the genome and are thus filtered. Although truncating some models, this is a necessary criterion, as other loci would be susceptible to erroneous extension. In the case of splicing events, unfiltered PSI values resulted in cross-gene splicing, reducing the number of *NGS* 5' UTRs to 15,911. Thus, increasing the splicing stringency, although possibly truncating several models, ensured the inclusion of several thousand more. *NGS* 5' UTR extension occurs when an erroneous splice junction extends upstream into an intergenic region of aggregated low coverage or there is background coverage from unannotated gene models or redundant read alignment.

From the above three classes, it is evident that the efficacy of this method relies intrinsically on the quality of the genome annotation, the depth of sequencing and the diversity of conditions sampled, and the stringency of mRNA-seq mapping. The mRNA-seq alignment originally performed for expression quantification was not performed with sufficient stringency for per-base pair resolution studies. Hefer *et al.* (in preparation) used default mapping parameters for the alignment of the paired-end reads. These sub-optimal alignments could have been improved by limiting the number of genomic alignments allowed, increasing the segment size and reducing the maximum intron length. Although such parameters would have better suited this particular study, rather than running a computationally expensive procedure of realignment, and for the sake of consistency, filtering criteria were employed to isolate high-quality, low-redundancy reads and high-confidence splice junctions (Supplementary Note 2.1, lines 1-15). Furthermore, the tapering nature of transcription initiation (Additional file 2.3) doesn't allow for background correction such as a local dynamic lambda that one would use in peak-finding algorithms (Feng *et al.*, 2011). Here we specify a permissive read as one possessing a `TopHat` mapping quality of at least two, which in `PHRED` probability conversion allows three or fewer possible matches against the

reference genome, as a procedure for background noise reduction. The library preparation for the mRNA-seq experiment did not specifically capture nor enrich for RNA by 5' termini. It was often observed that samples had non-identical transcription start sites, as per most 5' aligned read. Merging and aggregating alignments from all tissues and replicates increased the depth of coverage of 5' termini per locus, and this was essential for obtaining above-threshold representation of 5' UTR reads.

Curation of 5' UTRs, as with other genomic features (Supplementary Figure 2.1), should proceed in the context of new and improving experimental technologies and analytical capabilities, and the ensuing availability of novel empirical evidence. The empirical annotations derived in this study would be supplemented by a targeted TSSD analysis, which will again, validate, augment and improve the current dTSSs. A more appropriate library preparation design would be preliminary RNA captured by poly(A)+, subsequently enriched for 5' caps (de Hoon & Hayashizaki, 2008; Yamamoto *et al.*, 2009; Salimullah *et al.*, 2011; Zhao *et al.*, 2011; Schlüter *et al.*, 2013). 5' tags can be sequenced to reduce cost and ensure higher coverage at the 5' terminus on a limited budget. Such an approach would specifically enrich for only full length mature mRNA and long non-coding RNA. Although this provides only steady state RNA abundances and is not the best possible proxy for transcription rate, it will ensure enrichment of 5' termini of full-length transcripts and thus facilitate TSSD annotation, which in itself is replete with challenges. Lastly, the method that has been developed to delimit 5' UTRs is appropriate for 3' UTR annotation, and although a suitable and opportunistic curative resource, was excluded as it was outside of the scope of this study.

2.6 CONCLUSION

mRNA-seq is traditionally used for expression profiling and the detection of sequence variants such as SNPs and indels. It is an invaluable and widely used experimental strategy as it can address such questions simultaneously. The technology is particularly useful for genome-wide hypothesis generating research in both model and non-model organisms. Here, we extend the utility of mRNA-seq data and describe a method where it is used to delimit distal transcription start sites. Genome annotation curation is an essential and iterative process. ORCAE (Sterck *et al.*, 2012) allows for curation at a per-gene resolution, but requires timeous community involvement and the necessary knowledge of gene composition and structure to identify and correct misannotations. This high-throughput empirical curation

effort provides researchers with examinable evidence, both empirical and predicted, of TSSs. This will contribute to researchers' cognisance of alternative TSS possibilities while formulating hypotheses and designing wet and dry-lab procedures which target promoter and 5' UTR regions of this economically important fibre crop species.

2.7 REFERENCES

Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E. 2009. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biology* **10**: R79.

Creux NM, De Castro MH, Ranik M, Mathabatha MF, Myburg AA. 2013. Diversity and cis-element architecture of the promoter regions of cellulose synthase genes in Eucalyptus. *Tree Genetics & Genomes* **9**: 989–1004.

Creux N, De Castro M, Ranik M, Spokevicius A, Bossinger G, Maritz-Olivier C, Myburg Z. 2011. In silico and functional characterization of the promoter of a Eucalyptus secondary cell wall associated cellulose synthase gene (EgCesA1). *BMC Proceedings* **5**: P107.

Feng J, Liu T, Zhang Y. 2011. Using MACS to Identify Peaks from CHIP-Seq Data. *Current Protocols in Bioinformatics*. Chapter 2, Unit 2.14.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–1186.

Gremme G, Brendel V, Sparks ME, Kurtz S. 2005. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**: 965–978.

Haas BJ. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**: 5654–5666.

De Hoon M, Hayashizaki Y. 2008. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *BioTechniques* **44**: 627–632.

Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**: 451–464.

Kan Z, Gish W, Rouchka E, Glasscock J, States D. 2000. UTR reconstruction and analysis using genomically aligned EST sequences. *Proceedings: International Conference on Intelligent Systems for Molecular Biology*. 218–227.

Khrameeva EE, Gelfand MS. 2012. Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments. *BMC Bioinformatics* **13 Suppl 6**: S4.

- Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al. 2012.** The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40**: D1202–1210.
- Lenhard B, Sandelin A, Carninci P. 2012.** Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* **13**: 233–245.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009.** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Liu Y, Yin J, Xiao M, Mason AS, Gao C, Liu H, Li J, Fu D. 2013.** Characterization of Structure, Divergence and Regulation Patterns of Plant Promoters. *Journal of Molecular Biology Research* **3**: 23–36.
- Mizrachi E, Mansfield SD, Myburg AA. 2012.** Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54–62.
- Molina C, Grotewold E. 2005.** Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* **6**: 1471–2164.
- Ni T, Corcoran DL, Rach EA, Song S, Spana EP, Gao Y, Ohler U, Zhu J. 2010.** A paired-end sequencing strategy to map the complex landscape of transcription initiation. *Nature Methods* **7**: 521–527.
- Pervouchine DD, Knowles DG, Guigó R. 2012.** Intron-centric estimation of alternative splicing from RNA-seq data. *Bioinformatics* **29**: 273–274.
- Salimullah M, Mizuho S, Plessy C, Carninci P. 2011.** NanoCAGE: A high-resolution technique to discover and interrogate cell transcriptomes. *Cold Spring Harbor Protocols* **2011**: 5559.
- Schlüter J-P, Reinkensmeier J, Barnett MJ, Lang C, Krol E, Giegerich R, Long SR, Becker A. 2013.** Global mapping of transcription start sites and promoter motifs in the symbiotic α -proteobacterium *Sinorhizobium meliloti* 1021. *BMC Genomics* **14**: 156.
- Shahmuradov IA, Solovyev V V, Gammerman AJ. 2005.** Plant promoter prediction with confidence estimation. *Nucleic Acids Research* **33**: 1069–1076.
- Shandilya J, Roberts SGE. 2012.** The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochimica et Biophysica Acta* **1819**: 391–400.
- Soderlund C, Descour A, Kudrna D, Bomhoff M, Boyd L, Currie J, Angelova A, Collura K, Wissotski M, Ashley E, et al. 2009.** Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. *PLoS Genetics* **5**: e1000740.
- Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006.** Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology* **7 Suppl 1**: S10.1–12.

Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, et al. 2002. The generic genome browser: a building block for a model organism system database. *Genome Research* **12**: 1599–1610.

Sterck L, Billiau K, Abeel T, Rouzé P, Van de Peer Y. 2012. ORCAE: online resource for community annotation of eukaryotes. *Nature Methods* **9**: 1041.

Tanaka T, Koyanagi KO, Itoh T. 2009. Highly diversified molecular evolution of downstream transcription start sites in rice and Arabidopsis. *Plant Physiology* **149**: 1316–1324.

Thorvaldssdóttir H, Robinson JT, Mesirov JP. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in Bioinformatics* **14**: 178–192.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**: 1105–1111.

Valen E, Pascarella G, Chalk A, Maeda N, Kojima M, Kawazu C, Murata M, Nishiyori H, Lazarevic D, Motti D, et al. 2009. Genome-wide detection and analysis of hippocampus core promoters using DeepCAGE. *Genome Research* **19**: 255–265.

Washington NL, Stinson EO, Perry MD, Ruzanov P, Contrino S, Smith R, Zha Z, Lyne R, Carr A, Lloyd P, et al. 2011. The modENCODE Data Coordination Center: lessons in harvesting comprehensive experimental details. *Database* **19**: bar023.

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. 2009. Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis. *The Plant Journal* **60**: 350–362.

Zhang G, Guo G, Hu X, Zhang Y, Li Q. 2010. Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Research* **20**: 646–654.

Zhao X, Valen E, Parker BJ, Sandelin A. 2011. Systematic clustering of transcription start site landscapes. *PloS One* **6**: e23409.

Zuo Y-C, Li Q-Z. 2011. Identification of TATA and TATA-less promoters in plant genomes by integrating diversity measure, GC-Skew and DNA geometric flexibility. *Genomics* **97**: 112–120.

2.8 TABLES AND FIGURES

Table 2.1. Two-way comparison of 5' UTR lengths between *FGH*, *PASA* and *NGS* sources. Absolute frequency is highlighted as a heatmap (0–white, 16,641–red).

	Comparison	Absolute frequency	Percentage
FGH vs PASA	FGH > PASA	2,550	11.86
	PASA > FGH	675	3.14
	FGH = PASA	4,081	18.99
	FGH only	13,454	62.59
	PASA only	734	3.41
	Total	21,494	
PASA vs NGS	PASA > NGS	844	3.42
	NGS > PASA	5,463	22.13
	PASA = NGS	35	0.14
	PASA only	1,698	6.88
	NGS only	16,641	67.42
	Total	24,681	
FGH vs NGS	FGH > NGS	2,051	7.28
	NGS > FGH	13,431	47.67
	FGH = NGS	87	0.31
	FGH only	5,191	18.42
	NGS only	7,414	26.32
	Total	28,174	

Table 2.2. Number of NGS annotations omitted by exclusionary criteria.

Absolute frequency is highlighted as a heatmap (0–white, 8,465–red).

Exclusionary Criteria	Number of genes excluded
Overlapping gene model	2,556
Insufficient depth of coverage	8,465
Coverage into adjacent gene model	1,019
Splice event with adjacent gene model	680
Leaderless*	673
Total	13,393

*Leaderless refers to loci which, although passing all filtering criteria, had no mRNA-seq reads, thus expression evidence, upstream of *INIT*.

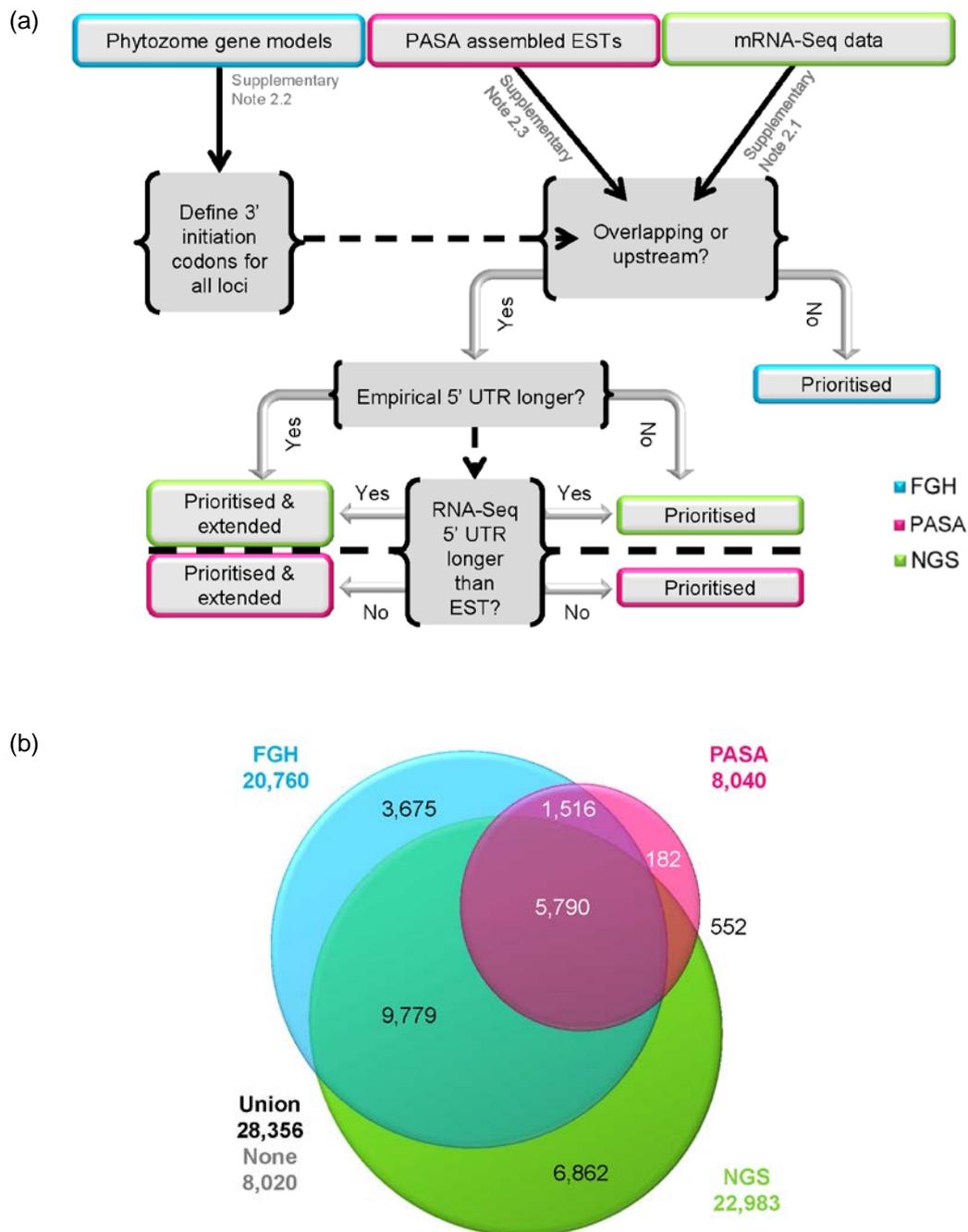


Figure 2.1. Use of data sources for the curation of *E. grandis* 5' UTRs (a) Flow chart overview of methods to determine the prioritisation and extension of *E. grandis* 5' UTR models. (b) Venn diagram showing the number of loci annotated by FGH, PASA and NGS, with the intersection of domains representing loci that were annotated by more than one source.

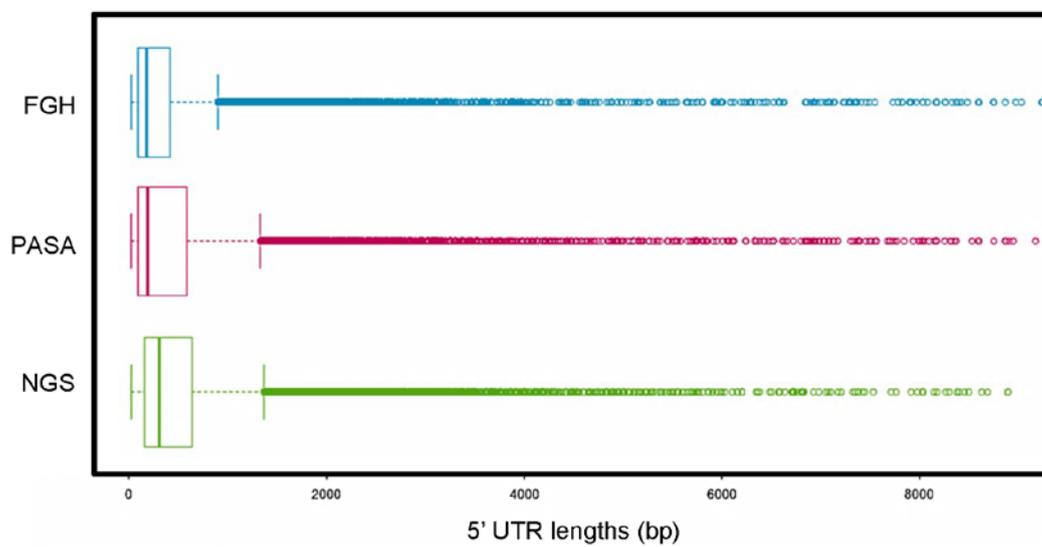


Figure 2.2. Length distributions of *FGH*, *PASA* and *NGS* determined 5' UTRs. The median divides the second and third quartile, while vertical lines show the 5th and 95th percentile, and dots show outliers. The second and third quartiles indicate that *FGH* annotations are shortest. *PASA* third and fourth quartile ranges are nearer that of *NGS* annotations, although the median value remains closer to that of *FGH*.

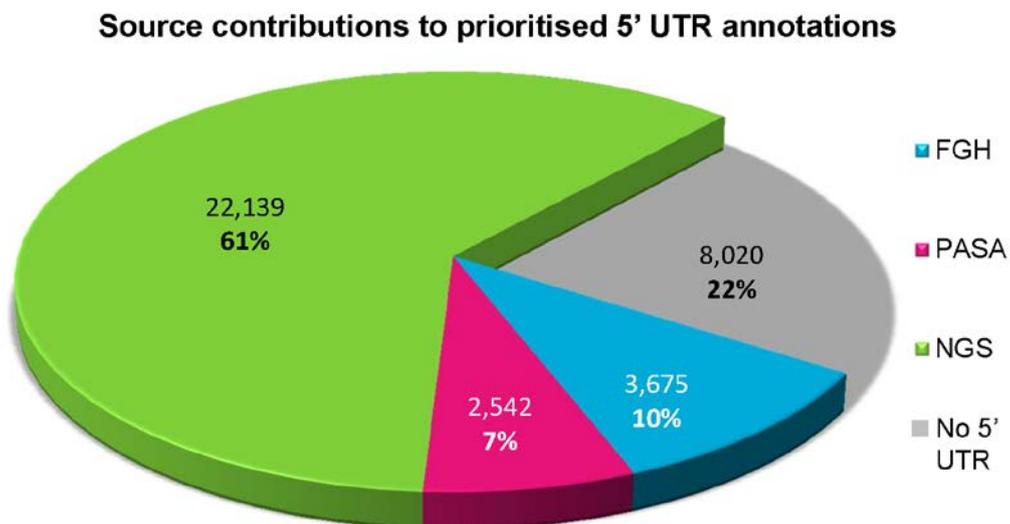


Figure 2.3. Final contributions of each source to the prioritised collection of *E. grandis* 5' UTR annotations. If a gene locus has an *NGS* or *FGH* 5' UTR annotation, the longer of the two is preferred. If there is only one empirical source, that is preferred above an *FGH* 5' UTR annotation, otherwise the *FGH* annotation is retained.

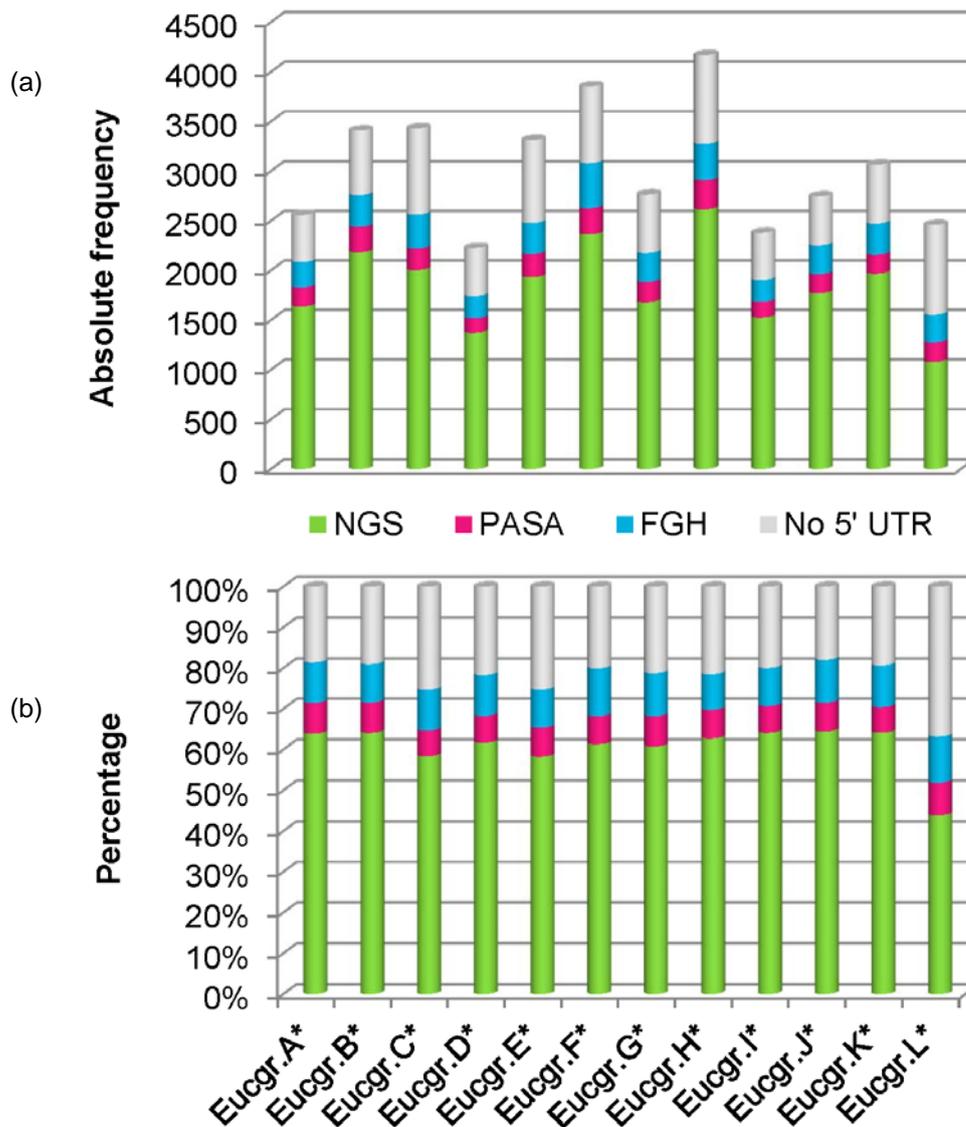


Figure 2.4. Percentage and absolute frequency of prioritised 5' UTR sources per scaffold. (a) the number of loci possessing prioritised FGH, PASA and NGS prioritised annotations per scaffold, and (b) the relative percentage of the prioritised contributions. The major scaffolds 1 to 11 are represented by the IDs Eucgr.A* to Eucgr.K* (* is a wildcard), where minor scaffolds 12 to 5379 are represented by IDs Eucgr.L*. Notice that scaffolds 1 to 11 NGS percentage fluctuates near 60%. The drop to 44% for the remaining scaffolds, Eucgr.L* is coupled with an increase in the percentage of loci with no 5' UTR annotation. This is an intuitively apt result considering the nature of the minor scaffolds, particularly their inability to assemble and the associated doubt coupled with those gene annotations.

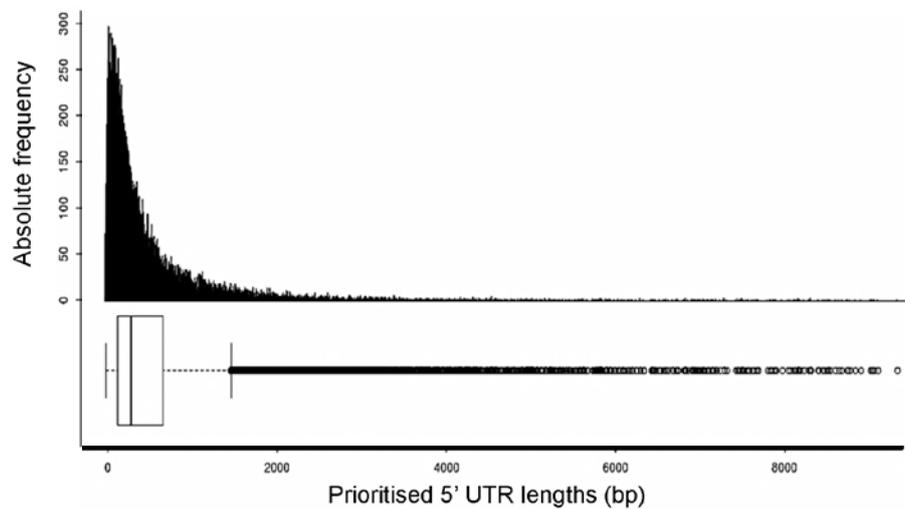


Figure 2.5. Distribution of prioritised *E. grandis* 5' UTR lengths. Absolute frequency is shown above the distribution for prioritised 5' UTR lengths. The box represents the median divided by the second and third quartile, while vertical lines show the 5th and 95th percentile, and dots show outliers.

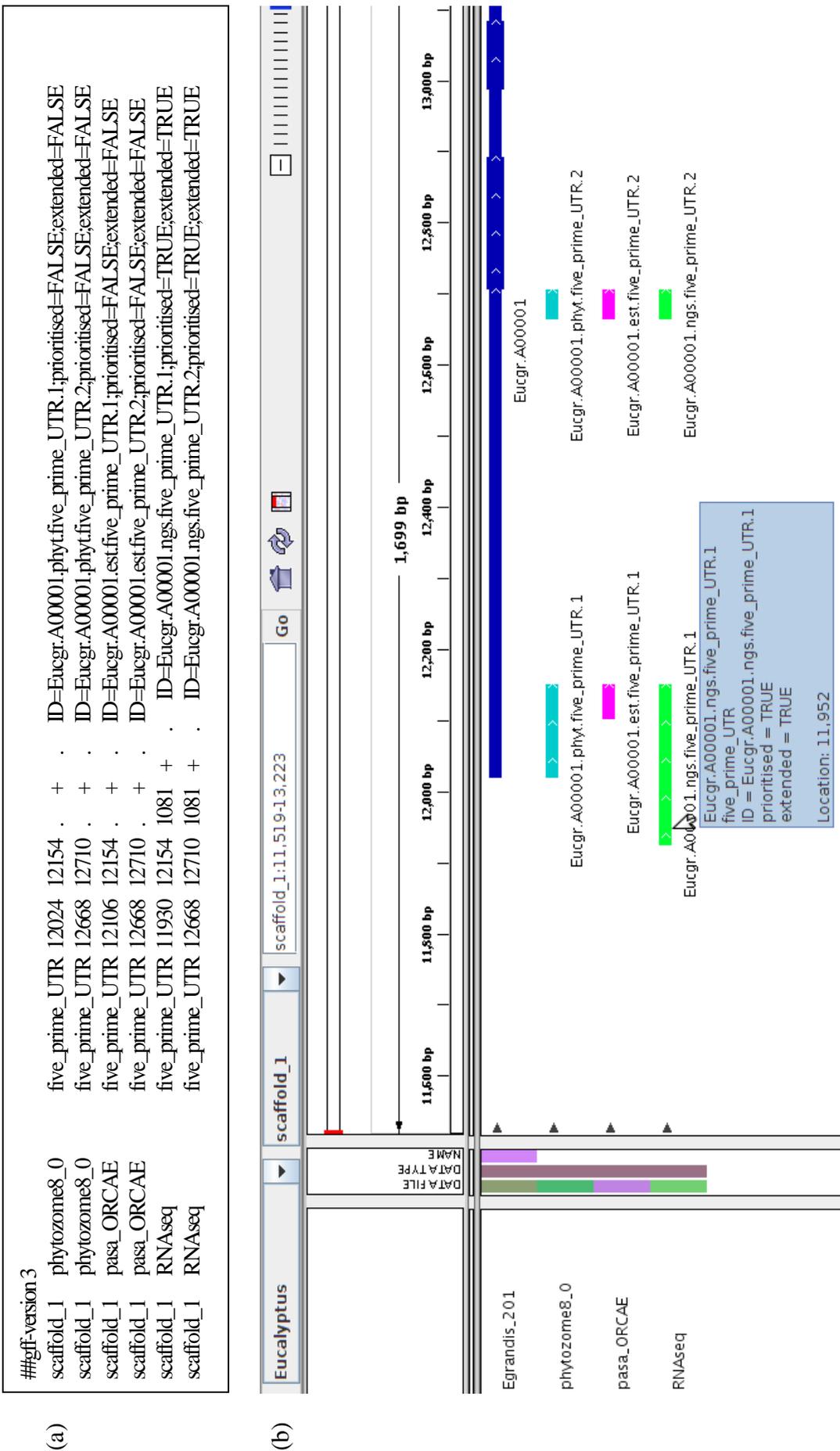


Figure 2.6. Representation of *E. grandis* 5' UTR gff file. (a) Gff3 record of Eucgr.A0001 and its corresponding IGV (Thorvaldsdóttir *et al.*, 2012) representation (b). Notice that the depth of coverage at *INIT* is recorded in the score column of *NGS* entries. This particular example shows that all three sources report the same splice junction. FgenesH (Solovyev *et al.*, 2006) has predicted the TSS upstream of any empirical evidence. mRNA-seq evidence suggests that the TSSD initiates further upstream than what the FGenesh TSS suggests. Thus, the mRNA-seq derived annotation is both extended and prioritized.

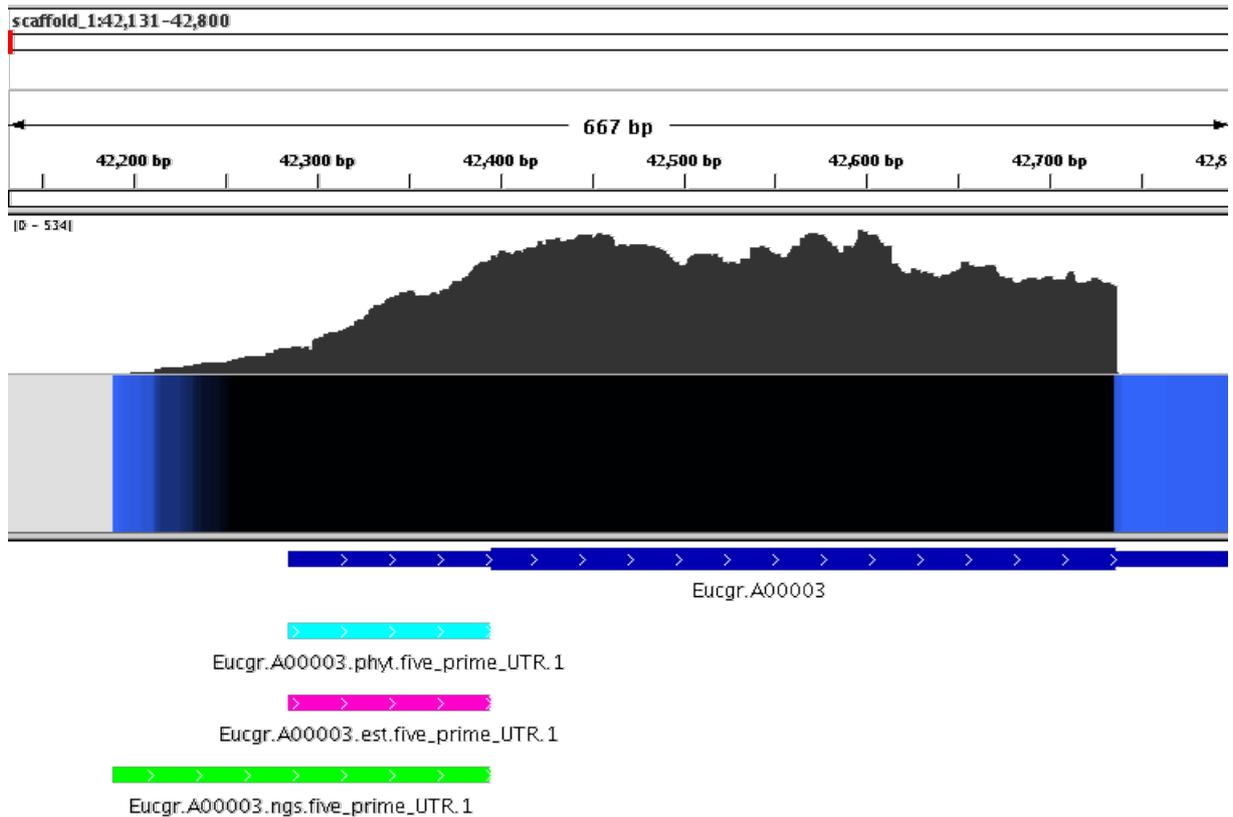


Figure 2.7. NGS data continues upstream of PASA and FGH annotations. Eucgr.A0003 (uncharacterized conserved protein (*DUF2358*)), shows 5' UTR annotation is congruent with the EST evidence. However, NGS evidence continues upstream from this and is considered both the extended and prioritised annotation. Tracks, from top to bottom, are as follows: 1) locus positioning, 2) bigwig mRNA-seq coverage displayed as histogram, 3) bigwig mRNA-seq coverage displayed as heat-map, 4) gene models (Egrandis_201_gene.gff3) from Phytozome, 5) separated (Supplementary Note 2.4) *FGH*, 6) *PASA* and 7) *NGS* 5' UTR annotations respectively.

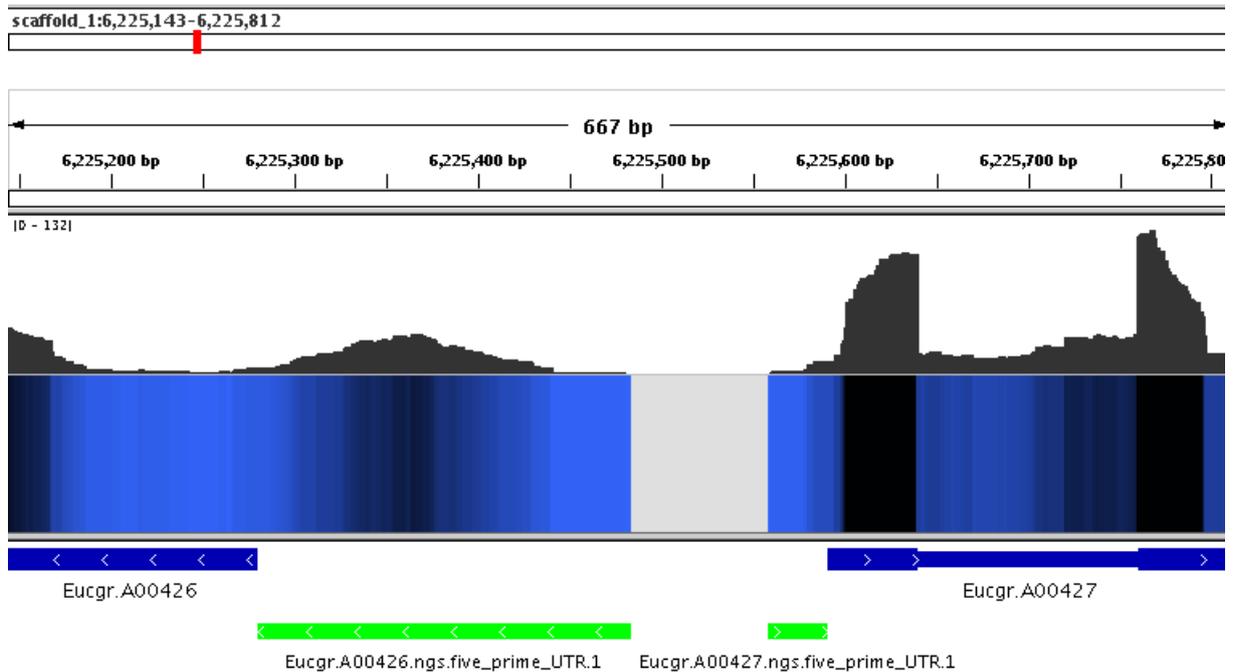


Figure 2.8. Close proximity head-to-head *E. grandis* genes are annotated with NGS 5' UTRs. Eucgr.A00426 (chloroplast outer envelope protein 37) and Eucgr.A00427 (uridine-ribohydrolase 2) contain NGS derived 5' UTR annotations only, successfully annotated despite the two genes being transcribed from opposite strands in close proximity (< 100 bp). Tracks, from top to bottom, are as follows: 1) locus positioning, 2) bigwig mRNA-seq coverage displayed as histogram, 3) bigwig mRNA-seq coverage displayed as heat-map, 4) gene models (Egrandis_201_gene.gff3) from Phytozome, 5) 5' UTR annotations respectively.

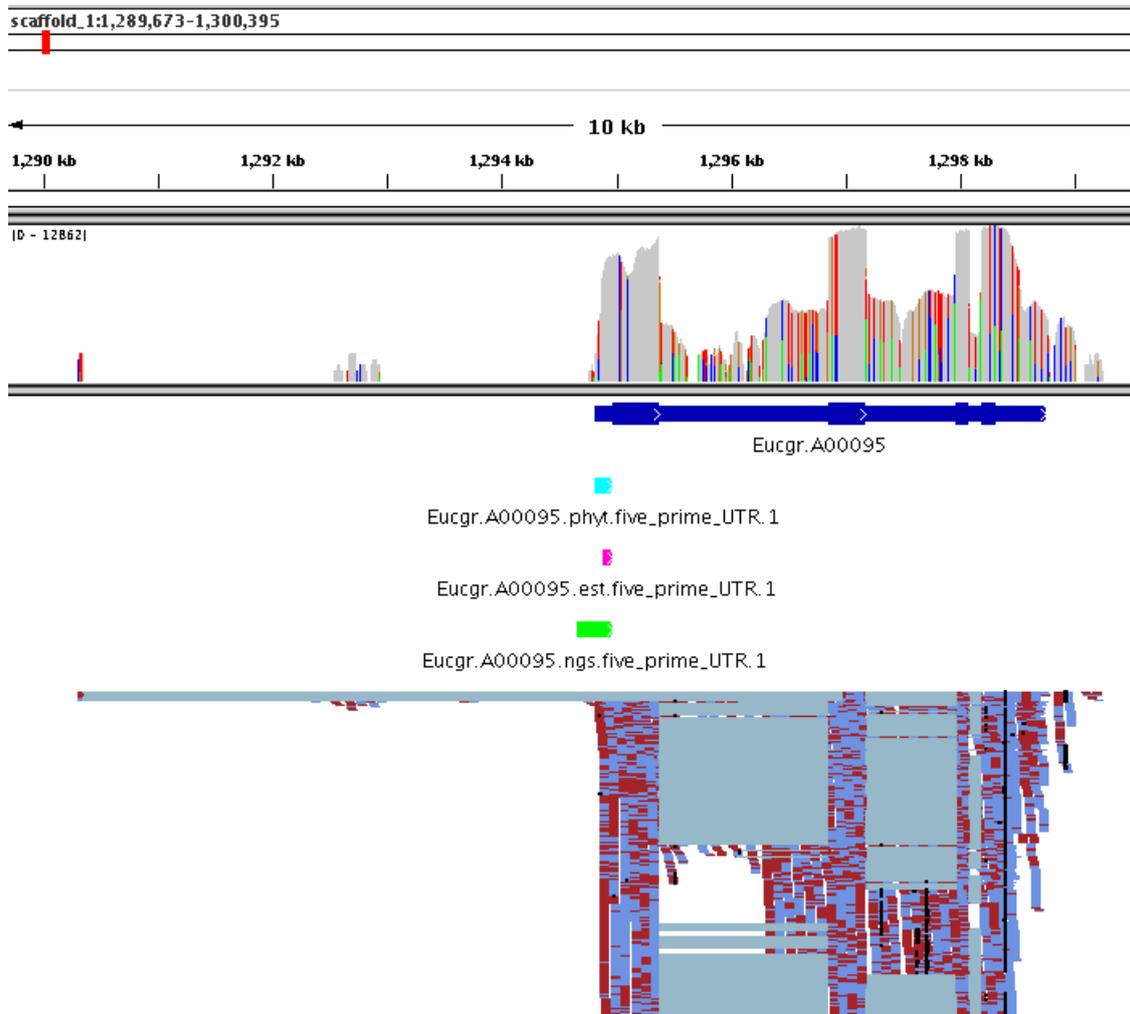
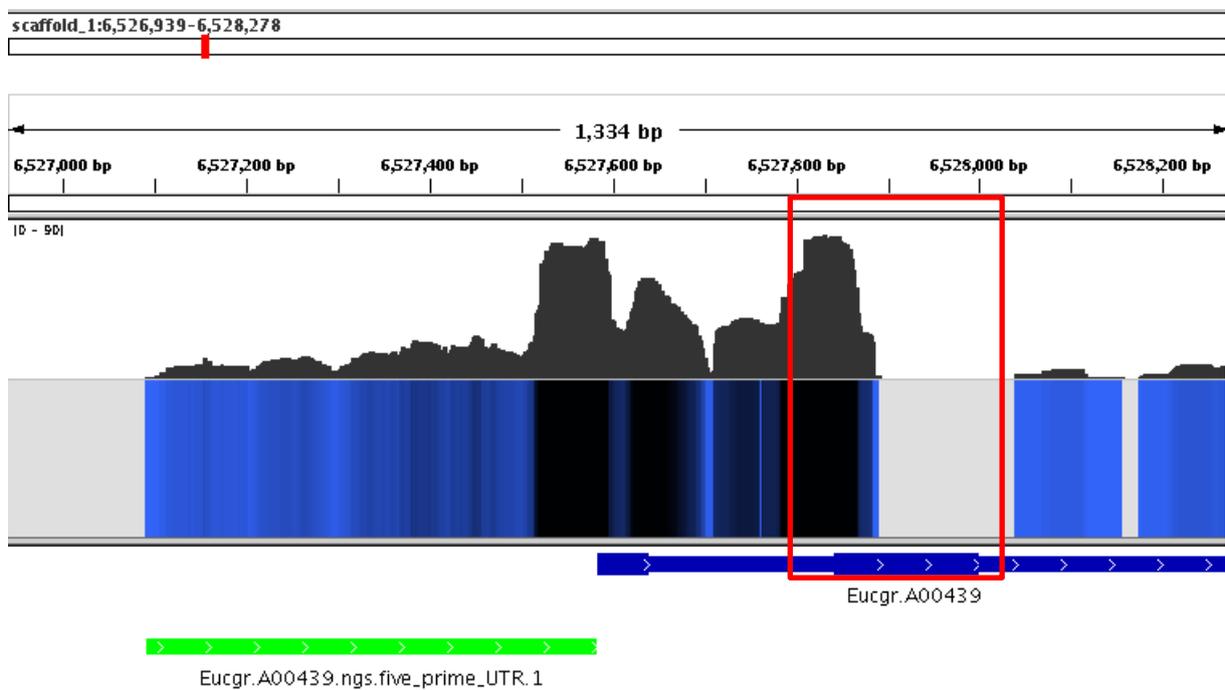
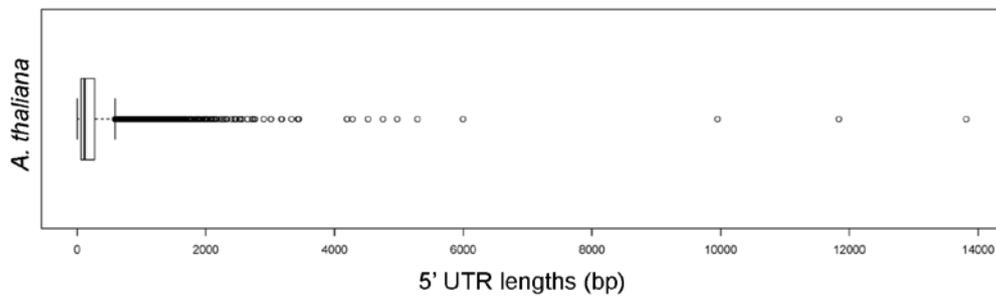


Figure 2.9. Filtering criteria remove spurious upstream splicing events from NGS 5' UTR annotations. Eucgr.A00095 (pfkB-like carbohydrate kinase family protein) aligned mRNA-seq data reports several splice events which extend several kb upstream from the last unspliced alignment. Using filtering criteria, these spliced reads were removed, thus terminating the 5' UTR array, as indicated by the green bar.

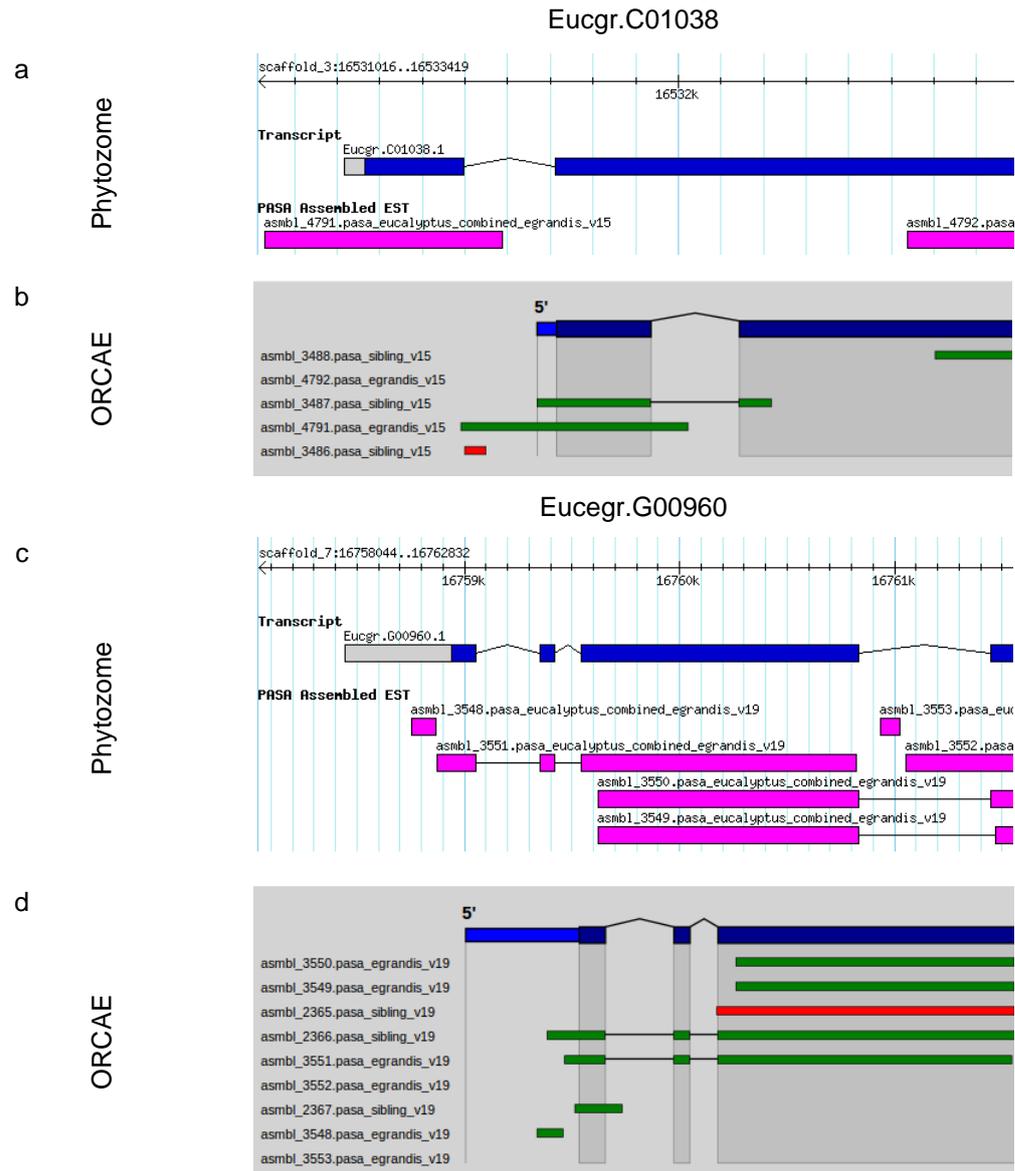
2.9 SUPPLEMENTARY MATERIAL



Supplementary Figure 2.1. Use of NGS data to curate gene annotations. NGS data can provide insight for omitted 5' UTRs and erroneous exon structure. Eucgr.A00439 (hypothetical protein) has no *FGH* or *PASA* annotated 5' UTR and there is discordance between the transcript evidence and the annotated exon structure.



Supplementary Figure 2.2 Distribution of *A. thaliana* 5' UTR lengths. 5' UTR lengths extracted from gff3 file, downloaded from Phytosome v9 (<ftp://ftp.jgi-psf.org/pub/comp/gen/phytozome/v9.0/Athaliana/>, Goodstein *et al.*, 2012; Lamesch *et al.*, 2012)



Supplementary Figure 2.3. Illustrations of predicted and empirical 5' UTR discordance. (a and b) Eucgr.C01038 (polyol/monosaccharide transporter 5) illustrates the truncation of the 5' UTR model, despite empirical evidence suggesting that transcription initiates upstream of the predicted TSS. (c and d) Eucgr.G00960 (phototropic-responsive NPH3 family protein) illustrates the extension of the 5' UTR model, with an absence of empirical evidence substantiating the predicted TSS. Notice that for Phytozome PASA-assembled ESTs (a and b), *E. grandis* and sister eucalypt ESTs are assembled together, whereas ORCAE (c and d) treats these as disparate inputs. For ORCAE aligned ESTs, those represented by green bars “agree” with the gene model, whereas those in red “disagree”. Both disagreeable alignments in the above examples indicate EST alignment to the reverse strand.

Supplementary Note 2.1. Pseudocode to delimit 5' UTRs from aligned mRNA-seq data. Open-source software is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R.

Description: 21 BAM files (mRNA-seq): 3x biological replicates for immature xylem, shoot tip, phloem, mature leaf, young leaf, flower and root tissues; aligned to *E. grandis* version 1 genome assembly (Hefer *et al.* in preparation; default parameters) using TopHat (Trapnell *et al.*, 2009)

Input: BAM files (174.7G)

Procedure:

```

1> FOR each BAM file
2>   Filter for mapping quality of 21
3>   IF PASS
4>   convert to SAM format1
5> Merge all SAM files into bulk_SAM
6> FOR each SAM entry
7>   Extract CIGAR string from SAM format
8>   Filter for intron size > 6500
9>   IF PASS
10>       Convert to bam; sort & index2
11>       Convert to wiggle3
12>       Define all possible splice junctions
13>       Quantify splice junctions4
14>       Filter splice junctions for PSI >= 0.4 and depth_of_coverage > 5
15>       Merge wiggles
16> FOR each Locus_ID IN Egrandis_loci
17>   Filter for depth_of_coverage >= 5 at INIT (derived in Supplementary Note 2.2)
18>   IF PASS
19>       Calculate normalisation_constant
20>       Initialise 5' UTR array
21>       FOR each position upstream of INIT
22>         IF depth_of_coverage > 0 and PASS FILTER_CRITERIA*
23>           Normalise and append to array
24>         ELSE IF depth_of_coverage = 0
25>           IF splice junction supports region and PASS FILTER_CRITERIA*
26>             Normalise and append to array
27>           ELSE IF no splice junction supports position
28>             Terminate array
29> FOR array IN Locus_ID_arrays
30>   Extract 5' UTR start, finish, splice junction coordinates and INIT depth_of_coverage
31>   Modify nomenclature and assign iterative IDs

```

Output: gff3 file (3.1M)

Validation: gff3-validator5
IGV6 inspection

***FILTER_CRITERIA:**

- Gene models do not overlap
- Continuous upstream coverage < 9000bp
- Continuous upstream coverage does not run into an adjacent gene model
- A splicing event does not span to an adjacent gene model
- A splicing event has passed filters as per LINE 14 and has $\{kc_{s_d}, kc_{s_a}\} > k \cdot \log_3(S)$ where $S = \max_{1 \leq y \leq s_d} \{c_y\}$

Open-source tools:

1. Samtools (Li *et al.*, 2009) samtools view -q 2
2. Samtools (Li *et al.*, 2009) samtools view -bSh | samtools sort | samtools index
3. Igvtools (Thorvaldsdóttir *et al.*, 2012) ./igvtools count
4. Bam2ssj (Pervouchine *et al.*, 2012) ./bam2ssj -cps "cps_generated_from_sam_in_LINE10" -maxlen 6500 -minlen
5. ModENCODE gff_validator (Washington *et al.*, 2011)
6. Integrative Genomics Viewer (Thorvaldsdóttir *et al.*, 2012)

Supplementary Note 2.2. Pseudocode to extract and modify 5' UTR annotations from Phytozome gene model annotations. Open-source software, when used, is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R; or the combination thereof.

Description: To extract and delimit the longest possible 5' UTR annotation per locus from the *E. grandis* gene-exon gff3 file from Phytozome.

Input: gff3 file (44M)

Procedure:

```

1> FOR each locus_PAC_ID IN Egrandis_loci
2>   Map locus_PAC_ID to locus_gene_ID
3> FOR each locus_gene_ID
4>   Define INIT as coordinate of most 3' initiation codon
5>   IF strand = '+'
6>     IF exon coordinates < INIT
7>       Delimit 5' UTR by those exons; to and not including INIT
8>   ELSE IF strand = '-'
9>     IF exon coordinates > INIT
10>      Delimit 5' UTR by those exons; to and not including INIT
11>   Modify nomenclature & assign iterative IDs

```

Output: gff3 file (3.7M)

Validation: gff3-validator1
IGV inspection2

Open-source tools:

1. ModENCODE gff-validator (Washington *et al.*, 2011)
2. Integrative Genomics Viewer (Thorvaldsdóttir *et al.*, 2012)

Supplementary Note 2.3. Pseudocode to modify EST tabular data (Sterck *et al.*, 2012) **and extract 5' UTRs.** Open-source software, when used, is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R; or the combination thereof.

Description:	To convert tabular <i>E. grandis</i> EST data from ORCAE into gff3 and delimit the 5' UTR annotations
Input:	tab file (21M)
Procedure:	<pre> 1> FOR each locus_ID IN Egrandis_loci 2> If EST_locus_ID = locus_ID 3> Filter alignments by locus_ID strand 4> IF PASS 5> Merge and sort EST coordinates 6> Delimit 5'UTR by coordinates to and not including INIT 7> Modify nomenclature & assign iterative IDs </pre>
Output:	gff3 file (1.8M)
Validation:	gff3-validator1 IGV inspection2
Open-source tools:	<ol style="list-style-type: none"> 1. ModENCODE gff-validator (Washington <i>et al.</i>, 2011) 2. Integrative Genomics Viewer (Thorvaldsdóttir <i>et al.</i>, 2012)

Supplementary Note 2.4. Script with example Linux commands to separate the merged gff3 file. Separate gff3 file into (a) its three source constituents, (b) those that have been prioritised, and (b) those that have been extended. A valid gff3 file requires the “#gff version3” header, which is ensured by using exclusionary commands.

- | | |
|-----|---|
| (a) | <pre>grep -ve "pasa" Egrandis_five_prime_UTRs.gff3 grep -ve "ngs" > FGH_Egr_five_prime_UTRs.gff3 grep -ve "phyt" Egrandis_five_prime_UTRs.gff3 grep -ve "ngs" > PASA_Egr_five_prime_UTRs.gff3 grep -ve "phyt" Egrandis_five_prime_UTRs.gff3 grep -ve "pasa" > NGS_Egr_five_prime_UTRs.gff3</pre> |
| (b) | <pre>grep -ve "prioritised=F" Egrandis_five_prime_UTRs.gff3 > prioritised_Egr_five_prime_UTRs.gff3</pre> |
| (c) | <pre>grep -ve "extended=F" Egrandis_five_prime_UTRs.gff3 > extended_Egr_five_prime_UTRs.gff3</pre> |

2.10 ADDITIONAL FILES

Additional file 2.1

2_1_Splice_junction_doner_and_acceptor_counts.xlsx

Output of bam2sjj as excel file, indicating splice junctions in 5' UTRs.

Additional file 2.2

2_2_Egrandis_five_prime_UTRs.gff3

Gff3 file of 5' UTRs annotated by FGH, PASA and NGS, with prioritised and extended tags.

Additional file 2.3

2_3_NGS_five_prime_utr_read_density.xlsx

Excel file of NGS 5' UTR lengths, and the density proportions of tapering annotations

CHAPTER 3 : Five core promoter classes drive transcription in *Eucalyptus grandis*

Ida C van Jaarsveld^{1,2}, Eshchar Mizrachi², Fourie Joubert¹ & Alexander A Myburg²

1 Bioinformatics and Computational Biology Unit, Department of Biochemistry, Genomics Research Institute (GRI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa

2 Department of Genetics, Genomics Research Institute (GRI), Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Private bag X20, Pretoria, 0028, South Africa

This chapter is formatted as a research article for a peer-reviewed journal (New Phytologist) to be submitted for review as a companion paper to “Genome sequence of *Eucalyptus grandis*: A global tree crop for fiber and energy” (Myburg *et al.*, in press). IC van Jaarsveld conceptualized the study, E Mizrachi facilitated the study design, IC van Jaarsveld designed the study and methodology, implemented the design, interpreted the results and wrote the manuscript. AA Myburg, F Joubert and E Mizrachi approved and supervised the study and reviewed the manuscript.

3.1 SUMMARY

- Core promoters are essential drivers of transcription and assemble regulatory inputs from diverse sources, such as transcription factors binding *cis*-regulatory elements and nucleosome occupancy. Metazoan core promoter elements are well described, yet plant promoters lack extensive characterization, most notably of how the putative elements regulate transcription initiation.
- We aimed to identify core promoter classes in *Eucalyptus grandis* and assess their functional significance. We use hexamer over-representation to determine significantly enriched hexamers, specifically for the distal transcription start site. Enriched hexamers were assigned to five core promoter classes based on co-occurrence and structural similarity. Each core promoter class was then assessed for expression level and specificity profiles, and Gene Ontology over-representation, to determine the functional associations of each core promoter class.
- We found five core promoter classes drive expression in *E. grandis*, TA, the canonical TATA-box, CT and GA, which are conserved in plants and possess tracts of dinucleotide repeats, W, containing the commonly enriched AAAAAA and TTTTTT hexamer constituents, and S – a class thus far unique to *Eucalyptus*, comprised of CCCCCC and GGGGGG homopolymer tracts. We found that, generally, S showed high and constitutive expression, while TA showed low and specific expression, and that each core promoter class differed in GO enrichment.
- We have annotated 78% of *E. grandis* core promoters, and expect distinct underlying mechanisms of transcriptional control for each promoter class. These insights are invaluable for the understanding of mechanisms of permissive transcription in plant species.

3.2 INTRODUCTION

Promoters are genomic features which initiate transcription by assembling regulatory inputs to control and modulate gene expression. The core promoter, the [-100,+50] region (Florquin *et al.*, 2005) overlapping and adjacent to the transcription start site (TSS), harbours *cis*-regulatory elements which bind the basal transcriptional machinery. This region integrates regulatory inputs which are typically provided by a complex network of *trans*-acting

transcriptional regulators binding to enhancer or repressor inputs in the proximal [-300,-100] or distal [-1000,-300] promoter. The entire 5' leader sequence of a gene is consequential for transcriptional regulation, including the 5' untranslated region (UTR) and core, proximal and distal promoter regions. This includes the region's nucleosome occupancy (Lee *et al.*, 2004; Jiang & Pugh, 2009; Rosin *et al.*, 2012) and histone modifications (Cianfrocco & Nogales, 2013; Kim & Marioni, 2013; Du *et al.*, 2013), which can either allow, compete with, or occlude the binding of regulatory factors with promoter elements.

Metazoan core promoter types have been described in terms of their nucleotide composition, transcription start site distribution (TSSD), nucleosome occupancy, epigenomic features and functional propensity (Lenhard *et al.*, 2012). By these integrative measures, three predominant promoter types have been described. Type I promoters are TATA-box enriched, have focused TSSDs, disordered nucleosome occupancy and occur in tissue-specific genes. Type II promoters are CpG enriched, have broad TSSDs, a nucleosome depleted region (NDR) upstream of the TSS and occur in constitutively expressed genes. The final class, Type III, has the most variable sequence structure, yet has a TSSD mid-range of Type I and Type II, and is specific to developmentally regulated genes (Engström *et al.*, 2007). While these types have been described in metazoans (Lenhard *et al.*, 2012), plant core promoters still lack comprehensive characterisation.

Core promoters in metazoans have been described extensively to possess sequence elements of varying length and consensus strength, yet stringent positional conservation (for review see Juven-Gershon, Hsu, Theisen, & Kadonaga, 2008 and Kadonaga, 2012). The most taxonomically prolific of these is the TATA-box (TATAWA). In metazoans, other prominent motifs are the BREu (SSRCGCC) and BREd (RTDKKKK) flanking the TATA-Box, DPE (RGWYVT), MTE (CSARCSSAAC) and Inr (YYANWYY, TSS underlined), with the co-occurrence of these elements being associated with broad transcription initiation patterns (Kadonaga, 2012). The only metazoan element observed to be conserved in the plant kingdom to date is the TATA-box (Yamamoto *et al.*, 2007a). Sequence composition at plant TSSs shows slight similarity to the Inr element, but is better annotated by the less stringent YR rule (TSS underlined) (Yamamoto *et al.*, 2007b, 2009).

The first detailed genome-wide characterization of promoter properties in plants was that of Molina and Grotewold (2005), who described 12,479 *Arabidopsis thaliana* core promoters.

The TATA box (TATAWAWA) was represented in 29% of the promoters analysed and clustered at position -32. This indicated that the TATA consensus was similar to that in metazoans, yet no other *cis*-elements were over-represented or significantly similar. Following this, Yamamoto et al. (2007) systematically compared promoters of *Arabidopsis*, rice (*Oryza sativa* spp. *japonica*), human and mouse by quantifying octamers and assessing their positional prevalence. Again, the TATA box was conserved across species. Significantly, this study showed that the pyrimidine rich sequences (Y Patch) over-represented in *Arabidopsis* (Yamamoto *et al.*, 2007b) was conserved in plants yet not in mammals.

Another interesting element in *Arabidopsis* is the GA repeat element (GAGAGA), occurring in 22% of the genes, and correlated with a broad transcription initiation patterns (Yamamoto *et al.*, 2009). The GA and TATA promoters are typically distinct from one another, with a weaker association of Y Patch and initiator occurrence with GA promoters than with TATA promoters. The dispersed initiation patterns driven by the GA element are similar to those from CpG islands in mammals, yet do not undergo such methylation. Genes driven by GA promoters show low expression specificity, such as is observed for housekeeping genes. Interestingly, the GA promoters in *A. thaliana* are associated with broad TSSDs, lack the canonical plant initiator YR (TSS underlined), and show constitutive expression. Bernard et al. (2010) described a TC-element in *Arabidopsis* and *O. sativa*, starkly similar to the Y Patch (TTCTTC as compared to TTCTTCTTC, respectively), yet topologically conserved at [-39,-26]. Again, the occurrence of TATA and TC promoter types was distinct, and TC did not occur with the initiator element. In addition to TC genome-wide and high-throughput promoter characterisation, lower-throughput analyses have identified the gene-level conservation of TC repeat promoter elements across species (Creux *et al.*, 2008).

A significant deficit in plant promoter research is the lack of functional characterisation of core promoter classes. The pervasive question remains: what drives transcription at TATA-less promoters in plants? A hexamer over-representation analysis of *Arabidopsis*, rice and soybean promoters (Maruyama *et al.*, 2012) showed that repeat sequences are over-represented in the promoter region [-1000,-1]. The majority of these repeats (TCTCTC, CTCTCT, AAAAAA, TTTTTT, GAGAGA, AGAGAG) have weak (2H) hydrogen bonds. Weak hydrogen bonds lend to less thermodynamic stability and lower energy requirements for DNA melting (Robb *et al.*, 2013; de Avila e Silva *et al.*, 2014) However in rice and

soybean, there are also a number of repeat sequences which have strong (3H) hydrogen bonds (GCGGCG, GGCGGC, CGCCGC, GGGGGG, CCCCCC, GGGCCC) which are over-represented in the promoter. This observed difference in DNA stability is significant and may indicate mechanisms including DNA structural conformation by which transcription is initiated in the absence of a TATA-box. An alternative hypothesis of core promoter elements exhibiting positional constraint is that these sequence features compromise the stability of the DNA at different degrees and propagate a continuous range of transcriptional control (Bernard *et al.*, 2010). The DNA sequence signatures of the core promoter may therefore result in physical conformations which enable the affinity-driven binding of the basal transcriptional machinery to initiate gene expression, including the sequence-specific positioning of nucleosomes.

GA short tandem repeats in the core promoter direct the structural deviation from the typical B-DNA conformation, to triple-stranded H-DNA (Han & de Lanerolle, 2008). H-DNA creates single stranded DNase 1-hypersensitive sites and might confer an open chromatin structure, thus permissive of regulatory element binding and transcription initiation. Han & de Lanerolle (2008) also show that GA repeats in H-DNA associate with histone acetylation and increased promoter activity. Heidari *et al.* (2012) showed that the length of the GA repeat sequence is highly mutable and drives inter-individual expression variation. In addition to the structural role suggested for GA·CT repeat regions, these sequences have also been shown to bind the transcriptional regulators GAGA-BINDING PROTEIN (GBP) (Sangwan & Brian, 2002; Kooiker *et al.*, 2005; Berger *et al.*, 2011; Simonini & Kater, 2014) indicating the possible duality of regulatory mechanisms by core promoter elements. Creux *et al.* (2008) showed that this repeat sequence is conserved at a gene level across several plant species, *A. thaliana*, *Eucalyptus grandis*, and *Populus trichocarpa*, particularly in secondary cell wall-related cellulose synthase promoters.

The *Eucalyptus grandis* genome release (Phytozome V1.0) provides a pivotal platform to study *cis* genetic mechanisms of transcriptional regulation. The growth, adaptability and wood chemistry traits that make *Eucalyptus grandis* a dominant plantation fibre crop are under strong transcriptional control (Mizrachi *et al.* 2010, 2012). The simultaneous availability of the newly sequenced *E. grandis* genome and quantitative expression evidence invite the first genome-wide analysis of promoter characteristics of a woody perennial, and the functional traits and expression specificity associated with them. Together with deep

RNA sequencing experiments across several tissues and developmental stages (Hefer et al., in preparation) and diverse *Eucalyptus* ESTs, we previously determined the distal transcription start sites (dTSS) for 78% of the *E. grandis* loci (Chapter 2; Myburg et al., in press). In this study we analyse the core promoters delimited by the dTSS annotations and aim to identify and functionally characterise core promoter classes. We hypothesise that discreet *E. grandis* core promoter types would be distinguishable by linear sequence arrangement and positional conservation. We further hypothesise that defined promoter types would have unique enrichment of functional associations and unique potential impact on gene expression level and tissue specificity. We found that *E. grandis* core promoters include those such as the TATA-box, which is conserved across kingdoms, and GA, which is conserved in *Arabidopsis*. Additionally we describe several promoter classes thus far unique to *Eucalyptus*, which may be supported across species in future plant promoters studies.

3.3 MATERIALS AND METHODS

3.3.1 Hexamer over-representation

Distal transcription start sites (dTSS) were inferred for 28,356 *E. grandis* genes from the 5' UTR annotations reported in Chapter 2 (Myburg *et al.*, in press). From these annotations, two promoter sequence data sets, core and distal, were extracted from the *E. grandis* genome (Figure 3.1a). The core promoter set θ contains soft-masked DNA sequence 100 bp upstream to 50 bp downstream of the dTSS [-100,+50] (Florquin *et al.*, 2005), whereas the distal promoter set D comprises 800 to 300 bp upstream of the dTSS [-800,-300]. The proximal promoter [-301,-101] was omitted for maximal discrimination, assurance of data independence and the disambiguation of core promoter and enhancer elements. In order to generate synthetic control data sets, fifth-order Markov models were generated from the i) core promoter set (MM_{θ}) and ii) core + distal promoter sets ($MM_{\theta D}$) using RSAT `oligo-analysis` (Thomas-Chollier *et al.*, 2008). Three replicate control core promoter sets $Q \in \{Q_1, Q_2, Q_3\}$ were then generated from MM_{θ} and another three $R \in \{R_1, R_2, R_3\}$ from $MM_{\theta D}$ (Figure 3.1). All control core promoter sets are equivalent in sequence length and number to the core promoter set θ , ensuring the highest level of hexamer and nucleotide frequency conservation in control promoter sets, whilst disrupting the positional conservation. For the test core promoter set θ and each of the six control core promoter sets; $N \in \{Q_1, Q_2, Q_3, R_1, R_2, R_3\}$, 145 six-bp segments were isolated with a 1 bp sliding window

as seg_1, \dots, seg_{145} . The equivalence of the window and k mer size with a 1 bp sliding window, although more computationally expensive, eliminate bias towards repeat sequences and allow a base-pair resolution of positional enrichment. For each $seg \in \{\theta_{seg_1}, \dots, R_{3_{seg_{145}}}\}$ and the distal promoter set D , the frequency of all possible 4,096 hexamer sequences $H \in \{AAAAAA, \dots, TTTTTT\}$ were calculated. Contingency tables were then derived (Supplementary Table 3.2), from which Fisher's exact tests were computed to test whether a given hexamer was significantly over-represented in the seg compared to D , the distal promoter. This resulted in an odds ratio and associated p -value for each seg , for both θ and control core promoter sets $N \in \{Q_1, Q_2, Q_3, R_1, R_2, R_3\}$. P -values were corrected for multiple comparisons using Bonferroni correction and all resultant q -values were deemed significant at < 0.01 .

To determine if a hexamer was over-represented, a prediction interval was calculated for the control sets, and if the test set exceeded the prediction interval, it was said to be over-represented. For each hexamer, the odds ratio distribution for θ was smoothed using a smoothing spline in R. For the control groups $Q \in \{Q_1, Q_2, Q_3\}$ and $R \in \{R_1, R_2, R_3\}$, the upper limit of 99% prediction interval $\{Q_{99}^{upper}, R_{99}^{upper}\}$ and median $\{\tilde{Q}; \tilde{R}\}$ were determined using a generalised linear model. We applied three different tiers of filtering criteria to identify the three types (*broad*, *spiked* and *low*) of hexamer over-representation, detailed in Supplementary Note 3.1. Briefly, we delimited the enrichment region for i) *broad*, as where the spline exceeds Q_{99}^{upper} and $q < 0.01$; ii) *spiked*, a position where the difference in the odds ratio and Q_{99}^{upper} exceeds $Q_{99}^{upper} - \tilde{Q}$ two-fold and $q < 0.01$; and iii) *low*, the spline exceeds the 90th percentile of Q and $q < 0.01$ (Figure 3.1). If a hexamer showed to be generally enriched across the length of the core promoter ($\tilde{Q} > 2.5$), R_{99}^{upper} was used for *broad* and *spiked*, and the 90th percentile of R for *low*, to ensure that significant hexamers showed a peak within the core promoter region [-100,+50]. For each enriched hexamer, the type, enrichment density and enrichment region are reported in Additional file 3.1. Additional filtering criteria for each type are detailed in Supplementary Note 3.1.

3.3.2 Hexamer clustering

A distance matrix was computed using the occurrence of significantly enriched hexamers per gene. Bootstrapped (n=1000) hierarchical clustering was performed using `pvclust` (Suzuki

& Shimodaira, 2006) and clusters were identified using a threshold of Approximately Unbiased *p*-value of 0.05. Clusters which shared structural properties and topological constraint were combined leaving five broad clusters, and thus core promoter classes, namely TA, CT, GA, W and S. GA and CT, although sharing characteristics, were kept separate to determine the extent to which *E. grandis* core promoters corroborate the Y Patch, GA class and CT elements already established in plant promoter literature.

3.3.3 Identifying gene groups

For each core promoter class, the comprising hexamers were searched, collectively, in all promoters in [-500,+500] using Regulatory Sequence Analysis Tools (RSAT) *dna-pattern* (Thomas-Chollier *et al.*, 2008) (Figure 3.1c, Supplementary Note 3.2). For each group a linear model was derived from the hexamer occurrence frequency in [-500,-100]. This gives an expectation of the collection of hexamers. Those positions in which the hexamer occurrence exceeded the 99.9% prediction interval were designated the enrichment region. All genes with one of the constitutive hexamers in the enrichment region were assigned as a class component. We thus defined gene groups. This verifies the hexamers that were identified in the stringent core region, not further upstream (can be downstream as this is a *distal* TSS) and ensures that the method eliminates the premature termination of a class region.

3.3.3 Expression analysis

Expression values were derived from *E. grandis* RNA-seq data, available at Eucgenie (<http://www.eucgenie.org/>, Hefer *et al.*, in preparation) describing expression in several diverse tissues collected from healthy, 6-year old *E. grandis* trees in 3X replicate (three independent trees). The mean FPKM values across replicates were recorded for shoot tip (ST), young leaf (YL), mature leaf (ML), flower (FL), root (RT), phloem (PH) and immature xylem (IX). The proportion of each tissue's expression to the total expression (FPKM) across tissues was also recorded per gene. For each core promoter class, the corresponding genes' expression levels and proportions were extracted. To determine core promoter class expression level trends, the maximum expression level distributions (recording the FPKM value of the tissue with highest expression for each gene, i.e. the maximum "expression potential") were compared using the Wilcoxon-Mann-Whitney test in pairwise comparisons,

with a “greater than” alternative hypothesis (Supplementary Note 3.4). For the expression specificity, the proportion distributions were tested as above. We also used Shannon Entropy (Schug *et al.*, 2005) to determine tissue specificity with a maximum value of 2.8 showing constitutive expression, and the minimum 0 showing tissue specificity, or regulated, expression (Supplementary Note 3.4).

3.3.4 Gene Ontology enrichment analysis

GO over-representation analysis was used to establish unique enrichment for functional associations. For each core promoter type, the gene sets were submitted to TopGO (Alexa *et al.*, 2006). For each core promoter type, using Fisher’s exact test, the top 40 over-represented GO values per GO category were computed, returning unadjusted and Bonferroni corrected *p*-values. Significantly over-represented GO terms (`classicFisher.p < 0.01`) were submitted to REVIGO (Supek *et al.*, 2011) for visualisation, using Lin (Lin, 1998) semantic similarity and *Arabidopsis thaliana* as the proxy for GO database size.

3.4 RESULTS

3.4.1 Three distribution types of hexamer enrichment in *E. grandis* core promoters

We used hexamer positional over-representation to identify DNA signatures enriched in the *E. grandis* core promoter region. Three distinct types of core promoter enrichment profiles were observed (Figure 3.1, Additional file 3.2). In the first type, *broad*, highly significant peaks of enrichment were observed both upstream and downstream of the dTSS consisting of heteropolymer or homopolymer tracts of DNA. Hexamers of the second type, *spiked*, displayed spiked peaks at or adjacent to the dTSS and possess a more complex nucleotide composition (e.g. Supplementary Figure 3.1a-d). A third type, *low*, contains hexamers of less significant enrichment and contains, amongst others, pyrimidine rich tracts described as constituents of the Y Patch of Yamamoto *et al.* (2007) (e.g. Supplementary Figure 3.1e-h). We defined core promoter classes by clustering those hexamers in the *broad* type which possess prominent and significant enrichment peaks in the core promoter region (Figure 3.1).

3.4.2 Five core promoter types of *E. grandis* can be defined by simple repeat sequences

Five classes of core promoter signatures were distinguished, including the TATA-box equivalent TA, CT and GA heteropolymer tracts, and W and S homopolymer tracts (Table 3.1). The TA class, present in 12.2% of *E. grandis* core promoters is comprised of both the strong (TATATA, Figure 3.2a) and weak (TATAAA, Figure 3.2b) form of the canonical binding consensus for TBP and is enriched where the TATA-box is positioned in both metazoans and other plant species. TA is enriched over [-55,-14] and is underrepresented at and downstream of the dTSS. Interestingly, and unique in promoter studies thus far, we found that TA has a bimodal distribution, with the major mode approximately 8 bp upstream of the minor mode.

CT and GA, present in 53% and 20% of promoters respectively (Table 3.1), predominantly occur downstream of the dTSS, despite being enriched from as far upstream as -50, and show a far wider distribution than that of TA. Remarkably, despite the overall prevalence of these features within the core promoter, both possess a phasic constraint within their respective frequency peaks. CTCTCT, a constituent of CT, is phasically constrained from [+3,+12] (Figure 3.2c), thus downstream of the dTSS and GAGAGA, a constituent of GA, is phasically constrained in [-14,+10] (Figure 3.2d), overlapping the dTSS. Both classes showed continued enrichment beyond +50, and extending the analysis to include [-500,+500] (Figure 3.1c, Supplementary Note 3.3) eliminated the premature termination of the CT and GA enrichment regions. CT and GA dinucleotide repeat lengths are indicated in Additional file 3.3.

The W class, in 27% of promoters (Table 3.1), deviates from strict homopolymerism, with the constituents AAAAAA, TTTTTT, and six hexamers which have an embedded AAAA homopolymer tract (Additional file 3.4). W hexamers occur predominantly at the dTSS. AAAAAA (Figure 3.2e) enrichment peaks at the dTSS, whereas TTTTTT (Figure 3.2f) displays a bimodal distribution, with the broader minor mode upstream at [-20,-7], and the sharp major mode at +1. The S homopolymer tract, the least abundant at only 4% of promoters, consists of CCCCCC (Figure 3.2g) and GGGGGG (Figure 3.2h) enrichment upstream and adjacent to the dTSS. These triple hydrogen bonded base pairs show an enrichment distribution with distinctly steep slopes.

The number of promoters possessing one of these core signatures is 21,565, representing 76% of the genes annotated with dTSS. The co-occurrence of core promoter signatures may indicate the combinatorial necessity for functionality, or alternative mechanisms of transcription at a single locus, and the possibility of alternative transcription start sites. Similarly, the depletion of co-occurrence may indicate disparate mechanisms of transcriptional activation. Of the 21,565 genes with annotated core signatures, 57% are annotated with a single core promoter class, 34% with two co-occurring classes and only 9% possess three or more signatures (Figure 3.4).

3.4.3 Core promoter classes are associated with different expression profiles

For each core promoter class, we measured and compared the maximum expression level (median of maximum FPKM recorded for individual genes annotated with that class) to discern whether expression level, a proxy for transcription initiation rate, is correlated with core promoter composition. The FPKM distributions are detailed in Table 3.2 and Figure 3.5a. We see that TA has the lowest median expression (212,700), CT, GA and W have moderate expression (medians 370,300, 334,500 and 356,100) respectively, and S has the highest expression with a median of 566,000. The outliers for each of the classes are considerable, with the median to maximum ratio <0.01 for all classes, showing that expression can be enhanced substantially despite the underlying core promoter, most likely driven by enhancer elements. Bonferroni-corrected Wilcoxon-Mann-Whitney tests (Figure 3.5b) show that S and TA core promoters are statistically more likely to drive high and low expression levels respectively.

Expression specificity was measured to discern whether core promoter classes might influence tissue specificity. Tissue expression proportions (percent of total FPKM per tissue) and the derived Shannon Entropy (Figure 3.6, 3.7) reveals that TA has the highest tissue specificity and that S is primarily constitutive. We also see that GA is more likely to confer tissue specificity than CT (Figure 3.7), and that CT and W are predominantly constitutive. The per-tissue expression proportions for each class (Additional file 3.5) suggest that there is no tissue which preferentially utilizes a core promoter class, but rather that specificity is conferred by combinatorial control with other *cis*-regulatory elements (Vandepoele *et al.*, 2009; Priest *et al.*, 2009; Irimia *et al.*, 2013).

3.4.4 Core promoter classes drive functionally distinct gene categories

Over-represented GO terms provide insight into the functional associations of each core promoter class. We found over-represented GO terms for all five core promoter classes (Additional file 3.6). CT had the most significant (`classicFisher.adj_p` < 0.05) GO terms (74), followed by W (52), both with terms from all three main categories. GA had 24 significant terms from only the Biological Process and Molecular Function categories. TA and S had only five and one Molecular Function terms, respectively. We found that 24 of the 36 *E. grandis* genes which carry the GO term “cellulose synthase activity” (GO:0016759) have a CT promoter and is thus over-represented in the CT class (`classicFisher.p` < 0.01, Figure 3.8a). As CT and GA are complementary sequences and display similar expression profiles, we assessed the co-occurrence of significant GO terms in these two classes to determine the functional overlap. We found that 40 GO terms co-occur, while 55 and 23 are unique to CT and GA respectively (Additional file 3.7). We also found that the GO over-representation of genes with the enriched canonical TATATA hexamer are enriched for both “response to stress” (GO:0006950) and “response to stimulus” (GO:0050896) (Figure 3.8b) corroborating current plant promoter literature (Yamamoto *et al.*, 2009, 2011; Zou *et al.*, 2011).

3.4.5 *Spiked* hexamers occur preferentially at the dTSS

The two remaining types, *spiked* and *low*, comprise 55 and 23 over-represented hexamers respectively (Additional file 3.8). Remarkably, all *spiked* hexamer enrichment is within 8 bps of the dTSS (Supplementary Figure 3.1a-d, Additional file 3.9). These results indicate that the dTSS region is conserved at a per-bp resolution within core promoter subclasses and could be critical for delimiting the first base of permissive transcription. Five A-rich *low* hexamers, AAAAAG AAAACC, AAACCC, CCAAAA, GGGAAA, are also enriched overlapping or in close proximity to the dTSS (Supplementary Figure 3.1e) and cluster together with AAAAAA of W (Additional file 3.10). Another *low* hexamer, CTATAA, mimics the distribution for the *broad* TA constituent, TATAAA (Supplementary Figure 3.1f). The remainder of the *low* class is comprised of pyrimidine residues matching those described as constituents in the Y Patch (Yamamoto *et al.*, 2007b), and show gradual enrichment in a 5' to 3' direction (Supplementary Figure 3.1g,h).

3.5 DISCUSSION

We previously identified dTSSs for 28,356 genes in the newly sequenced *E. grandis* genome (Chapter 2, Myburg *et al.*, in press). We found that empirical evidence of transcription regularly exceeded the predicted 5' UTR (82%), asking the question of what drives transcription further upstream than expected. This provided an opportunity to assess the core promoter signatures at these dTSSs, and determine what DNA sequences are associated with, and putatively delimit, the first base of permissive transcription. Previous plant studies suggest that two discrete types, the TATA-box and GA class are conserved in plant promoters, and that only the TATA-box, of the highly conserved metazoan core promoter element repertoire (TATA-box, DPE, BREu, BREd, MTE, Inr), is conserved across kingdoms. In this study, we have identified five putative *E. grandis* core promoter classes (Table 3.1) using hexamer over-representation and co-occurrence clustering. Of *E. grandis* genes annotated with a dTSS, 76% possess one or more core promoter class hexamer constituents in their respective enrichment region. Using available mRNA-seq expression data (Hefer *et al.*, in preparation) across seven diverse tissues of *E. grandis*, we describe trends in expression level and tissue specificity for each core promoter class, and a gradient of constitutive versus specific expression across classes, specifically the constitutive and highly expressed S class, and the specific and lowly expressed TA class. Together with Gene Ontology enrichment, these sources of information provide valuable insight into transcriptional regulation in *Eucalyptus* and the possible underlying mechanisms conserved across species.

This study hinges on the premise of sequence enrichment inferring conservation, and conservation inferring functionality (Yamamoto *et al.*, 2009). A distinct motive for this study was to determine those hexamers which showed an enrichment profile within the strict confines of the core promoter region and to exclude general *cis*-regulatory elements. We thus used a two-step approach in identifying core promoter classes. The first step used a strict and narrow region around the core promoter, and we used only real distal promoter sequence as a control for maximum signal to noise ratio, ensuring independence of each cell in contingency tables, and to eliminate the identification of *cis*-regulatory enhancer elements. For maximum discriminatory power we used both an independent upstream region and multiple synthetic control samples for the assessment of over-representation. The second step ensured that each over-represented hexamer was not enriched upstream of the core promoter, indicating

enhancer properties, and that those elements still enriched at +50 of the core promoter were not terminated prematurely. The strict filtering criteria in this study, favoured specificity over sensitivity, and highly enriched hexamers of low frequency were filtered. These elements may contribute to establishing permissive transcription, or regulate the temporospatial specificity of expression, but were beyond the scope of this study and are better suited to low-throughput analyses.

The TATA-box, conserved across kingdoms (Yamamoto *et al.*, 2007a; Lenhard *et al.*, 2012), is the most highly conserved core promoter element. It is thus not surprising that we found both the canonical TATATA and common variant TATAAA within the *E. grandis* TA class, and serve to corroborate the dTSSs identified in Chapter 2 (Myburg *et al.*, in press). Our analyses suggest that the *E. grandis* TA core promoter class is similarly involved in stress and stimulus response (Figure 3.8b). Although our results indicate that TA genes are lowly expressed (average maximum FPKM = 1,036,000, compared to the overall average maximum FPKM = 1,269,400), these results are from expression data for healthy non-stressed trees, and we thus do not expect stress-induced expression. As “response to stress” (GO:0006950) and “response to stimulus” (GO:0050896) are over-represented, it may be that TATA-box genes are, in general, only highly expressed when eliciting a response. It is noteworthy that all hexamer constituents of TA display a bimodal distribution, which is thus far unique in core promoter studies. Although a possible artefact of dTSS annotation, this is the only bimodally distributed class and thus warrants further study as to whether TBP binds each mode with equal affinity, and the inherent effect on transcription.

The CT promoter class is the most prevalent in *E. grandis* (53%, Table 3.1), and shows similarity to both the Y Patch (Yamamoto *et al.*, 2007b) and TC_[-39,-26]-PLMs (Bernard *et al.*, 2010) described in *A. thaliana* (43% and 19% of promoters tested in each study, respectively) and corroborated in *O. sativa* (Yamamoto *et al.*, 2007a; Bernard *et al.*, 2010) and *Glycine max* (Maruyama *et al.*, 2012). In this study, the *E. grandis* CT class is distinct from the other elements that constitute the Y Patch which were found to be only marginally enriched (type *low*), but without positional constraint in the *E. grandis* core promoter (Additional file 3.8). A CT element has previously been described in the secondary cell wall-related cellulose synthase (*CesA*) genes of *A. thaliana*, *E. grandis* and *P. trichocarpa* (Creux *et al.*, 2008), and the element was associated with seven candidate wood quality candidate genes, including *CesA3*, in *E. globulus* (Acuña *et al.*, 2012). Together with evidence of enrichment of

“cellulose synthase activity” (GO:0016759; `classicFisher.p` = 0.005) of the CT class in this study (Figure 3.8a), we posit that the CT element may be distinct from the Y Patch and enriched in genes associated with cellulose biosynthesis during secondary cell wall formation in *Eucalyptus*.

GA repeat sequences, complementary to CT in the DNA double-helix, have likewise been described to occur in the promoters of plant (Yamamoto *et al.*, 2009) and metazoan species (Heidari *et al.*, 2012). The GA repeat sequences found in 21.6% of *A. thaliana* promoters by Yamamoto *et al.* (2009) occur in the equivalent region of *E. grandis* GA promoters. The *E. grandis* GA promoters show stronger functional associations than those in *Arabidopsis* (Yamamoto *et al.*, 2009) and are predominantly constitutive with moderate expression levels. Metazoan GA promoters are prevalent in developmental genes (Heidari *et al.*, 2012), and are similarly enriched in developmental genes in *E. grandis*, with over-represented GO terms such as “tissue development” (GO:0009888), shoot system development” (GO:0022621) and “developmental process” (GO:0032502), amongst others (Additional file 3.6). Previous analyses have described CT (Yamamoto *et al.*, 2007b; Bernard *et al.*, 2010) and GA (Yamamoto *et al.*, 2009) dinucleotide repeats as separate promoter classes in plants. However, the dinucleotide repeat lengths and their positioning indicate that both classes are capable of forming H-DNA (Han & de Lanerolle, 2008) and binding GAGA-binding factors (Berger *et al.*, 2011), both of which regulate transcription (Berger & Dubreucq, 2012). It is possible that CT and GA possess the same underlying transcriptional mechanism and could, on a functional basis, be combined into one. Although *E. grandis* CT and GA promoter classes share 40 over-represented GO terms and have similar expression level and specificity profiles, we refrained from merging them to study *E. grandis* core promoter corroboration of the Y Patch, GA class and CT elements already established in plant promoter literature.

Core promoter classes yet to be defined in other plant promoters are W and S. These classes are similar in that they are homopolymer tracts occurring in close proximity to the dTSS (Figure 3.9). AAAAAA and TTTTTT are commonly described as enriched in plant promoters (Yamamoto *et al.*, 2009; Azad *et al.*, 2011; Maruyama *et al.*, 2012), but to our knowledge, this is the first reported instance of within-core enrichment. A·T repeat sequence confers DNA rigidity and is enriched in the linker DNA adjacent to nucleosomes (Jiang & Pugh, 2009), whilst being depleted within nucleosomes (Mrázek *et al.*, 2011). It has also been shown that TSSs cluster tightly around the 5' end of the +1 nucleosome in

Saccharomyces cerevisiae (Rhee & Pugh, 2012). The A/T enrichment observed upstream of the dTSS may establish a nucleosome depleted region (NDR) for the binding of the pre-initiation complex, and is delimited by the within-core enrichment of W. The TCT initiator element in *Drosophila melanogaster* exhibits similarly strong nucleosome positioning as the transcriptional mechanism (Kadonaga, 2012). TCTTTT and TTCTTT, two *spiked* hexamers overlapping the dTSS (Supplementary Figure 3.2) and clustering with TTTTTT (Additional file 3.10) support the possibility of an upstream NDR and the precise positioning of the +1 nucleosome as the core transcriptional mechanism in *E. grandis* W promoters. Unlike W, which shows general enrichment in plant promoters, of the plant promoters described thus far, S within-core enrichment is unique to *E. grandis*. Although proteins have been shown to bind C-repeats and have transcriptional activity (Xu & Goodridge, 1999; Ehrenkauffer *et al.*, 2009), it is more likely that the distinct physiochemical properties, that being high disrupt energy and increased DNA stability (Breslauer *et al.*, 1986), delimit the dTSS, although this requires further investigation (Florquin *et al.*, 2005).

We have used hexamer over-representation to identify five *E. grandis* core promoter classes. We found support for the TATA-box and GA plant promoter classes, and evidence supporting the presence, if not the position and distribution, of Y Patch constituents. The low-abundance CA class (Yamamoto *et al.*, 2009) was not detected in this study. The most abundant promoter class in *E. grandis* is CT, previously described as a constituent of the Y Patch (Yamamoto *et al.*, 2007b). We posit that, not only is CT distinct from the Y Patch, but that it may utilise the same underlying mechanisms as the GA class (Figure 3.9). Finally, two novel classes, W and S, may direct structural morphology of DNA at the dTSS and regulate transcription by nucleosome positioning (Lee *et al.*, 2004; Narlikar *et al.*, 2007; Rosin *et al.*, 2012) and DNA stability (Breslauer *et al.*, 1986), respectively. A unique feature of this study is the specific use of dTSSs, thus the most 5' position of empirically discernable transcription, to define core promoters. Rather than determining sequence enrichment for peak TSS selection, which has been shown to be highly stochastic (Kaern *et al.*, 2005), we aimed to determine sequence enrichment associated with the start of permissive transcription. Together with the putative structural mechanisms of regulation suggested in this study, we provide valuable insight for models of transcription initiation where the core promoter acts as a binary on/off switch and *cis*-regulatory elements temporospatially modulate the expression level and specificity to derive the complex phenotypes observed in plant species.

3.6 ACKNOWLEDGMENTS

We would like to acknowledge Anna Kersting for providing the *E. grandis* Gene Ontology Universe and Ontology mapping files, Nannette Coetzer and Johann Swart for writing and maintaining the Galaxy wrapper for TopGO respectively, and Charles Hefer for computing and providing the *E. grandis* gene expression data.

3.7 REFERENCES

- Acuña C, Villalba P, García M, Pathauer P, Hopp E, Marcucci Poltri S. 2012.** Microsatellite markers in candidate genes for wood properties and its application in functional diversity assessment in *Eucalyptus globulus*. *Electronic Journal of Biotechnology* **15**: 2–2.
- Alexa A, Rahnenführer J, Lengauer T. 2006.** Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600–1607.
- De Avila e Silva S, Forte F, T S Sartor I, Andrighetti T, J L Gerhardt G, Longaray Delamare AP, Echeverrigaray S. 2014.** DNA duplex stability as discriminative characteristic for *Escherichia coli* $\sigma(54)$ - and $\sigma(28)$ - dependent promoter sequences. *Biological : Journal of the International Association of Biological Standardization* **42**: 22–8.
- Azad AKM, Shahid S, Noman N, Lee H. 2011.** Prediction of plant promoters based on hexamers and random triplet pair analysis. *Algorithms for Molecular Biology* **6**: 19.
- Berger N, Dubreucq B. 2012.** Evolution goes GAGA: GAGA binding proteins across kingdoms. *Biochimica et Biophysica Acta* **1819**: 863–868.
- Berger N, Dubreucq B, Roudier F, Dubos C, Lepiniec L. 2011.** Transcriptional regulation of *Arabidopsis* LEAFY COTYLEDON2 involves RLE, a *cis*-element that regulates trimethylation of histone H3 at lysine-27. *The Plant Cell* **23**: 4065–4078.
- Bernard V, Brunaud V, Lecharyn A. 2010.** TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics* **11**: 166–180.
- Breslauer KJ, Frank R, Blöcker H, Marky L a. 1986.** Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences of the United States of America* **83**: 3746–3750.
- Cianfrocco MA, Nogales E. 2013.** Regulatory interplay between TFIID's conformational transitions and its modular interaction with core promoter DNA. *Transcription* **4**: 1–7.
- Creux NM, Ranik M, Berger DK, Myburg AA. 2008.** Comparative analysis of orthologous cellulose synthase promoters from *Arabidopsis*, *Populus* and *Eucalyptus*: evidence of conserved regulatory elements in angiosperms. *The New Phytologist* **179**: 722–737.

Du Z, Li H, Wei Q, Zhao X, Wang C, Zhu Q, Yi X, Xu W, Liu XS, Jin W, et al. 2013. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa L. japonica*. *Molecular Plant* **6**: 1463–1472.

Ehrenkauf GM, Hackney J a, Singh U. 2009. A developmentally regulated Myb domain protein regulates expression of a subset of stage-specific genes in *Entamoeba histolytica*. *Cellular Microbiology* **11**: 898–910.

Engström PG, Ho Sui SJ, Drivenes O, Becker TS, Lenhard B. 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Research* **17**: 1898–1908.

Florquin K, Saeys Y, Degroeve S, Rouzé P, Van de Peer Y. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Research* **33**: 4255–4264.

Han Y-J, de Lanerolle P. 2008. Naturally extended CT.AG repeats increase H-DNA structures and promoter activity in the smooth muscle myosin light chain kinase gene. *Molecular and Cellular Biology* **28**: 863–872.

Heidari A, Nariman Saleh Fam Z, Esmaeilzadeh-Gharehdaghi E, Banan M, Hosseinkhani S, Mohammadparast S, Oladnabi M, Ebrahimpour MR, Soosanabadi M, Farokhashtiani T, et al. 2012. Core promoter STRs: novel mechanism for inter-individual variation in gene expression in humans. *Gene* **492**: 195–198.

Irimia M, Maeso I, Roy SW, Fraser HB. 2013. Ancient *cis*-regulatory constraints and the evolution of genome architecture. *Trends in Genetics* **29**: 521–528.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics* **10**: 161–172.

Juven-Gershon T, Hsu J-Y, Theisen JW, Kadonaga JT. 2008. The RNA polymerase II core promoter - the gateway to transcription. *Current Opinion in Cell Biology* **20**: 253–259.

Kadonaga JT. 2012. Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology* **1**: 40–51.

Kaern M, Elston TC, Blake WJ, Collins JJ. 2005. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics* **6**: 451–464.

Kim JK, Marioni JC. 2013. Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology* **14**: R7.

Kooiker M, Airoidi CA, Losa A, Manzotti PS, Finzi L, Kater MM. 2005. BASIC PENTACYSSTEINE1, a GA Binding Protein that induces conformational changes in the regulatory region of the homeotic *Arabidopsis* gene SEEDSTICK. *The Plant Cell* **17**: 722–729.

Lee C-K, Shibata Y, Rao B, Strahl BD, Lieb JD. 2004. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nature Genetics* **36**: 900–905.

- Lenhard B, Sandelin A, Carninci P. 2012.** Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* **13**: 233–245.
- Lin D. 1998.** An information-theoretic definition of similarity. *Proceedings of the 15th International Conference on Machine Learning*: 296–304.
- Maruyama K, Todaka D, Mizoi J, Yoshida T, Kidokoro S, Matsukura S, Takasaki H, Sakurai T, Yamamoto YY, Yoshiwara K, et al. 2012.** Identification of cis-acting promoter elements in cold- and dehydration-induced transcriptional pathways in *Arabidopsis*, rice, and soybean. *DNA Research* **19**: 37–49.
- Molina C, Grotewold E. 2005.** Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* **6**: 1471–2164.
- Mrázek J, Chaudhari T, Basu A. 2011.** PerPlot & PerScan: tools for analysis of DNA curvature-related periodicity in genomic nucleotide sequences. *Microbial Informatics and Experimentation* **1**: 13.
- Narlikar L, Gordân R, Hartemink AJ. 2007.** A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Computational Biology* **3**: e215.
- Priest HD, Filichkin SA, Mockler TC. 2009.** Cis-regulatory elements in plant cell signaling. *Current Opinion in Plant Biology* **12**: 643–649.
- Rhee HS, Pugh BF. 2012.** Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature* **483**: 295–301.
- Robb NC, Cordes T, Hwang LC, Gryte K, Duchi D, Craggs TD, Santoso Y, Weiss S, Ebright RH, Kapanidis AN. 2013.** The transcription bubble of the RNA polymerase-promoter open complex exhibits conformational heterogeneity and millisecond-scale dynamics: implications for transcription start-site selection. *Journal of Molecular Biology* **425**: 875–885.
- Rosin D, Hornung G, Tirosch I, Gispan A, Barkai N. 2012.** Promoter nucleosome organization shapes the evolution of gene expression. *PLoS Genetics* **8**: e1002579.
- Sangwan I, Brian MRO. 2002.** Identification of a soybean protein that interacts with GAGA element dinucleotide repeat DNA. *Plant Physiology* **129**: 1788–1194.
- Schug J, Schuller W-P, Kappen C, Salbaum JM, Bucan M, Stoeckert CJ. 2005.** Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biology* **6**: R33.
- Simonini S, Kater MM. 2014.** Class I BASIC PENTACYSTEINE factors regulate HOMEBOX genes involved in meristem size maintenance. *Journal of Experimental Botany* **65**: 1455–1465.
- Supek F, Bošnjak M, Škunca N, Šmuc T. 2011.** REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One* **6**: e21800.

Suzuki R, Shimodaira H. 2006. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**: 1540–1542.

Thomas-Chollier M, Sand O, Turatsinze J-V, Janky R, Defrance M, Vervisch E, Brohée S, van Helden J. 2008. RSAT: regulatory sequence analysis tools. *Nucleic Acids Research* **36**: W119–27.

Turatsinze J-V, Thomas-Chollier M, Defrance M, van Helden J. 2008. Using RSAT to scan genome sequences for transcription factor binding sites and *cis*-regulatory modules. *Nature Protocols* **3**: 1578–1588.

Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. 2009. Unraveling transcriptional control in *Arabidopsis* using *cis*-regulatory elements and coexpression networks. *Plant Physiology* **150**: 535–546.

Xu G, Goodridge a G. 1999. Function of a C-rich sequence in the polypyrimidine/polypurine tract of the promoter of the chicken malic enzyme gene depends on promoter context. *Archives of Biochemistry and Biophysics* **363**: 202–212.

Yamamoto YY, Ichida H, Abe T, Suzuki Y, Sugano S, Obokata J. 2007a. Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Research* **35**: 6219–6226.

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T. 2007b. Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**: 1471–2164.

Yamamoto YY, Yoshioka Y, Hyakumachi M, Obokata J. 2011. Characteristics of core promoter types with respect to gene structure and expression in *Arabidopsis thaliana*. *DNA Research* **5**: 333.

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. 2009. Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *The Plant Journal* **60**: 350–362.

Zou Y, Huang W, Gu Z, Gu X. 2011. Predominant gain of promoter TATA Box after gene duplication associated with stress responses. *Molecular Biology and Evolution* **28**: 2893–2904.

3.8 TABLES AND FIGURES

Table 3.1. Core promoter classes detected and defined in *E. grandis*. Constituent over-represented hexamers, the number of genes with those over-represented hexamers and the region of over-representation are shown. The occurrence (spatial) distribution (full size figures in Additional file 3.2) shows the shape of promoter class distribution from -500 to +500, with the dTSS position indicated by the dotted white line.

Type	Hexamers	Number of genes	region	Occurrence distribution — Spline (0.7) — 99.9% Pred. Int.
TA	TATATA ATATAT TATAAA	3,463 (12.2%)	[-55,-14]	
CT	CTCTCT TCTCTC	15,007 (52.9%)	[-76,+397]	
GA	GAGAGA AGAGAG	5,697 (20.1%)	[-33,+203]	
W	TTTTTT AAAAAA AAAAAT GAAAAA AGAAAA GAAAAT CAAAAA AAAAAC	7,774 (27.4%)	[-11]; [-2,+3]	
S	CCCCCC GGGGGG	1,042 (3.7%)	[-36,+5]	

Table 3.2. FPKM distribution for each promoter class. FPKM values are from the highest-expressed tissue per gene. Median FPKM values are highlighted as a heatmap (212,700–white, 566,000–red).

	Min	1st Quartile	Median	Mean	3rd Quartile	Max	Median /Max %	No. not expressed
TA	189	45,650	212,700	1,036,000	773,600	132,700,000	0.16	97
CT	99	86,610	370,300	1,253,000	1,052,000	482,500,000	0.08	296
GA	248	69,050	334,500	1,241,000	1,034,000	350,600,000	0.10	145
W	228	79,220	356,100	1,218,000	1,016,000	350,600,000	0.10	262
S	607	183,200	566,000	1,599,000	1,450,000	88,540,000	0.64	8

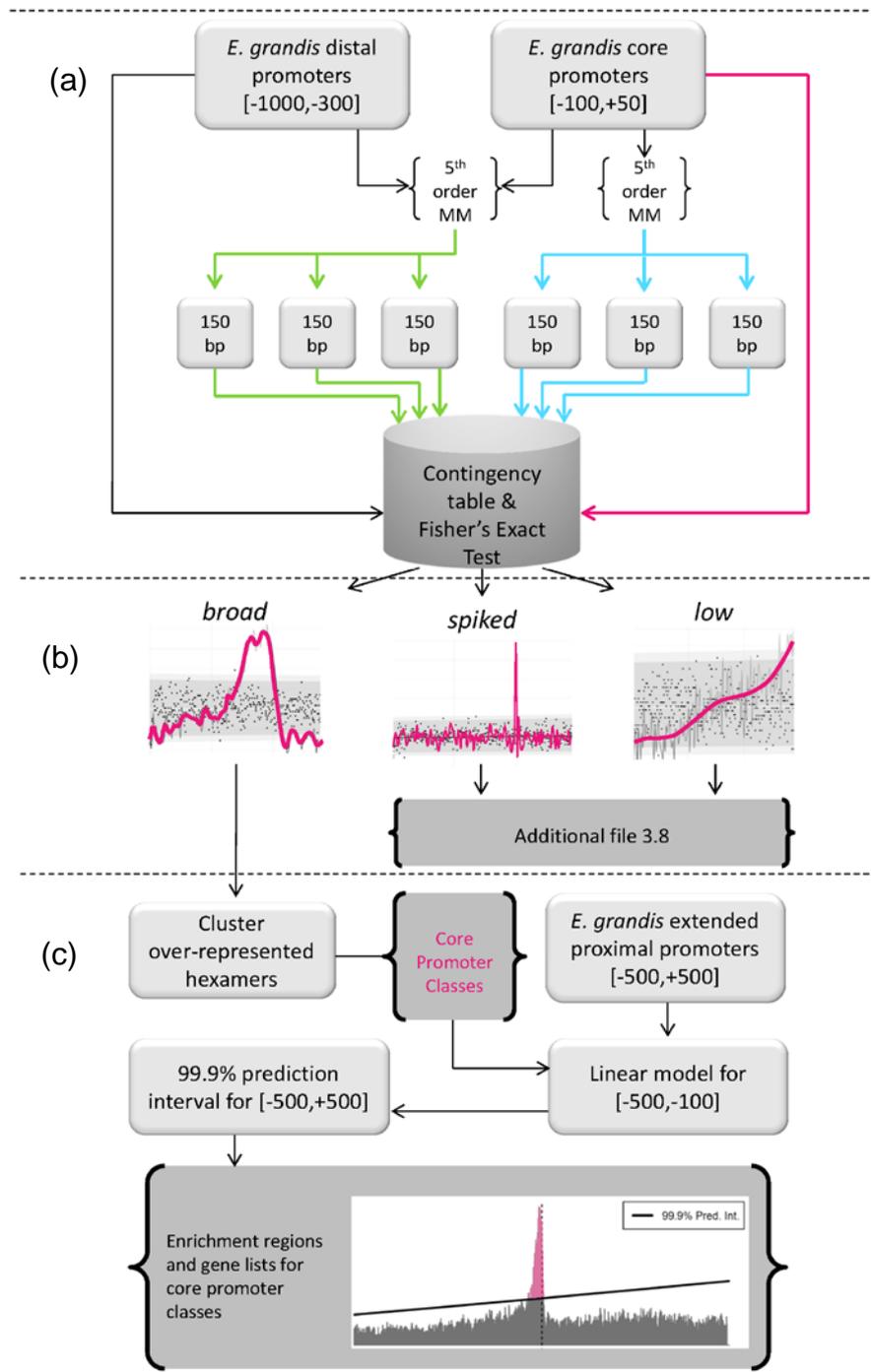


Figure 3.1. Overview of methodology. Procedure to (a) generate control sets, and identify (b) enriched hexamers and (c) core promoter classes.

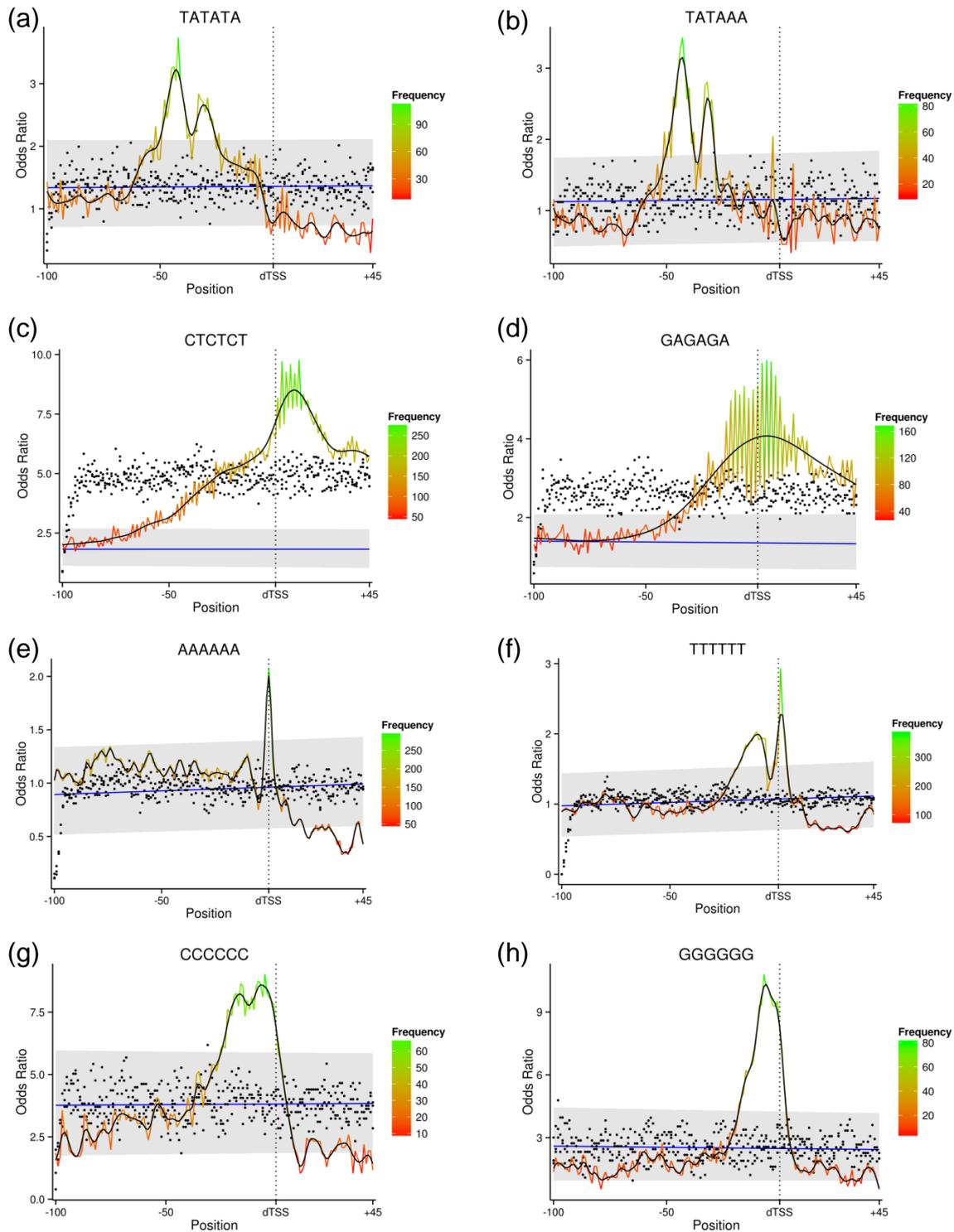


Figure 3.2. Odds Ratio distribution of over-represented hexamers determined by Fisher's Exact Test. (a) TATATA, (b) TATAAA, (c) CTCTCT, (d) GAGAGA, (e) AAAAAA, (f) TTTTTT, (g) CCCCCC, and (e) GGGGGG. The coloured line indicates the Odds Ratio per position, and is coloured, as a third axis, according to the frequency, or raw count, per position. The start of each tested hexamer position is indicated on the x-axis. The black line indicates the smoothed spline of the Odds Ratio values. The dots represent the odds ratios of the core control sets $Q \in \{Q_1, Q_2, Q_3\}$. The blue line and the grey shaded area are the linear model and derived 99% prediction interval; determined by the core $Q \in \{Q_1, Q_2, Q_3\}$ (a,b,e-h) and core+distal $R \in \{R_1, R_2, R_3\}$ (c,d) control sets.

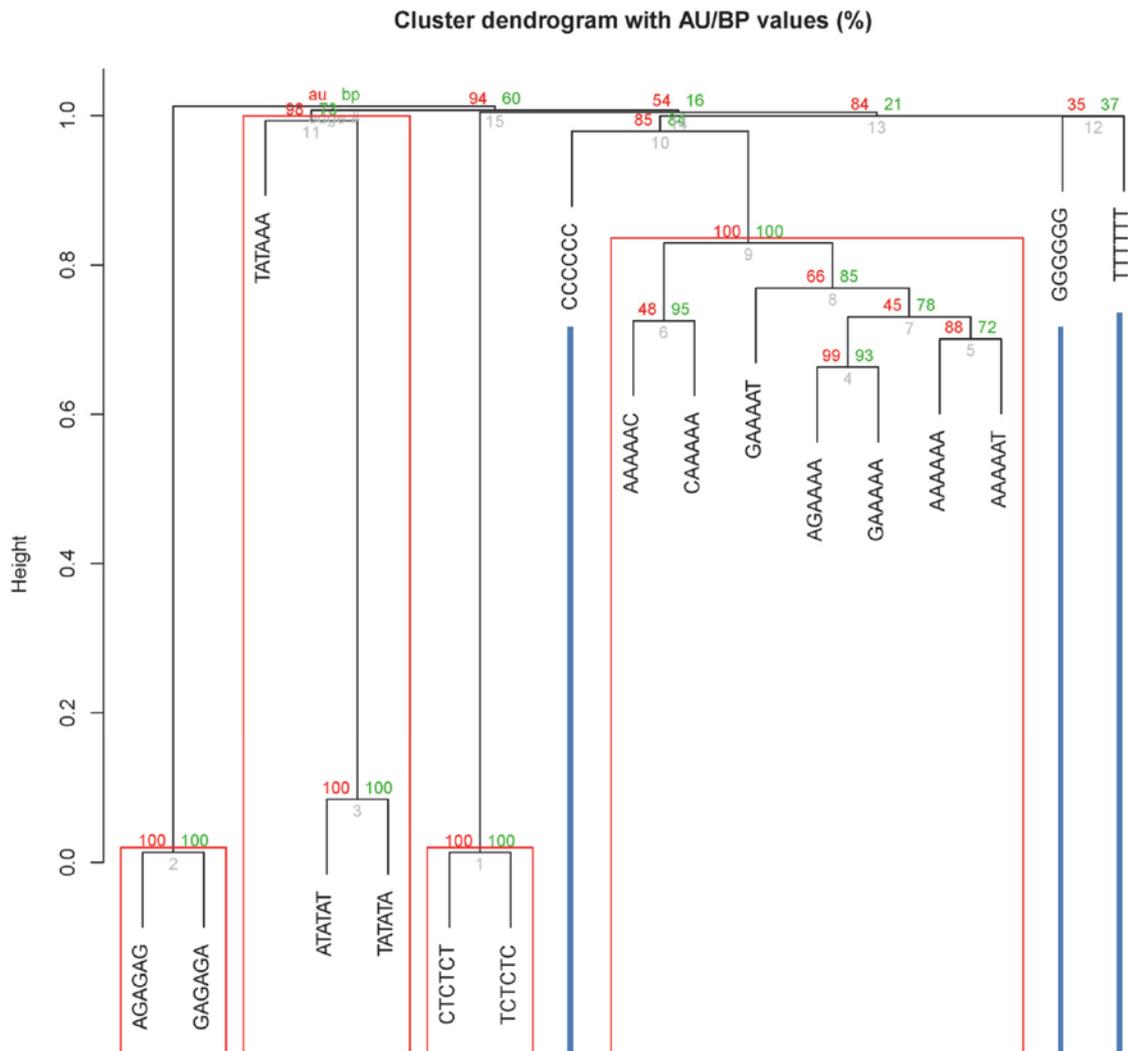


Figure 3.3. Bootstrapped hierarchical clustering (n=1000) showing hexamer co-occurrence across promoters. Red and green values show the Approximately Unbiased and Bootstrap Probability p -values. Grey values indicate node number. Red boxes indicate GA, TA, CT and A (of W) clusters with Approximately Unbiased p -value ≥ 0.95 . Blue lines indicate unclustered hexamers, specifically C and G (of S) and T (of W).

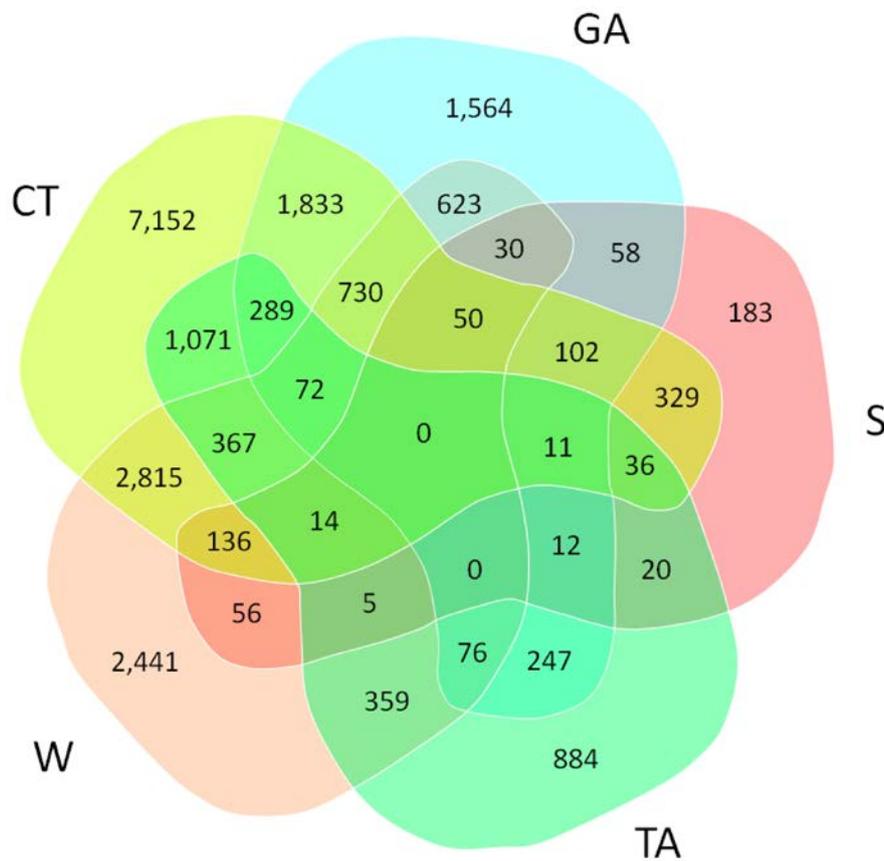
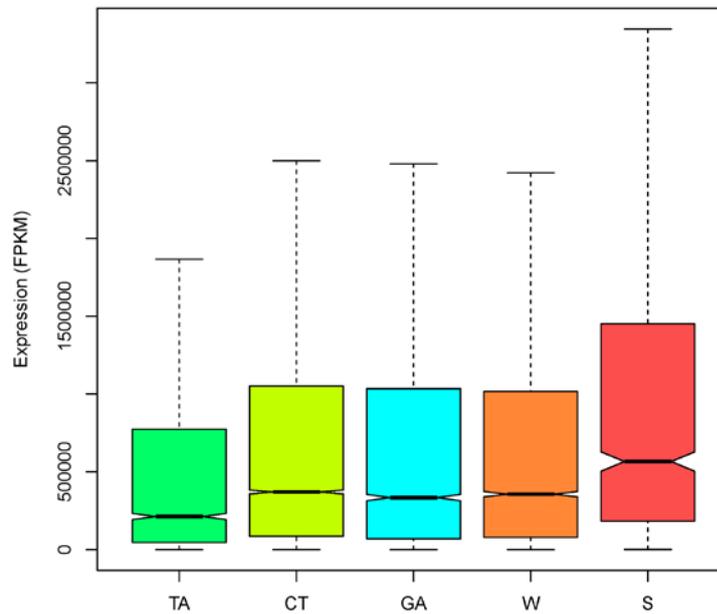


Figure 3.4. Venn diagram of core promoter class co-occurrence per promoter. 57% of promoters are annotated with a single core class, 34% with two core classes, 9% with three or more. Individual promoter data for Venn diagram cells are provided in Additional file 3.11.

(a)



(b)

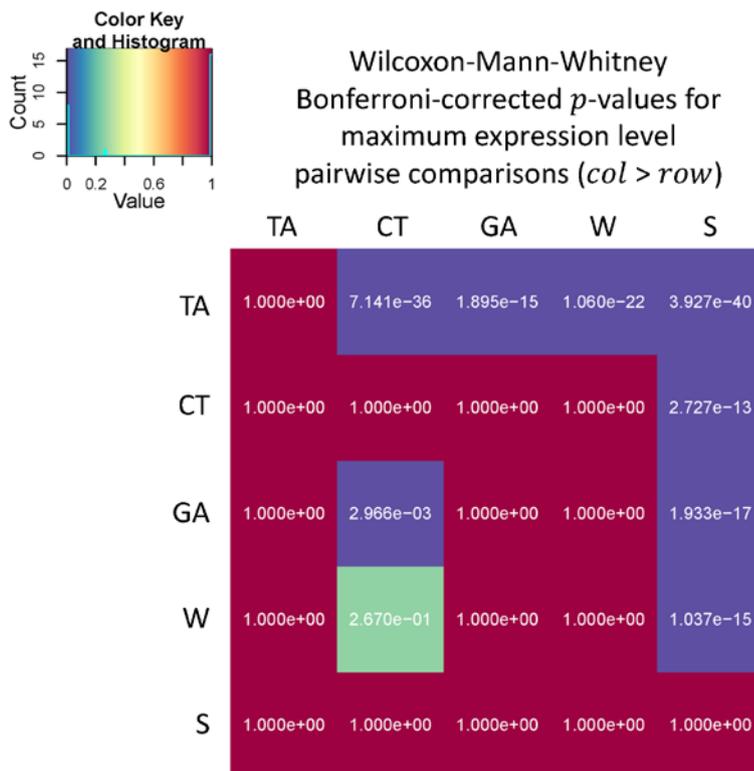
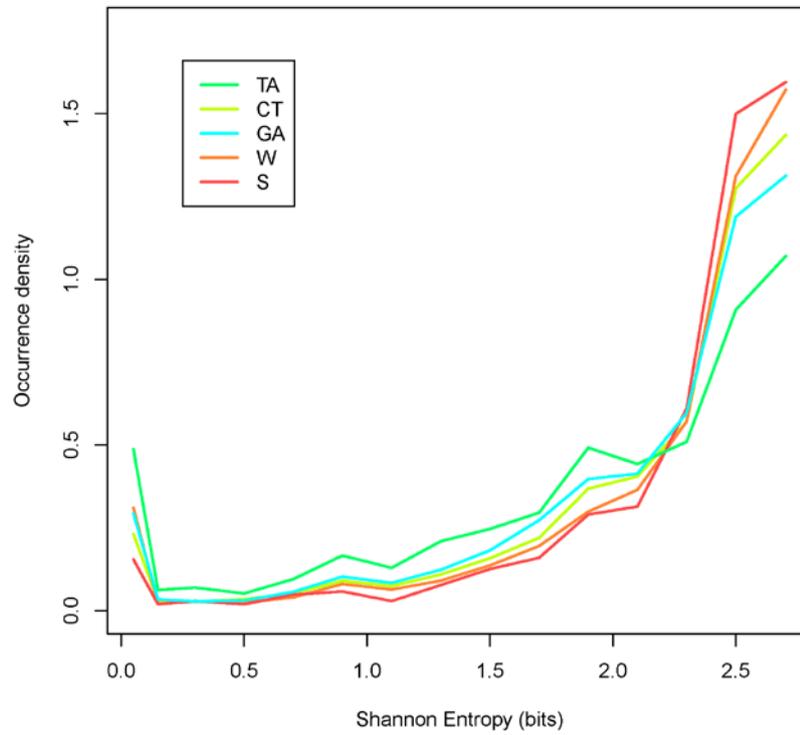


Figure 3.5. Expression profiles of *E. grandis* core promoter classes. (a) Distribution of maximum FPKM values (per gene) per promoter class. Outliers have been suppressed. (b) Bonferroni-adjusted p -values of the two-sample Wilcoxon-Mann-Whitney tests for the pairwise comparison of the expression level (FPKM) of each column promoter class (col) with the expression level of each row promoter class (row) with the alternative hypothesis that $col > row$.

(a)



(b)

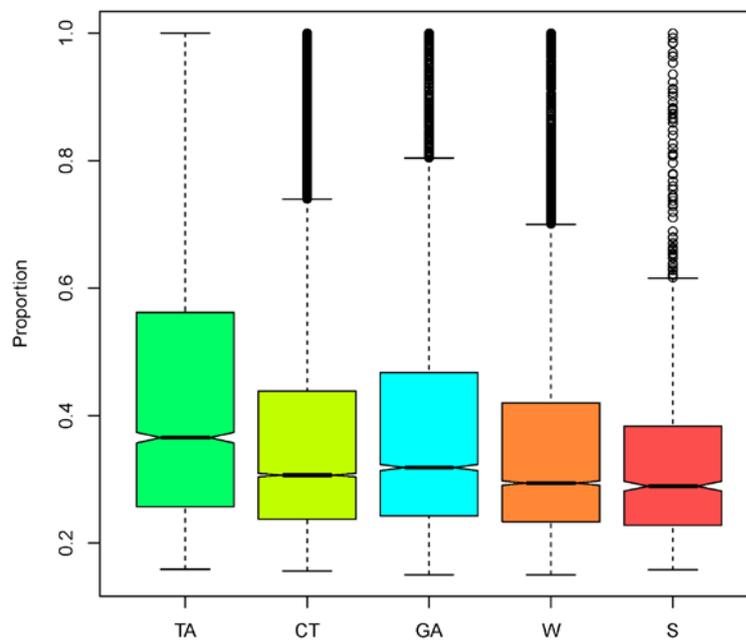


Figure 3.6. Expression specificity of *E. grandis* core promoter classes. (a) Density of Shannon Entropy values per core promoter class. (b) Distribution of maximum tissue expression as a proportion of total expression.

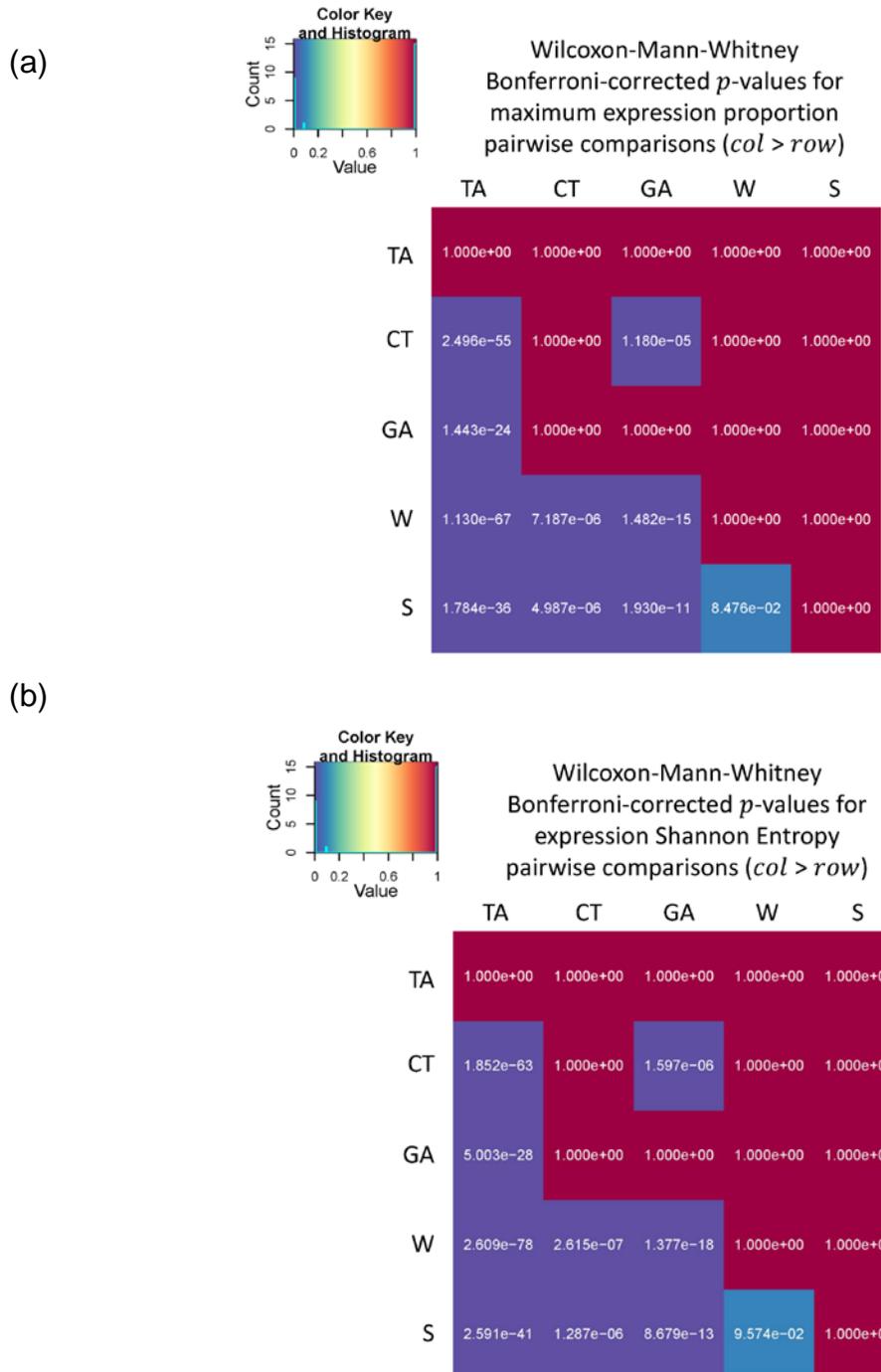
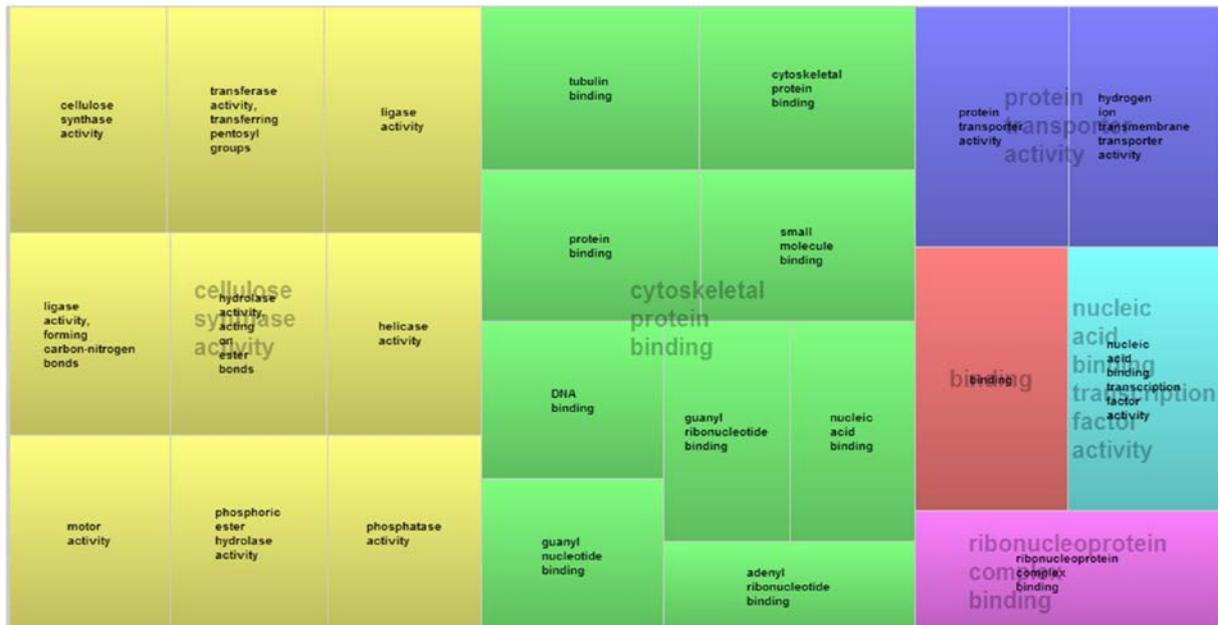


Figure 3.7. Expression specificity of core promoter classes. Bonferroni-adjusted p -values of the two-sample Wilcoxon-Mann-Whitney tests performing the (a) pairwise comparison of the expression specificity proportion of each column promoter class (col) with each row promoter class (row) with the alternative hypothesis that $col > row$; and (b) pairwise comparison of the tissue expression Shannon Entropy of each column promoter class (col) with each row promoter class (row) with the alternative hypothesis that $col < row$.

(a)



(b)



Figure 3.8. REVIGO TreeMap display of GO terms enrichment for two promoter classes. (a) over-represented Molecular Function GO terms of the CT promoter class and (b) over-represented Biological Process GO terms of TATATA, a constituent of the TA promoter class.

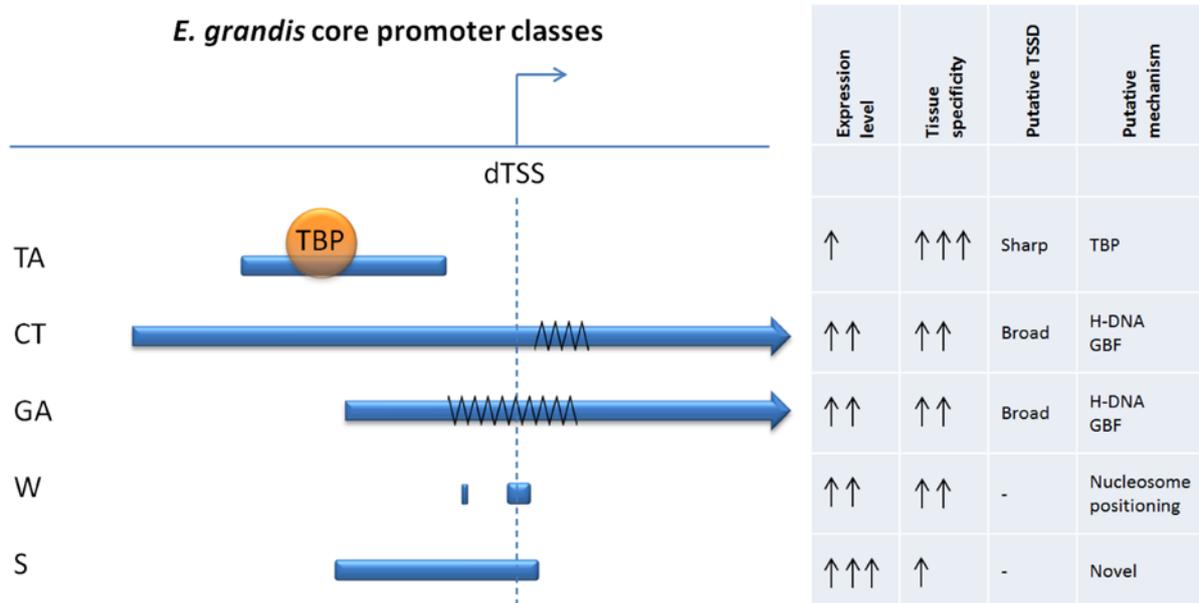


Figure 3.9. Summary of core promoter classes defined in *E. grandis*. The schematic shows DNA sequence from -100 to +50 and summarizes the expression level, tissue specificity, putative TSSD distribution and putative mechanism of transcriptional regulation. TBP = TATA BINDING PROTEIN; GBF = GAGA BINDING FACTORS.

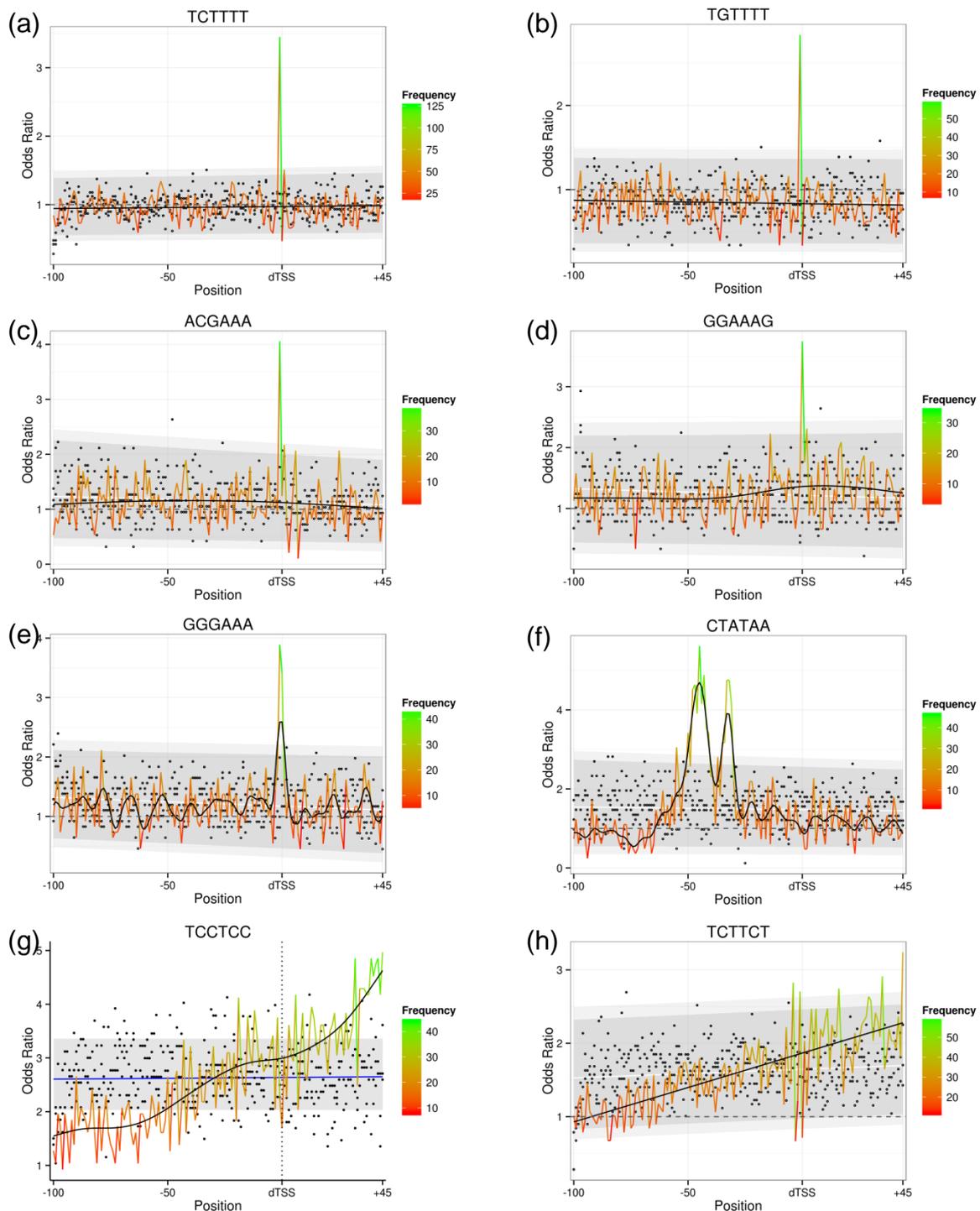
3.9 SUPPLEMENTARY MATERIAL

Supplementary Table 3.1. Formulaic contingency table for Fisher's Exact Test of hexamer over-representation. AAAAAA is used as an example. $\{K_1, \dots, K_Z\}$ = overlapping non-N 6mers

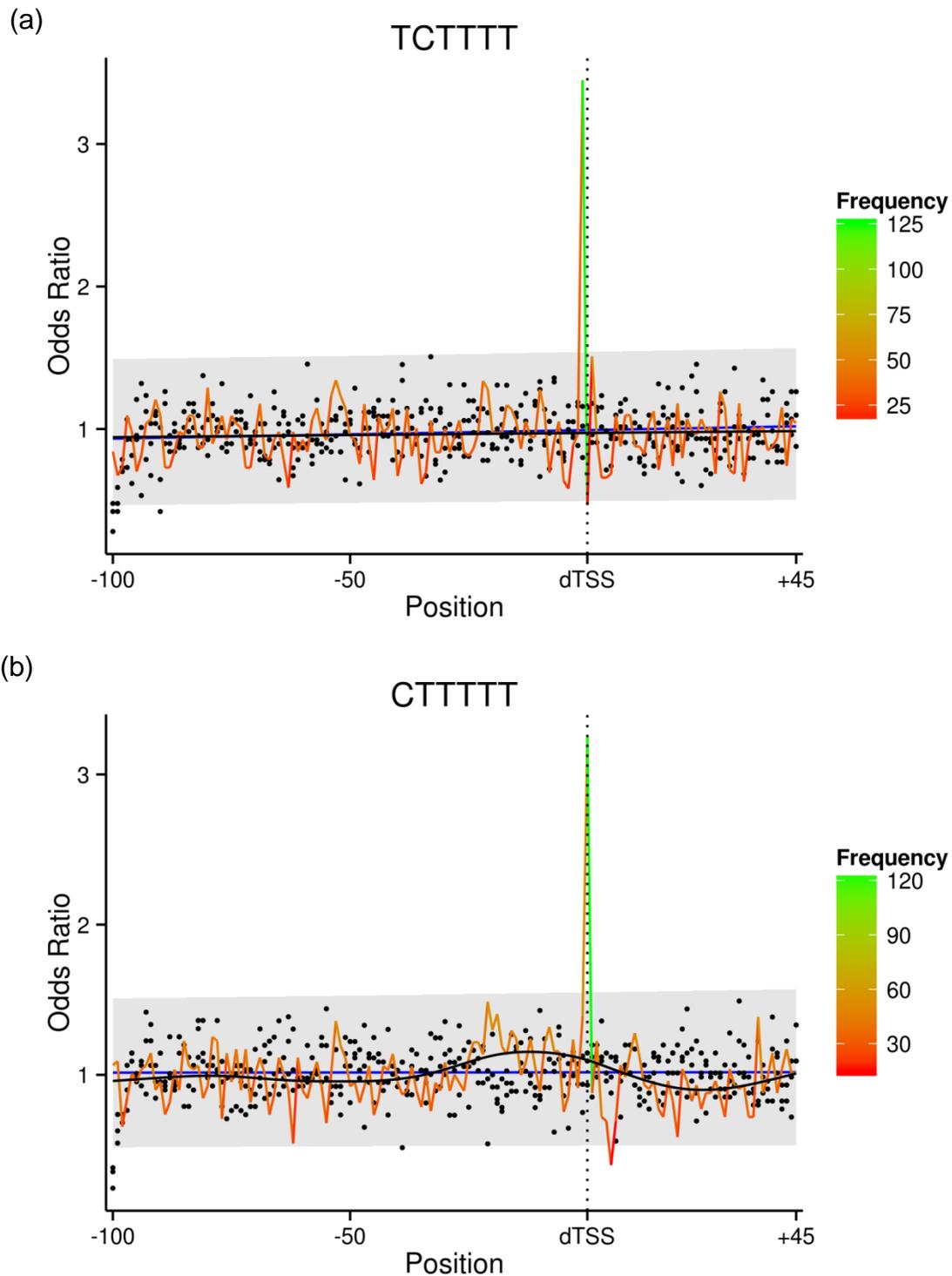
	<i>seg</i>	Distal
Observed	$\sum_{i=1}^Z \{seg_x(K_i) seg_x(K_i) = H_{AAAAAA}\}$	$\sum_{i=1}^Z \{dis(K_i) dis(K_i) = H_{AAAAAA}\}$
Expected	$\sum_{i=1}^Z \{seg_x(K_i)\}$	$\sum_{i=1}^Z \{dis(K_i)\}$

Supplementary Table 3.2. Text contingency table for Fisher's Exact Test of hexamer over-representation. Example using *AAAAAA* of $H \in \{AAAAAA, \dots, TTTTTT\}$.

	<i>seg</i>	Distal
Observed	Number of AAAAAAs found in the 6 bp segment of the test or control core promoter sets	Number of overlapping AAAAAAs found in the distal promoter set
Expected	Total number of non-N hexamers found in the segment	Total number of non-N overlapping hexamers in the distal promoter set



Supplementary Figure 3.1. Hexamer over-representation for types *spiked* and *low*. *Spiked* types (a) TCTTTT, (b) TGTTTT, (c) ACGAAA and (d) GGAAAG occur in close proximity to the dTSS and show marked and precise enrichment. *Low* types show (e) GGGAAA enrichment near the dTSS, (f) CTATAA, indicating C is enriched upstream of the TA constituent TATAAA, and (g) TCCTCC and (h) TCTTCT, pyrimidine rich Y Patch constituents showing gradual enrichment in a 5' to 3' direction. The start of each tested hexamer position is indicated on the x-axis. The black line indicates the smoothed spline of the odds ratio values. The dots and blue line represent the odds ratios of the core control sets $Q \in \{Q_1, Q_2, Q_3\}$ and the derived linear model respectively. The grey shaded area represents the 99% prediction interval (a-f) and the 10th to 90th percentile (g,h) determined by the core control set $Q \in \{Q_1, Q_2, Q_3\}$



Supplementary Figure 3.2. Enrichment of *spiked* hexamers at the dTSS. (a) TCTTTT and (b) CTTTTT are enriched at the dTSS supporting the initiator element \underline{TCT} (Kadonaga, 2012).

Supplementary Note 3.1. Pseudocode for the procedure to determine over-represented hexamers. Open-source software, when used, is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R; or a combination thereof.

Description: To determine the type, if any, of positional over-representation of hexamers in core promoters.
Input: For each hexamer, Fisher's Exact Test odds ratios (*OR*) and *q*-values for each *seg* $\in \{seg_1, \dots, seg_{145}\}$ for θ ; $Q \in \{Q_1, Q_2, Q_3\}$ and $R \in \{R_1, R_2, R_3\}$
Procedure:

```

1> FOR each hexamer IN  $H \in \{AAAAAA, \dots, TTTTTT\}$ 
2>   Filter for maximum frequency  $\geq 20$  and minimum q-value  $< 0.05$  in  $\theta$ 
3>   IF PASS
4>     Select core control data set Q as control C
5>     Linear model of maximum Q OR values ( $lm.Q_{max}$ )
6>     Linear model of median Q OR values ( $lm.Q_{med}$ )
7>     IF 95% prediction interval for  $lm.Q_{med} > 2.5$  and  $\sum Q$  frequencies  $> 30,000$ 
8>       Select core + distal control data set R as control C
9>       Linear model of maximum R OR values ( $lm.R_{max}$ )
10>       $lm.C_{max} = lm.R_{max}$ 
11>     ELSE
12>       $lm.C_{max} = lm.Q_{max}$ 
13>     Filter for broad_FILTER_CRITERIAa
14>     IF PASS
15>       Assign as broad constituent
16>       Write to output and plot figure
17>     ELSE
18>       Filter for spiked_FILTER_CRITERIAb
19>       IF PASS
20>         Assign as spiked constituent
21>         Write to output and plot figure
22>       ELSE
23>         Filter for low_FILTER_CRITERIAc
24>         IF PASS
25>           Assign as low constituent
26>           Write to output and plot figure
27>         ELSE
28>           Continue onto next hexamer

```

Output: Enriched sequence, enrichment region and enrichment score as R object, .csv file and figure for each over-represented hexamer
Validation: Supplementary Note 3.3

^a***broad_FILTER_CRITERIA***

1. Maximum frequency ≥ 65
2. *q*-value < 0.01
3. Area between smoothed spline and 99% prediction interval of $lm.C_{max}$ (a) > 0 (Additional file 3.12a)

^b***spiked_FILTER_CRITERIA***

1. Maximum frequency ≥ 20
2. *q*-value < 0.01
3. Single spike which shows at least double the enrichment of a second spike, if any.
4. Spike height (*k*) / 99% prediction interval height of $lm.C_{max}$ (b) > 2 (Additional file 3.12b)

^c***low_FILTER_CRITERIA***

1. Maximum frequency ≥ 40
2. *q*-value < 0.05
3. Area between smoothed spline and 90th percentile of control *C* (a) > 0 (Additional file 3.12c)

Supplementary Note 3.2. Procedure to cluster over-represented hexamers. Open-source software, when used, is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R; or the combination thereof.

Description:	Procedure to determine core promoter classes, given a list of enriched hexamers, and their respective enrichment profiles.
Input:	Extended promoter sequence [-500,+500] for all genes with a dTSS.
Procedure:	<ol style="list-style-type: none"> 1> FOR each significant hexamer <i>sig</i> IN <i>broad</i> 2> Search for which genes contain <i>sig</i> in its enrichment region 3> From gene occurrence create distance matrix of hexamer sequence and gene occurrence 4> Cluster hexamers based on gene co-occurrence ($AUp < 0.05$)¹ 5> Manually inspect hexamer clusters and unclustered hexamers 6> IF similar physical DNA properties and overlap of enrichment regions 7> Merge hexamers/clusters 8> Assign core promoter class 9> ELSE 10> Assign core promoter class
Output:	Core promoter classes and their hexamer constituents
Open-source tools:	<ol style="list-style-type: none"> 1. Pvclost (Suzuki & Shimodaira, 2006) <code>pvclost(method.hclust="single",nboot=1000); pvrect(alpha=0.95)</code>

Supplementary Note 3.3. Procedure to determine core promoter class enrichment regions and gene lists. Open-source software, when used, is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R; or the combination thereof.

Description:	Procedure to determine the enrichment region of a core promoter class and annotate gene groups, given the constituent hexamers.
Input:	Significant hexamers in each core promoter class Extended promoter sequence [-500,+500] for all genes with a dTSS
Procedure:	<ol style="list-style-type: none"> 1> FOR each group of significant hexamers <i>sigs</i> IN {TA, CT, GA, W, S} 2> Search for cumulative occurrence of <i>sigs</i> in extended promoter¹ 3> Linear model for occurrence frequency in [-500,-100] (<i>lm</i>) 4> Use <i>lm</i> to determine 99% prediction interval for [-500,+500] 5> IF frequency > 99% prediction interval 6> Classify region of enrichment 7> FOR each gene's promoter 8> IF promoter possesses enriched hexamers in class enrichment region 9> concatenate gene ID to core promoter class gene list
Output:	Enrichment region, graph of occurrence and enrichment region and gene list
Open-source tools:	<ol style="list-style-type: none"> 1. dna-pattern; Regulatory Sequence Analysis Tools (http://rsat.ulb.ac.be/, Turatsinze <i>et al.</i>, 2008; Thomas-Chollier <i>et al.</i>, 2008)

Supplementary Note 3.4. Procedure to determine expression level and specificity. Open-source software, when used, is indicated below with full parameter usage description. All other parsing and analysis modules/pipelines execute using Bash, Make, Awk, Sed, Python or R; or the combination thereof.

Description:	Procedure to determine trends in core promoter class expression level and specificity through pairwise comparison of classes using Wilcoxon-Mann-Whitney (WMW) tests.
Input:	Gene list for each core promoter class Mean FPKM values for shoot tip (ST), young leaf (YL), mature leaf (ML), flower (FL), root (RT), phloem (PH) and immature xylem (IX) Tissue proportions of total expression per gene
Procedure:	<ol style="list-style-type: none"> 1> FOR each core promoter class gene list 2> Extract FPKM values 3> FOR each gene 4> Extract maximum FPKM value 5> Perform WMW “greater than” test for each pairwise comparison of maximum expression distributions 6> FOR each core promoter class gene list 7> Extract expression proportion values 8> FOR each gene 9> Determine Shannon Entropy¹ 10> Extract maximum proportion value 11> Perform WMW “greater than” test for each pairwise comparison of maximum expression proportion 12> Perform WMW “less than” test for each pairwise comparison of Shannon Entropy distributions
Output:	WMW q -values for pairwise comparison of core promoter classes for i) expression level, ii) expression proportion and iii) specificity Shannon Entropy.
Open-source resources:	<ol style="list-style-type: none"> 1. Code adopted from http://davetang.org/muse/2013/08/28/tissue-specificity/

3.10 ADDITIONAL FILES

Additional file 3.1

3_1_Hexamer_over_representation_profiles.xlsx

Excel formatted file detailing type, density and region of significant hexamers

Additional file 3.2

3_2_Enrichment_distributions_full_size.zip

Compressed folder of core promoter class enrichment distributions in [-500,+500]

Additional file 3.3

3_3_CT_and_GA_repeats.xlsx

Excel workbook with frequency of CT and GA repeat lengths in the core promoter

Additional file 3.4

3_4_A_rich_W_class_hexamers.zip

Compressed folder of A rich significant hexamers which are constituents of W

Additional file 3.5

3_5_Heatmaps_of_expression_proportions_per_class.pdf

PDF file of heatmaps showing expression proportions between tissues for each gene of each core promoter class.

Additional file 3.6

3_6_TopGO_results.zip

Zipped folder containing i) Excel workbook of TopGO results, ii) raw TopGO output and iii) REVIGO TreeMap representation of over-represented GO terms for each GO category.

Additional file 3.7

3_7_CT_and_GA_co_occurring_GO_terms.xlsx

Excel workbook with co-occurring GO terms between CT and GA, showing their significance in both classes.

Additional file 3.8

3_8_Enrichment_distribution_figures_for_spiked_and_low.zip

Compressed folder of significant hexamers of types *spiked* and *low* respectively

Additional file 3.9

3_9_Narrow_type_distance_from_dTSS.xlsx

Excel workbook showing the distance of *spiked* peaks from the dTSS

Additional file 3.10

3_10_Cluster_of_broad_spiked_and_low_hexamers_dendrogram.pdf

PDF of the dendrogram generated from `pvclust` hierarchical clustering of all significant hexamers combined.

Additional file 3.11

3_11_Genes_for_each_promoter_class.zip

Compressed folder of gene lists (.txt) for each core promoter class

Additional file 3.12

3_12_Graphs_for_broad_spiked_and_low_filtering_criteria.pdf

PDF with example distributions and respective schematics of the filtering criteria used to define *broad*, *spiked* and *low* types.

CHAPTER 4 : Concluding Remarks

4.1 A GENOME-WIDE STUDY OF TRANSCRIPTION IN *EUCALYPTUS GRANDIS*

Eucalyptus grandis is an economically important hardwood fibre crop species. The cellulose yield from *E. grandis* hardwood is of predominant industrial value in pulp and paper production and various other chemical cellulose applications (Mizrachi *et al.*, 2012). The draft *E. grandis* genome assembly and annotation (Phytozome V1.0) were made available in January 2011 (Myburg *et al.*, in press), allowing for genome-wide and base pair-resolution analyses of genome structure and function for downstream applications such as enhanced growth and wood properties. The results of these analyses are, however, dependent on the accuracy of the assembly (5,379 scaffolds of which 85% are in 12 large chromosome-level scaffolds) and gene model predictions (36,376). Key biological processes underlying fibre secondary cell wall properties, including the deposition of lignin, cellulose and hemicellulose, are under strong transcriptional control (see Hussey *et al.*, 2013 for review). An important approach for studying transcriptional control is the genome-wide characterisation of core promoters (Molina & Grotewold, 2005), an approach made possible in *E. grandis* by the availability of the genome annotation as well as extensive cDNA sequence resources (EST and RNA-Seq).

In Chapter 2, *E. grandis* 5' UTRs were empirically curated using available transcript evidence, specifically EST and high-throughput mRNA-Seq data. This included ~2.9 million *E. grandis* ESTs (produced by 454 RNA sequencing) and paired-end mRNA-Seq data from seven developing tissues of *E. grandis* trees. The current Phytozome (Goodstein *et al.*, 2012) *E. grandis* 5' UTR models were analysed and the concurrence of empirical evidence and FGenesH (Solovyev *et al.*, 2006) predictions was assessed. A prioritised set of 5' UTR models was selected from the above three sources. It was expected that including and prioritizing empirically substantiated models would enrich the total number and quality of *E. grandis* 5' UTR models. An important limitation of using mRNA-seq data, by itself, was the inability to confidently determine TSSDs. 5'UTRs with long tails of low coverage were identified, but it was indeterminable whether this was from true biological transcript variation, or technical artefacts from either the library preparation or read alignment. Nevertheless, *in lieu* of TSSD annotations, the TSSs inferred by these models are referred to as distal TSSs (dTSS) as they represent the most 5' position of transcription evidence. This annotation provides insight into, rather than what position stochastically favours transcription, what position is the first base of, and thus delimits, permissive transcription.

In Chapter 3, the curated dTSSs were used to extract and annotate core promoters in *E. grandis* by defining core promoter classes and assessing their functional characteristics. Five core promoter classes were distinguished, each putatively comprising different underlying mechanisms of transcriptional control, including TBP affinity (TA), ssDNA extrusion (CT, GA), nucleosome positioning (W) and inherent DNA stability (S). This study used a lexical enrichment analysis to determine conserved DNA signatures in core promoters across the genome, and allowed the inference of physiochemical properties driving transcription initiation based on the properties of these bp interactions. The converse has also been applied, using physiochemical properties to predict promoters (Florquin *et al.*, 2005), validating physiochemical dominance in delimiting permissive transcription. As with all genome-wide enrichment studies, input data can contain noise and tarnish both sensitivity and specificity. This can be controlled to an extent by the appropriate selection of test and background data, but often one metric has to be favoured over the other. This study favoured specificity and thus tried to limit the number of false positive enriched hexamers. Despite this, it is possible that some genes have been assigned to a core promoter class erroneously (false positive), where the sequence composition is by chance and not as a result of driving the core expression and rendering transcription permissive. This may be a result of error introduced in both the core promoter gene group assignment by combining several hexamers to a class, or in the assignment of dTSSs. Although our results corroborate the presence of both the canonical (TATATA) and weaker variant (TATAAA) of the TATA-box, these results require further investigation. It is possible that the bimodality is an artefact of the dTSS annotation, with predicted 5' UTRs having a strong bias for TATA-box signatures, and thus the two modes representing empirical and predicted 5' UTRs. Alternatively, this could provide insight into the mechanisms of the more specific TSS selection (Yamamoto *et al.*, 2009) exhibited by TBP bound TATA-box promoters, suggesting they may show some phasic constraint. While this study finds only 12% of promoters with a canonical TBP consensus, it is possible that W promoters, possessing regions with poly-T tracts, may be coerced into ssDNA and capable of binding TBP (Ahn *et al.*, 2012). Finally, Venters & Pugh (2013) show that 85% of TBP-bound sequence have zero to three mismatches from the human TATA-box consensus TATAWAWR. This suggests that while we have found the strongest consensus (TATAWA) for *E. grandis*, there are likely less stringent sequences which bind TBP, albeit it with lower affinity (Savinkova *et al.*, 2013).

The most valuable contribution of this research has been the novel sequence-driven insight into the possible mechanisms of permissive transcription in plant species, and more specifically in charismatic megaflores such as *Eucalyptus*. Metazoan promoters are characterised as TATA and methylated CpG promoters respectively (Lenhard *et al.*, 2012). Plant promoters are known to lack CpG enrichment and are similarly devoid of the associated DNA methylation patterns thought to permit active transcription. Until this point, studies on core promoters featured analyses on DNA regions defined by the prominent or peak TSSs. In the only plant study to date which incorporated the true range of transcription initiation sites as a TSSD (Yamamoto *et al.*, 2009), the GA repeat element was discovered in *A. thaliana*, whilst remaining undetected when using traditional TSS annotations (Molina & Grotewold, 2005; Yamamoto *et al.*, 2007). Our promoter analysis uses dTSSs to determine those promoter features which make DNA permissive to transcription. In doing so, we have corroborated the GA repeat element, as well as the TATA-box, and described other core promoter classes which are thus far unique to *E. grandis*, but may be discovered with similar approaches in other plants. Importantly, this study has shown that other mechanisms putatively control the binding and assembly of the PIC in the absence of consensus driven affinity of TBP, particularly that of putative nucleosome positioning. It is expected that future research will show these features to be conserved between at least plant species and that the novel enrichment observed in this study is a result of the definition of the core promoter by the dTSS.

4.2 FUTURE PERSPECTIVES

This study has used full-length mRNA-seq for the empirical curation of dTSSs and sequence over-representation for functional characterization of core promoter classes, providing valuable insight into *E. grandis* transcriptional mechanisms. The core promoters were functionally characterised by GO analysis and expression level and specificity metrics and thus achieved the aims outlined in the first chapter. There are, however, many technologies, and applications thereof, which can address targeted transcriptional research at a finer resolution. For the identification of permissive transcription and TSSDs, combined 3' and 5' capture and mapping of full-length transcripts will improve the understanding of transcription initiation stochasticity and TSSDs. For a finer, single-cell resolution of transcription initiation, applications such as single-cell RNA-seq and RNA fluorescence *in situ* hybridization (Kim & Marioni, 2013). This study suggests that nucleosome positioning is

critical in determining the first base of permissive transcription, specifically for the W core promoter class. Nucleosome studies including those of positioning (Brown *et al.*, 2013) and histone modifications (Kaplan *et al.*, 2009; Rach *et al.*, 2011; Du *et al.*, 2013) will be necessary to either confirm or refute these and provide further insight into sequence specific characteristics of nucleosome affinity and how these govern transcription by TF binding competition or nucleosome depleted regions. DNA methylation and studies providing insight for non-B-DNA structure and the epigenetic architecture will provide context on how the core promoter exerts its function in four dimensional space within the nucleus, governed by physiochemical interactions.

Functional characterisation of expression level, as inferred by RNA steady-state abundance, can be fully explored by a growing compendium of expression data across diverse tissues, and in several stress and environmental responses, whereas in this study, only tissue specificity is considered. Measuring both the expression level (eg. microarray, qPCR, RNA-seq) and the regulatory genotype (DNA-seq of promoters and 5' UTRs) of a population, allows per-individual resolution in the detection of natural promoter variation and the impact on DNA-TF affinity and transcription. *Cis*-elements or enhancers are expected to modulate expression for regulated genes, however, it is possible that core promoter classes have alternative likelihoods of using enhancer elements for expression modulation, with strong nucleosome positioning being negatively correlated with enhancer modulation (Kaplan *et al.*, 2009; Jeziorska *et al.*, 2009). The combinatorial control of core promoter and *cis*-regulatory elements diversifies outputs, and co-expression studies which identify regulons of genes modulated by the same TF or semi-hierarchical network (Mentzen & Wurtele, 2008; Vandepoele *et al.*, 2009; Zheng *et al.*, 2011), or ChIP-seq experiments (Jothi *et al.*, 2008; Mason *et al.*, 2010; Håndstad *et al.*, 2011) can be used to characterise the TFBSs.

This study draws comparison to other plant species' core promoters based on findings reported in plant promoter literature. To remove bias and variation, a cross-species comparison using the same methodology and rigor would not only corroborate findings, but identify species or clade specific core promoter signatures and putative transcriptional mechanisms. Comparisons made between core promoters of orthologous and paralogous genes would provide further insight on core promoter evolution. Finally, the integration of all these resources and data types will begin to holistically describe transcription initiation (Ernst *et al.*, 2010), stochasticity and mechanisms of permissive and modulated transcription, not

only in *E. grandis*, but across all plant species through a true Systems Genetics approach. Understanding the complex phenomenon of successful transcription, amongst many other applications, will facilitate research towards cellulose-enriched fibre cells in *Eucalyptus* plantation species.

4.3 ACKNOWLEDGMENTS

I would like to acknowledge all members of the Forest Molecular Genetics group at the University of Pretoria for their critical discussion of *Eucalyptus* biology. I would also like to acknowledge Alexander A. Myburg, Eshchar Mizrahi and Fourie Joubert for their insight and review of this manuscript.

4.4 REFERENCES

Ahn S, Huang C-L, Ozkumur E, Zhang X, Chinnala J, Yalcin A, Bandyopadhyay S, Russek S, Unlü MS, DeLisi C, et al. 2012. TATA binding proteins can recognize nontraditional DNA sequences. *Biophysical Journal* **103**: 1510–1517.

Brown CR, Mao C, Falkovskaia E, Jurica MS, Boeger H. 2013. Linking stochastic fluctuations in chromatin structure and gene expression. *PLoS Biology* **11**: e1001621.

Du Z, Li H, Wei Q, Zhao X, Wang C, Zhu Q, Yi X, Xu W, Liu XS, Jin W, et al. 2013. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa* L. Japonica. *Molecular Plant* **6**: 1463–1472.

Ernst J, Plasterer HL, Simon I, Bar-Joseph Z. 2010. Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Research* **20**: 526–536.

Florquin K, Saeys Y, Degroeve S, Rouzé P, Van de Peer Y. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Research* **33**: 4255–4264.

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, et al. 2012. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research* **40**: D1178–1186.

Håndstad T, Rye MB, Drabløs F, Sætrom P. 2011. A ChIP-Seq benchmark shows that sequence conservation mainly improves detection of strong transcription factor binding sites. *PLoS One* **6**: e18430.

Hussey SG, Mizrahi E, Creux NM, Myburg AA. 2013. Navigating the transcriptional roadmap regulating plant secondary cell wall deposition. *Frontiers in Plant Science* **4**: 325.

- Jeziorska DM, Jordan KW, Vance KW. 2009.** A systems biology approach to understanding *cis*-regulatory module function. *Seminars in Cell & Developmental Biology* **20**: 856–862.
- Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. 2008.** Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Research* **36**: 5221–5231.
- Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J, et al. 2009.** The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**: 362–366.
- Kim JK, Marioni JC. 2013.** Inferring the kinetics of stochastic gene expression from single-cell RNA-sequencing data. *Genome Biology* **14**: R7.
- Lenhard B, Sandelin A, Carninci P. 2012.** Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nature Reviews Genetics* **13**: 233–245.
- Mason MJ, Plath K, Zhou Q. 2010.** Identification of context-dependent motifs by contrasting ChIP binding data. *Bioinformatics* **26**: 2826–2832.
- Mentzen WI, Wurtele ES. 2008.** Regulon organization of *Arabidopsis*. *BMC Plant Biology* **8**: 99.
- Mizrachi E, Mansfield SD, Myburg AA. 2012.** Cellulose factories: advancing bioenergy production from forest trees. *New Phytologist* **194**: 54–62.
- Molina C, Grotewold E. 2005.** Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* **6**: 1471–2164.
- Rach EA, Winter DR, Benjamin AM, Corcoran DL, Ni T, Zhu J, Ohler U. 2011.** Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genetics* **7**: e1001274.
- Savinkova L, Drachkova I, Arshinova T, Ponomarenko P, Ponomarenko M, Kolchanov N. 2013.** An experimental verification of the predicted effects of promoter TATA-box polymorphisms associated with human diseases on interactions between the TATA boxes and TATA-binding protein. *PloS One* **8**: e54626.
- Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006.** Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology* **7 Suppl 1**: S10.1–12.
- Vandepoele K, Quimbaya M, Casneuf T, De Veylder L, Van de Peer Y. 2009.** Unraveling transcriptional control in *Arabidopsis* using *cis*-regulatory elements and coexpression networks. *Plant Physiology* **150**: 535–546.
- Venters BJ, Pugh BF. 2013.** Genomic organization of human transcription initiation complexes. *Nature* **502**: 53–58.

Yamamoto YY, Ichida H, Matsui M, Obokata J, Sakurai T, Satou M, Seki M, Shinozaki K, Abe T. 2007. Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics* **8**: 67.

Yamamoto YY, Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J. 2009. Heterogeneity of *Arabidopsis* core promoters revealed by high-density TSS analysis. *The Plant Journal* **60**: 350–362.

Zheng X, Liu T, Yang Z, Wang J. 2011. Large cliques in *Arabidopsis* gene coexpression network and motif discovery. *Journal of Plant Physiology* **168**: 611–618.