Genome re-annotation and DNA motif identification in Brassicaceae species

Inaugural-Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Vimal Rawat

(aus Delhi, India)

Köln, 2017

Diese Arbeit wurde am Max-Planck-Institut für Züchtungsforschung in Köln in der Abteilung für Entwicklungsbiologie durchgeführt.

Max-Planck-Institut für
Pflanzenzüchtungsforschung

IMPRS Cologne

Berichterstatter:

Prof. Dr. George Coupland

Prof. Dr. Thomas Wiehe

Prüfungsvorsitzender:

Prof. Dr. Martin Hülskamp

Tag der mündlichen Prüfung: Januar 19, 2016

New opinions are always suspected, and usually opposed, without any other reason but because they are not already common.

— John Locke (English Philosopher)

Table of Contents

Chapter 3

List of Figures

List of Tables

List of Abbreviations

ABA response element like	ABREL
Average	Avg.
Basic Local Alignment Search Tool	BLAST
BLAST-like alignment tool	BLAT
Base pair	bp
Biological rhythms analysis software system	BRASS
CCA1 binding site	CBS
Cross Correlation	CC
Circadian clock associated 1	CCA1
Complementary DNA	cDNA
Coding sequence	CDS
Chromatin immunoprecipitation	ChIP
Chromatin immunoprecipitation-sequencing	ChIP- seq
Conserved non-coding sequence	CNS
Cis-regulatory element	CRE
Cis-regulatory module	CRM
Circadian Time	CT
Coefficient of variation	CV
Diurnal DNA motif	DDM
DNase I hyper-sensitive site	DHS
Deoxyribonucleic acid	DNA
Evening Element	EE
Early flowering	ELF
Expectation maximization	EM
Expressed Sequence Tag	EST
Fold change	FC
False discovery rate	FDR
Fragment Per Kilobase of transcript per Million mapped reads	FPKM
genomic DNA	gDNA

Gene Feature Format	GFF
Gene ontology	GO
High confidence	HC
hidden Markov model	HMM
Hours	hrs.
Heat stressed	HS
Hormone Up and Down box	HUD box
Isoform sequencing	Iso-seq
Kilobase	kb
Low confidence	LC
Late elongated hypocotyl	LHY
Long terminal repeat	LTR
Lux arrhythmo	LUX
Markov cluster algorithm	MCL
Morning element	ME
Megabase	Mb
micro RNA	miRNA
Model-based Periodicity Screening	MoPS
Massive parallel sequencing	MPS
Next-generation sequencing	NGS
Open reading frame	ORF
Protein box	PBX
Principal component analysis	PCA
Polymerase chain reaction	PCR
Phylogenetic Footprinting	PF
Position-specific scoring matrix	PSSM
Position weight matrix	PWM
Relative amplitude error	RAE
Recovered	REC
Ribonucleic acid	RNA
RNA-sequencing	RNA-seq

Reverse transcription-PCR	RT-PCR
Sequencing by synthesis	SBS
Starch synthesis box	SBX
Shoot apical meristem	SAM
Single-molecule real-time	SMRT
Sequencing by Oligo Ligation Detection	SOLiD
Standard error in measurement	SEM
Telebox	TBX
Transposable element	TE
Transcription factor	TF
Transcription factor binding sites	TFBS
Timing of CAB expression 1	TOC1
Time-point specific diurnal DNA motif	tpsDDM
transfer RNA	tRNA
Transcription start site	TSS
Untranslated region	UTR
Whole genome sequencing	WGS
Wild-type	WT

Summary

The DNA sequence analysis field has experienced a paradigm shift caused by the drastic reduction in the sequencing cost and time. With the availability of several reference genome assemblies, understanding of structural and functional aspects of genomes has started growing. Annotating a reference genome is the first and very crucial step that ensures its efficient usability to serve as a community resource. Unlike coding regions, non–coding regions do not translate into proteins but still play a central role in development and physiology of an organism by regulating gene expression. Identification and annotation of these regions are only initial steps, equally interesting and even more rewarding is to decipher the interplay between these two components of a genome. Identification of *cis*-regulatory elements (CREs), the functional components of the non-coding genome, is paramount to our understanding regarding the gene expression regulation. The role of CREs in regulating rhythmic (diurnal) expression of thousands of genes has been reported in several plants species (including *Arabidopsis thaliana*) but still only a few CREs have been reported so far.

In the first project, using extensive RNA-sequencing data, I substantially improved the annotation and usability of a Brassicaceae species, *Arabidopsis lyrata.* Gene model coordinates for over 90% genes are corrected, with improved UTRs (untranslated regions) annotation. Over 2,000 genes are now annotated as transposable element (TE)-related genes and around 8% annotated with alternate transcripts. With hundreds of cases of gene-merge and gene-split, improved annotation also corrects coding space of the genome. Experimentally validated data for several such cases strongly supported updated annotation, highlighting the importance of employing species-specific RNA-sequencing data for genome annotation.

In the second project, I compared time-series transcriptomics data for two Brassicaceae species, *Arabidopsis thaliana* and *Arabis alpina.* Around 30% genes were found under the control of diurnal regulation in both species. An interesting finding regarding phase-shift of the circadian clock genes and their direct targets was also observed. Gene Ontology term enrichment analysis suggested that diurnal genes associated to carbohydrate metabolism are the most affected by this phase shift while light-signaling associated genes are the least affected. I also demonstrated the usefulness of Phylogenetic shadowing to identify enriched CREs in the diurnal genes. Using several recently

assembled Brassicaceae genomes, I analyzed the conservation patterns in promoters of orthologous diurnal genes. In total, I identified 54 and 45 DNA motifs for *Arabidopsis thaliana* and *Arabis alpina* respectively. Over 65% motifs were found common for both species including previously reported six motifs. Based on recently published open chromatin data, around 30% of the DNA motifs revealed protected sites from an endonuclease (DNase I), indicating their potential role as protein-binding sites. Several phase-specific co-occurring DNA motifs pairs were found conserved in both species, including previously known Evening Element (EE) and ABA Response Element Like (ABREL) pair, underlining the broad conservation of *cis*-regulation of diurnal expression.

Zusammenfassung

Das Feld der DNA-Sequenzanalyse hat sich, vor allem durch die drastisch gesunkenen Sequenzierungskosten sowie durch den verminderten Zeitbedarf, stark gewandelt. Mit der Verfügbarkeit mehrerer Referenzgenomassemblierungen hat ein wachsendes Verständnis der strukturellen und funktionellen Aspekte des Genoms begonnen. Die Annotation eines Referenzgenoms ist dem entsprechend ein erster wichtiger Schritt, der einer effizienten Nutzung als gemeinschaftlicher Ressource dient. Im Gegensatz zum codierenden Teil des Genoms wird der nicht-codierende Teil nicht in Proteine übersetzt, spielt aber dennoch eine zentrale Rolle in der Regulierung der Genexpression und damit in der Entwicklung und Physiologie von Organismen. Mit der Identifizierung und Annotation dieser Teile des Genoms ist jedoch nur ein erster Schritt getan. Darüber hinaus ist die Entschlüsselung des Zusammenspiels von codierenden und nicht-codierenden Bereichen eine ebenso interessante wie aufschlussreichere Fragestellung. Die Identifizierung von *Cis*-Regulatorischen Elementen (CREs) sowie deren Funktion in der Genregulation und Expression ist entscheidend für das Verständnis des nicht-codierenden Teils des Genoms. Für die tagesrhythmische Expression tausender von Genen in verschiedenen Pflanzenarten (einschließlich *Arabidopsis thaliana*) spielen die CREs eine zentrale Rolle, dennoch sind bis heute nur wenige CREs beschrieben.

Im ersten Teil meiner Arbeit wurde durch die Einbeziehung von umfangreichen RNA-Sequenzdaten, ist es mir gelungen die Annotation und deren Benutzerfreundlichkeit für eine Brassicaceae Art, *Arabidopsis lyrata*, wesentlich zu verbessern. Für mehr als 90 % der Gene haben sich Genmodellkoordinaten aufgrund der verbesserten „Un-Translated Region“-basierten Annotation verändert. Tausende Gene sind dadurch als „Transposable Element“ annotiert worden, zudem ist für rund 8 % der Gene alternative Transkription identifiziert worden. Hunderte Gene wurden entweder mit anderen Genen zu einem Gen verbunden oder voneinander getrennt, so konnte die Annotation des codierten Teils entscheidend verbessert und korrigiert werden. Diese Verbesserung konnte durch experimentelle Daten für mehrere Gene belegt werden, was die Bedeutung von artspezifischen RNA-Sequenzdaten für die Genannotation deutlich macht.

Im zweiten Teil dieser Arbeit habe ich Daten aus Transkriptionszeitreihen von zwei Brassicaceae Arten, *Arabidopsis thaliana* und *Arabis alpina,* verglichen. Dabei konnte ich zeigen, dass rund 30 % der Gene dieser Arten tagesrhythmisch exprimiert werden.

Zwischen den Arten wurde eine interessante Verschiebung der Phase von rhythmisch zirkulierenden Genen beobachtet. Eine ontologische Analyse bezüglich des vermehrten Auftretens von tagesrhythmisch exprimierten Genen zeigt, dass Kohlenhydratstoffwechsel-assoziierte Gene am stärksten in ihrer Phase verschoben, Lichtsignal-assoziierte Gene hingegen am wenigsten beeinflusst sind. Darauf aufbauend wurden mittels „Phylogenetic Shadowing“ CREs gesucht die vermehrt in der tagesrhythmischen Genregulation vorkommen. Dabei war es möglich die Konservierungsmuster in orthologen Promotoren der tagesrhythmischen Gene anhand von mehreren kürzlich assemblierten Brassicaceae Genomen zu analysieren. So wurden 54 beziehungsweise 45 DNA-Motive für *Arabidopsis thaliana* und *Arabis alpina* gefunden, wobei die beiden Arten mit über 65 % übereinstimmten - inklusive sechs bekannter Motive. Basierend auf öffentlich zugängliche „Open Chromatin“ Daten wurde festgestellt, dass circa 30 % der DNA-Motive einen Schutz vor Endonuklease (DNase I) zeigen, was eine mögliche Rolle als Proteinbindungsstellen nahelegt. Mehrere zusammen auftretende und phasenspezifische DNA-Motiv-Paare wurden in beiden Arten gefunden, darunter bereits bekannte wie das „Evening Element“ und „ABA-Response-Element“ Paar, denen konservierte tagesrhythmische *cis*-regulierte Expression zugrunde liegt

Chapter 1

Introduction

1.1 Plant genome sequencing

The release of reference genome of *Arabidopsis* was a major milestone in plant biology (The Arabidopsis Genome Initiative, 2000) Being the first plant genome and the third multicellular genome to be sequenced after nematode *Caenorhabditis elegans* (The C. elegans Sequencing Consortium, 1998) and insect *Drosophila melanogaster* (Adams et al., 2000), it had a huge contribution towards beginning of plant genomics era. In *Arabidopsis*, functional and evolutionary studies became possible with the availability of high-quality reference genome (Koornneef & Meinke, 2010).

The recent introduction of high-throughput sequencing technologies dramatically reduced the difficulty, time, and cost associated with it. With these advancements in sequencing, plant community has witnessed a sharp rise in the successfully completed several plant genome projects, including fairly large and economically important plants such as rice ((Goff et al., 2002); (Yu et al., 2002.); (International Rice Genome, 2005), soybean (Schmutz et al., 2010), maize (Schnable et al., 2009), chickpea (Varshney et al., 2013) and wheat (Wheat Genome Sequencing Initiative, 2014).

1.2 Advancement in sequencing strategies

Sequencing for the *Arabidopsis* reference sequence was performed using (semi-automated) dideoxy Sanger sequencing ((Sanger et al., 1977); (Hunkapiller et al., 1991)). This method is extremely time consuming and expensive. In 2005, a new sequencing technology, sequencing by synthesis (SBS) developed by 454 Life Sciences made sequencing cheaper and quicker ((Ruparel et al., 2005); (Margulies et al., 2005)). In this method, DNA sequences are determined by synthesis or addition of nucleotides to the complementary DNA strand rather than chain-termination (as in dideoxy Sanger sequencing method). In 2006, Solexa Inc. (acquired by Illumina in early 2007) released Genome Analyzer® that was also based on SBS technology (Margulies et al., 2005). In the same year, Agencourt personal genomics (acquired by Applied Biosystems in 2007) launched SOLiD (Sequencing by Oligo Ligation Detection) sequencer, based on a polony technology (sequencing by ligation) (Shendure, 2005). This method involves the use of DNA ligase enzyme to determine the nucleotide at a given position in an oligonucleotide. These three sequencers were the most typical examples of new massively parallel sequencing (MPS) system of next-generation sequencing (NGS) technology. Simultaneous sequencing of spatially separated DNA templates in a massively parallel fashion facilitated quick sequencing (Shendure & Ji, 2008).

Sanger sequencing (1st generation method) and SBS (2nd generation method) both required prior *in vivo* amplification (molecular cloning) or *in vitro* (by polymerase chain reaction (PCR)), while 3rd generation sequencing methods (PacBio RS sequencer launched by Pacific Biosciences) have no such requirement of prior amplification as sequencing of single molecules is performed ((Eid et al., 2009); (Rothberg et al., 2011)). Apart from the amplification free approach, another striking feature of the 3rd generation sequencing methods was sequencing of longer reads (up to dozens of kilobases (kbs)) compared to 2nd generation sequencing platforms (up to 700 base pairs (bps)).

1.3 The impact of whole genome sequencing (WGS)

The decision of selecting a species for sequencing is taken after considering several criteria such as scientific or economic importance, the size of the research community, genome size, ploidy level, availability of genetic and physical maps, etc. A substantial fraction of sequenced plant genomes belong to crop species and have been sequenced for particular research purposes by large and active research communities ((Wheat Genome Sequencing Initiative, 2014); (Schnable et al., 2009); (Mayer et al., 2012); (Velasco et al., 2010); (Shulaev et al., 2011); (Slotte et al., 2013); (Willing et al., 2015)). For last few years, this trend is being challenged and projects like **Genome10K** (https://genome10k.soe.ucsc.edu/) launched with the objective of sequencing a genomic zoo, a collection of DNA sequences representing the genomes of 10,000 vertebrate species, roughly one from every vertebrate genus (Genome 10K Community of Scientists, 2009).

As more and more genomes are getting sequenced, novel biological aspects are getting elucidated, such as questions on adaptation or evolution ((Dassanayake et al., 2011); (Slotte et al., 2013); (Willing et al., 2015)). With addition of each new genome, our understanding of genome biology increases and sometimes, the previous hypotheses are refined or re-defined. For example, the genome of banana (*Musa acuminata*) and tomato

(*Solanum lycopersicum*) enhanced understanding of not only whole-genome duplications but also, its role in shaping the evolution of monocot and dicot plants ((D'Hont et al., 2012); (Sato et al., 2012)).

1.4 Genome re-sequencing and population-scale studies

Once the genome sequence for a species is available, it becomes possible to catalog sequence variations with associated biological consequences. This includes naturally occurring sequence variations and mutations introduced by random mutagenesis ((Page & Grossniklaus, 2002); (Østergaard & Yanofsky, 2004); (Schneeberger & Weigel, 2011)). Identification of these sequence differences is crucial to connect genotype to phenotype. Though, sequencing a genome has become quicker, inexpensive and less complicated, assembling a genome is still a challenging task ((Schatz et al., 2010); (Earl et al., 2011); (Salzberg et al., 2012)). Therefore, sequencing individual genomes and mapping sequencing data to get an estimate of sequence variation became popular (known as re-sequencing). Re-sequencing not only simplified sequence variation detection but also, made sequencing with lower depth possible. This is essential to accommodate multiple genomes within the same cost. This approach is either applied to whole genome context (whole genome re-sequencing) ((Ossowski et al., 2008); (Huang et al., 2009)) or to specific loci of interest (targeted enrichment re-sequencing) ((Gnirke et al., 2009); (Mamanova et al., 2010)).

Re-sequencing involves mapping of randomly fragmented millions of DNA pieces (typically around 100 bp long) back to reference sequence with high accuracy. When dealing with small pieces of DNA, distinguishing sequencing/assembly errors from real sequence variations is a non-trivial task. While alignment tools such as Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990), BLAST like Alignment tool (BLAT) (Kent, 2002) are fast and powerful but not specialized for mapping enormous amount of short read data. Several NGS read alignment algorithms introduced recently to counter these challenges. These methods mostly based on indexing either read or reference sequence to gain speed ((Lin et al., 2008); (Li et al., 2008b); (Jiang & Wong, 2008); (Schatz, 2009); (Li et al., 2008)). The read alignment to a single genome can give rise to "reference bias", a bias for discovering sequence variation in highly similar regions with reference sequence. Computation tools such as GenomeMapper (Schneeberger et al.,

2009) along with work by (Gan et al., 2011)) is worth mentioning in this context. In the year 2009, Schneeberger et al. demonstrated that simultaneous alignment of short read data against several genomes not only provide access to highly diverged regions that are difficult to map otherwise, but also limits alignment errors near Indels. In the same year, Gan et al. highlighted the importance of analyzing genomic data along with transcriptomics data to interpret functional consequences associated with sequence variation in “reference-free” manner. Recent advancements in alignment methods resulted in developing a reference-sequence-free approach to benefit non-model species ((Nordström et al., 2013); (Ratan et al., 2010); (Iqbal et al., 2012)). The possibility to re-sequencing genomes led to ambitious projects like the 1001 *Arabidopsis* genome project focusing on the population dynamics by sequencing hundreds of individual genomes ((Weigel & Mott, 2009); (Cao et al., 2011);).

1.5 Genome annotation: Coding and non-coding regions of a genome

1.5.1 Genome annotation: Coding regions of a genome

Once a high-quality genome sequence is available, next step is to annotate the gene models (referred to as coding region of the genome). Genome annotation broadly divided into two distinct phases. The first phase includes structural annotation of genome that involves, precise identification of sequence elements such as introns, exons, start codon, stop codon, etc. The second phase is functional annotation where the aim is to assign biological function to genomic elements.

Identification and masking repeat sequences are the initial steps of the genome annotation. Repeat masking is important to inform gene prediction tools to exclude these regions from gene prediction. Tools used for repeat identification either involves homology-based searches such as LTR_Finder (Xu & Wang, 2007), RepeatMasker (http://www.repeatmasker.org/webrepeatmaskerhelp.html), Censor (Kohany et al., 2006) etc. or *de novo* library-based search such as RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html), RepeatScout (Price, et al., 2005) etc. Frequently a combination of both approaches is used for repeat identification and masking.

Ab initio gene tools (predictors) offer a fast and comprehensive solution for screening the genome for potential protein coding regions ((Norton & York, 2003); (Brent, 2005)). Prediction is done for common features of a protein-coding gene such as start codon (ATG), stop codon (TAA, TGA, TAG), open reading frame (ORF), intron-exon boundaries and sometimes even polyadenylation sites (Salamov & Solovyev, 2000). Present-day *ab initio* predictors are mostly based on hidden Markov models (HMMs) or more complex and improved versions of it (e.g. GHMM) ((Burge & Karlin, 1997); (Lukashin & Borodovsky, 1998); (Korf, 2004); (Salamov & Solovyev, 2000); (Stanke & Waack, 2003)). However, these tools have practical limitations in predicting UTRs and alternative isoforms. Moreover, despite using sophisticated models for gene prediction, *ab initio* methods suffer from a high false positive rate (Norton & York, 2003). Once transcriptional and protein sequence evidences became available, gene prediction tools were updated to utilize experimental data for further improvement in accuracy (Augustus (Stanke & Waack, 2003), SNAP (Korf, 2004), FGENESH (Salamov & Solovyev, 2000)). Supplementing *ab initio* tool with (experimental) evidence data substantially improved its accuracy in filtering false positives. Traditionally, experimental data such as Expressed Sequence Tags (ESTs), cDNA or protein data are used to supplement prediction tool (Guigó et al., 2006). The newest addition to this list is RNA-sequencing (RNA-seq) data. RNA-seq data is commonly used in two ways for gene prediction (Garber, 2011). First, *de novo* assembly of RNA-seq reads, followed by transcripts mapping using long read mapping tools such as Velvet (Zerbino & Birney, 2008). Alternately, RNA-seq reads can be aligned to the genome sequence using split read aligners such as TopHat (Trapnell et al., 2009), followed by transcripts reconstructing tools like Cufflinks. **Figure 1** summarizes various evidences layers used to improve gene model (Robertson et al, 2010).

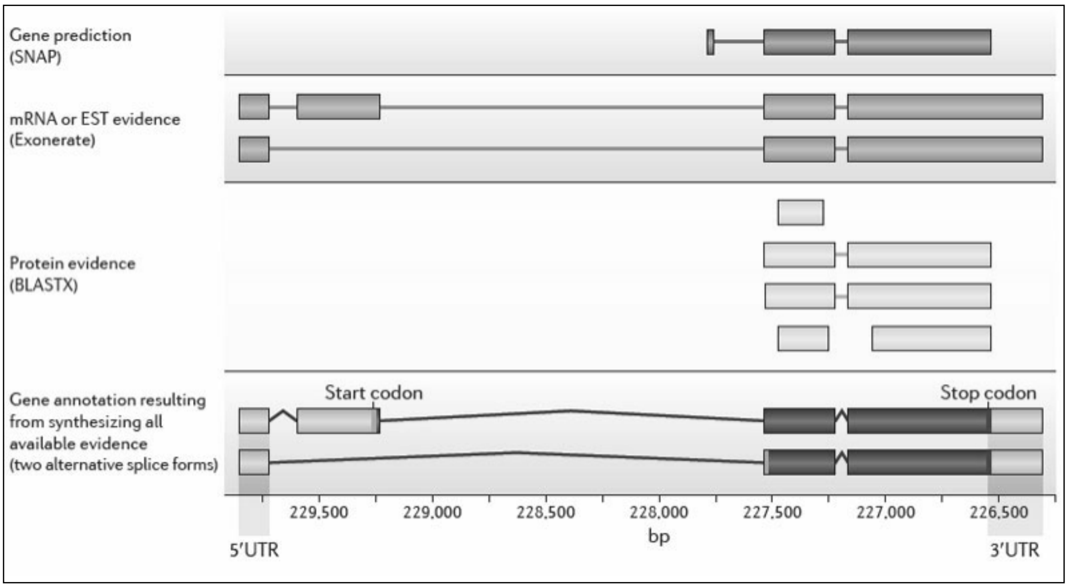

Gene prediction (SNAP)
mRNA or EST evidence (Exonerate)
Protein evidence (BLASTX)
Gene annotation resulting from synthesizing all available evidence (two alternative splice forms)
Start codon
Stop codon
229,500
229,000
228,500
228,000
227,500
227,000
226,500
bp
5′UTR
3′UTR

Figure 1 | Various layers in gene model annotation

Source: (Yandell & Ence, 2012)

In second phase, genes can be functionally annotated by employing several approaches including sequence similarity searches, protein-protein interactions and functional assignment based on protein 3D folding or gene expression profiles. The most straightforward way to infer gene function is based on sequence similarity, using database-searching programs such as BLAST (Altschul et al., 1990) or PSI-BLAST (Altschul et al., 1997). Once a high scoring match is found in database, functional annotation is transferred to the query sequence. The success of this approach highly depends upon completeness and accuracy of information present in the database(s). Several databases are available for functional annotation assignment: RefSeq NCBI database (Pruitt et al., 2012), UniProt/Swiss-Prot (Wu et al., 2006), Gene Ontology database (Ashburner et al., 2000), Pfam (Finn et al., 2006) etc.

Even for annotated genomes, additional data (e.g. additional RNA-seq data) can be used to improve existing genome annotations ((Li et al., 2011); (Eckalbar et al., 2013); (Darwish, Shahan, Liu, Slovin, & Alkharouf, 2015); (Rawat et al., 2015)). Next challenge is to make such improved information available with version information. In a comprehensive review, Steven L Salzberg underlined the need of setting “guidelines” or “community-wide accepted standards” for genome sequencing and annotation projects (Salzberg, 2007). His emphasis was on setting a wiki for updating genome annotation similar to what the *Arabidopsis* community successfully demonstrated through TAIR (https://www.arabidopsis.org/).

In chapter 2 of the thesis, I will present a study conducted to improve annotation of *Arabidopsis lyrata*, which was recently published in PLoS One (Rawat et al., 2015).

1.5.2 Genome annotation: Non-coding regions of a genome

Genome annotation efforts primarily focus on sequence with coding potential but a substantial fraction of the genome remains un-annotated. For instance, in gene dense *Arabidopsis* genome around 70% genome is not annotated (Lamesch et al., 2012). However, information about the regulation of genes is encoded in this, so-called non-coding regions which is now has become an integral part of genome annotations.

Unlike coding regions, identification of non-coding regions is not straightforward due to high variation in location, size, and sequence composition. To identify functional non-coding elements, promoter regions of co-expressed genes can be compared ((Aerts et al., 2003); (Aerts et al., 2004); (Sarkar & Maitra, 2008); (Wrzodek et al., 2010); (Gao et al., 2013)). Another common approach to identify functional non-coding regions is to use evolutionary constraints through comparative sequence analysis. One prerequisite for this approach is the availability of multiple closely related genome sequences. Using comparative sequence analysis, several attempts have been made to uncover conserved non-coding sequences (CNS) in yeast (Kellis et al, 2003), insect ((Stark et al., 2007); (Siepel et al., 2005)), worm ((Siepel et al., 2005)), vertebrates ((Siepel et al., 2005); (Bejerano et al., 2004); (Boyle et al., 2008)) and plants ((Hupalo & Kern, 2013); (Lyons et al., 2008)).

Brassicaceae is an economically important family of flowering plants, including the model plant *Arabidopsis*. Recent advances in sequencing technology and reduced cost of sequencing, facilitated full genome sequencing and assembly of several Brassicaceae genomes ((Wang et al., 2011); (Dassanayake et al., 2011); (Wu et al., 2012); (Slotte et al., 2013); (Haudry et al., 2013); (Kitashiba et al., 2014); (Lobréaux et al., 2014); (Willing et al., 2015)). With the availability of several sequenced and annotated genomes, comparatively small genome size (Johnston et al., 2005) and large research community contributing to functional information, Brassicaceae seems ideal for non-coding region analysis. In the year 2013, Haudry and coworkers conducted a well-designed study with nine Brassicaceae genomes (six previously sequenced species and three new species) to identify 90,000 conserved non-coding sequences (CNSs). In this study, they could show that in *Arabidopsis* around 4% of the genome (4.5 Mb) is evolving under selection constraint and resides close to transcription start site (TSS) wih no coding potential identified (Haudry et al., 2013). One more interesting finding of this study was about similar CNS regions in *Arabidopsis*, despite nearly 40% smaller genome than *Arabidopsis lyrata*. (Hu et al., 2011). Besides, being the first large-scale attempt to discover and quantify CNS in Brassicaceae this study also highlighted the importance and abundance of non-coding elements.

1.5.2.1 Transcription regulation by *cis* elements

In plants, transcriptional regulation of gene expression is mainly controlled at gene promoters through *cis*-acting elements. ((Meshi & Iwabuchi, 1995); (Singh, 1998); (Liu et al., 1999); (Kaufmann et al., 2010)). Transcriptional regulation by transcription factors (TFs) binding in the promoter region of a gene, is a widely explored mechanism (Wray et al., 2003).

Arabidopsis has around 6-10% of its genes coding for transcription factors, which underlines the importance and complexity of transcriptional regulation by TFs ((Riechmann et al., 2000); (Qu & Zhu, 2006)). Even if bound by the identical protein, TF binding sites (TFBSs or *cis*-regulatory elements (CREs)) are not identical ((Palaniswamy et al., 2006); (Priest et al., 2009)). To represent such complex sequence preference, all TFBSs are used and referred to as DNA motif. In plants, CREs are short (generally conserved) motifs of 5 - 20 nucleotides and usually found upstream of genes (Rombauts et al., 2003). However CREs have also been found downstream of the TSS, for instance, in the 1st intron of the gene itself ((Zhang et al., 2012); (Sieburth & Meyerowitz, 1997); (Sheldon et al., 2002); (Kooiker et al., 2005)). A single promoter is typically composed of many CREs allowing for different combinations of TFs to mediate different expression responses of a gene.

1.5.2.2 Identification of *cis*-elements in promoter sequence

The main goal of the *cis*-regulatory analysis is to locate and annotate CREs. This knowledge then, can be transferred to a broader context for better understanding of gene regulation. Various experimental and computational methods have been employed for identification of *cis*-elements.

Classical DNA footprinting experiment was one of the first attempts to identify regions in promoter bound by regulatory proteins (Galas & Schmitz, 1978). An extension of this approach is DNaseI-sequencing ((Crawford et al., 2004); (Boyle et al., 2008)). This method is used to discover the regulatory regions by sequencing of regions sensitive to DNase I cleavage in genome-wide manner. Analysis of DNase I hypersensitive sites (DHS) has revealed novel binding sites and has proven extremely useful for discovery of CREs ((Crawford et al., 2004); (Boyle et al., 2008)).

A more specific method, chromatin immunoprecipitation (ChIP), involves crosslinking of DNA to a specific (already known) DNA-binding protein followed by isolation step using a specific antibody. The DNA bound to the protein can then be identified, using microarray chips (ChIP-Chip) (Ren et al., 2000) or by direct sequencing (ChIP-seq) (Johnson et al., 2007). Resolution-wise ChIP-seq outperforms ChIP-Chip, but the precise binding location of transcription factor remains difficult to determine using either method. As an essential step, most of the studies with ChIP-seq experiment follow a computational motif finding step to pinpoint the precise binding locations and sequence preference. Although variability of most binding motifs and variable affinity make the motif finding challenging (Park, 2009). Due to several such efforts, binding location and sequence preferences of many *Arabidopsis* transcription factors are known including TGA, Hy5, PIL5, SEP3, SVP, FLC, PRR5, PRR7 and CCA1 ((Fonseca et al., 2010); (Lee et al., 2007); (Wu et al., 2012); (Kaufmann et al., 2009); (Gregis et al., 2013); (Deng et al., 2011); (Nakamichi et al., 2012); (Liu et al., 2013); (Nagel et al., 2015)). Moreover, recent efforts have been made to study changes in binding preferences of TFs when they work in combination (Mateos et al., 2015).

However, all these techniques have their limitations. DNase I-seq analysis works genome-wide, but the footprints do not provide information on the bound protein whereas, immune precipitation techniques require specific antibodies ((Tsompana & Buck, 2014); (Park, 2009)). Besides, for large-scale assay these techniques are expensive and sometimes lack properly defined controls (Park, 2009). Additionally, due to high complexity, it is sometimes difficult to implement these techniques in specific contexts.

Because of these limitations, computational approaches are a good alternative or supplement to CREs identification. Computational approaches are defined broadly into two groups; (i) identification of instances of known TFBS and (ii) *de novo* identification of unknown DNA motifs.

(i) The accuracy of identification largely depends on how (well) the motif is defined. One approach to define motifs with degenerated consensus sequences (using IUPAC representation) but this lacks information about the likelihood of observing alternate nucleotide on various sequence positions. Most common way to define a motif is a position weight matrix (PWM) or position-specific scoring matrix (PSSM) ((Stormo et al., 1982); (Stormo, 2000)). A PWM or PSSM is a 4 X N matrix (for DNA), where the four rows represent DNA bases (A, T, G, and C) and N is the length of TFBS. Elements of PWM reflect the likelihood of observing a particular nucleotide at that particular position, which is usually done after correction for a compositional bias of the background genome. Search for existing motif that is the identification of instances of the motif, seems fairly straightforward, but low complexity and degeneracy in the sequence of known motifs make searches prone to false positives. Recent developments in PWM matching approaches are majorly based on index-based algorithms. These approaches involve pre-processing of the target sequence(s) into index structure (mostly suffix tree) which then can be used for quick search of PWM match (Beckstette et al., 2006). Another approach is the online search approach where a simple sequential search is performed over the target sequence ((Liefooghe et al., 2009); (Salmela & Tarhio, 2007); (Korhonen et al., 2009)).

TFs are frequently expressed in several different tissues (or cell types) and still manage to coordinate tissue-specificity via different interacting co-regulators. Therefore, identification of single TF binding profiles, in isolation is not sufficient for deciphering complex transcriptional networks. Likewise, CREs are generally clustered into some relatively small stretches (a few hundred bps), forming *cis*-regulatory module (CRM). Computational approaches to identify relevant co-occurring TFBSs (potential CRM) have been developed to address this problem (Cister: (Frith et al., 2001.); ClusterBuster : (Frith et al., 2003) ; CisModule : (Zhou & Wong, 2004) ; ModuleSearcher : (Aerts et al., 2004); Clover: (Frith, 2004); ModuleDigger : (Sun et al., 2009);,COPS : (Ha et al., 2012)).

(ii) A frequently used approach for *de novo* discovery of (*cis*) regulatory elements is to find sequence elements with high occurrence in upstream regions of a set of genes regulated in the same way. Genes with similar functional annotation, co-expressed genes (in same species) or orthologous genes from closely related species are explored for *de novo* motif identification. Promoter sequence alignments of closely related species were quite successful in the discovery of regulatory elements, commonly known as phylogenetic footprinting ((Tagle et al., 1988);(Cliften et al., 2003)). Based on the same principle, phylogenetic shadowing ((Boffelli et al., 2003.), (Hong et al., 2003)) has later been developed to explore shorter (and more refined) regions of promoters of closely related species to reduce combinatorial complexity. Both approaches have limitations; co-expressed genes obtained from microarray or RNA-seq experiments represent steady-state mRNA levels and not necessarily provide co-regulated gene set. Moreover, wrong ortholog assignment, missing ortholog, ortholog diversification (species-specific duplication or evolution) and low (and in short stretches) conservation in promoters limit accuracy of these approaches.

There is a long list of tools available for *de novo* DNA motif identification (MEME: (Bailey & Elkan, 1994); AlignACE: (Roth et al., 1998); RSAT : (van Helden, André, & Collado-Vides, 1998); Weeder : (Pavesi, Mauri, & Pesole, 2001); RSAT suite : (Thomas-Chollier et al., 2008); CisFinder : (Sharov & Ko, 2009); rGADEM (R package) : (Mercier et al., 2011)). These tools are mostly based on two approaches; first is an alignment-free method, involving searches for k-mers (words) with specified number of mismatches overrepresented as compared to the background. The second group of methods is based on either expectation maximization (EM) for motif elicitation or

Gibbs search; MEME and AlignACE belong to this group ((Bailey & Elkan, 1994); (Lawrence et al., 1993)).

In thesis chapter 3, I present a study to discover DNA motifs enriched in the promoter of diurnally expressed gene in *Arabidopsis* and *Arabis alpina*.

1.5.2.3 Importance of *cis* elements in diurnal expression

All organisms experience the daily change in light and temperature due to rotation of the earth. Synchronized gene expression in response to these cyclic environmental changes is referred to as diurnal gene expression. Plants sense the time of the day and prepare themselves even before the light become unavailable and temperature drops at night. For this, they need to synchronize expression of genes according to time of a day. This contributes to rhythmic gene expression, which is believed to be driven through an extensive network of diurnal and clock-regulated transcription factors (TFs) and their corresponding CREs ((Dunlap, 1999); (Harmer et al., 2000); (McClung, 2006); (Harmer et al., 2009); (Greenham & Mcclung, 2015)).

1.5.2.4 *Cis* elements enriched in genes under diurnal/circadian regulation

One of the earliest report connecting *cis* elements with rhythmic gene expression reported an AT-rich oligonucleotide, AAAATATCT, commonly known as Evening Element (EE) in promoters of evening-phased co-expressed cyclic genes (Harmer et al., 2000). Subsequent experiments confirmed that its presence is enough to drive periodic evening-phased expression in genes. EE is bound by CCA1, a core component of the circadian clock. Later, several circadian clock-related (and diurnal) *cis*-regulatory elements have been discovered and described including another similar AT-rich element, CCA1 binding site (CBS), AAAAAATCT (Michael & McClung, 2002) HUD box (Hormone Up and Down box, G-box (CACGTG; (Giuliano et al., 1988)) etc. ((Hudson & Quail, 2003); (Covington et al., 2008)). Knowledge about circadian or diurnal CRE was enormously advanced with numerous time-course microarray-experiments, clustering of genes according to the expression peak, and subsequent CRE identification using enrichment analysis (Covington et al., 2008). Detailed analysis of phase-specific gene expression, revealed CREs specific to different phases of a day; Morning Element (AACCACGAAAAT) enriched in promoter of morning-phased genes, Telo-box (AAACCC) and protein box (ATGGGCC) enriched in mid-night phased genes, and GATA element was found to be enriched in afternoon and evening-phased genes ((Covington et al., 2008); (Covington et al., 2008)). Besides *Arabidopsis*, CREs are also found in the cycling transcriptome of a monocot (rice) and a dicot (poplar) species. All major classes of diurnal CREs, including morning (ME, GBOX), evening (EE, GATA) and midnight (PBX/TBX/SBX) elements were found in both species (Filichkin et al., 2011). This study provided evidence of conserved diurnal *cis* regulation between mono and dicotyledonous species (Filichkin et al., 2011).

Chapter 2

Improving annotation of *Arabidopsis lyrata* genome

"Strive for continuous improvement, instead of perfection."

Kim Collins (World champion sprinter, 2003)

Motivation and result summary

Arabidopsis lyrata is a close relative of *Arabidopsis* and frequently serves as an out-group in evolutionary studies. Additionally, it is also a model species for research on adaptation and molecular evolution. Even though its reference sequence is one of the few genome assemblies exclusively based on high-quality dideoxy sequencing data, its gene annotation was generated with limited RNA sequencing data. In the context of several on-going projects, we struggled with weaknesses of its current genome annotation.

Re-annotation of the genome using extensive RNA-seq data corrected the coordinates of around 90% gene models and introduced alternate isoforms for over 2,000 gene models. This updated annotation includes hundreds of previously wrongly splitted and merged gene models, some of which were experimentally validated. Based on the RNA-seq data derived from a heat stress experiment, I also describe, how the new annotation enables an advanced analysis of differentially expressed isoforms in *A. lyrata*.

Contents of this chapter are published in PLoS One, 2015 with the title "Improving annotation of *Arabidopsis lyrata* with RNA-seq data".

2.1 Introduction

A. lyrata is a predominantly self-incompatible, perennial plant species from the Brassicaceae family that diverged from *Arabidopsis* approximately 10 million years ago (Hu et al., 2011). Despite its evolutionary closeness, surprisingly its genome size is around one and a half times larger than *Arabidopsis* genome ((Johnston et al., 2005); (Lysak et al., 2009). Besides *Arabidopsis*, *A. lyrata* is the only species from Brassicaceae family with a reference assembly exclusively based on high-quality dideoxy sequencing. This 207 Mb *A. lyrata* reference assembly attributed the genome size difference to the accumulation of many small deletions in the *A. thaliana* genome, primarily in non-coding regions and transposable elements (TEs) (Hu et al., 2011). Moreover, *A. lyrata* has undergone recent genome expansion due to activity of transposable elements (TEs), in particular, Copia long terminal repeat (LTR) retrotransposons ((Hu et al., 2011); (Slotte et al., 2013); (Willing et al., 2015)) which is the basis for species-specific patterns in DNA methylation (Seymour et al., 2014).

With evolutionary closeness with *Arabidopsis* and fully assembled genome, *A. lyrata* serves as an important out-group for comparative evolutionary studies within *Arabidopsis* ((Schneeberger et al., 2011); (Cao et al., 2011); (Long et al., 2013)). Moreover, recent advances in sequencing technologies have also facilitated the full genome sequencing and assembly of an increasing number of Brassicaceae genomes and their close relatives ((Slotte et al., 2013); (Willing et al., 2015); (Wang et al., 2011); (Dassanayake et al., 2011); (Wu et al., 2012); (Haudry et al., 2013); (Kitashiba et al., 2014); (Lobréaux et al., 2014); (Liu et al., 2014)), which, projected Brassicaceae as a good candidate family for comparative genomics. Intra- as well as inter-species comparisons still heavily rely on the high-quality genome annotations. Nowadays, high-quality annotations have become essential even in the non-model species.

The current genome annotation of the *A. lyrata* describes 32,670 genes, which were predicted using a combination of *ab initio* gene prediction, homology to known proteins sequences and expression data from related species (Hu et al., 2011). Even though the gene models were analyzed for their expression support using RNA-seq data (38% gene models were supported by expression data), gene prediction methods integrating RNA-seq alignment information were not developed at the time genome annotation was generated. In a recent study, Haudry and colleagues supplemented the original annotation with additional putatively transcribed regions to study the conservation of non-coding sequences among related Brassicaceae species (Haudry et al., 2013). They integrated the results of additional *ab initio* gene predictions, RNA-seq data alignments, and homology searches against the genes of *Arabidopsis* to mask potentially un-annotated coding sequences and regions that recently lost coding potential due to mutations.

Building upon the major efforts of the initial annotation of *A. lyrata* genome (version-1 from hereon) I have updated the gene models using RNA-seq samples from different tissues under stress and wild-type (WT) conditions. Improved annotation ("version-2" hereon) has changed/updated the coordinates of 29,141 out of original 32,670 gene models, removed 1,286 and added 1,295 new models. This update corrected coding region of hundreds of gene models, which were wrongly merged or split in version-1 and also separated genes harboring annotated TE (in coding region).. Additionally, I have analyzed the transcriptional response of *A. lyrata* rosette tissue to heat stress to show the improved utility of version-2 for the identification of differential isoform usage and pre-mRNA splicing.

2.2 Results and Discussion

2.2.1 Improving *A. lyrata* genome annotation using RNA-seq data

In the collaboration with the laboratories of Dr. Ales Pecinka (Ahmed Abdelsamad, Björn Pietzenuk) and Prof. Detlef Weigel, (Danelle Seymor and Daniel Koenig), we sequenced the transcriptome of various *A. lyrata* aerial tissues, including whole rosettes, dissected shoot apices, complete inflorescences, along with vegetative rosettes exposed to cold and heat stress (**see Materials and Methods**). In total, 290.1 million, strand unspecific, single-end short reads were generated using Illumina sequencing technology after poly-A purification. Short reads were aligned to *A. lyrata* reference assembly (Hu et al., 2011) using Bowtie v2.1.0 (Langmead & Salzberg, 2012) and the splice junction mapper TopHat v2.0.9 (Trapnell et al., 2009) (**see Materials and Methods**). Approximately 75% (146.8 million) of the reads aligned uniquely and were used for further analyses (**Appendix I**). Over 10% of the reads aligned to putative intergenic regions with no potential coding region annotated, strongly indicated that some gene models might have been missed in the version-1 annotation. Visual inspection of these intergenic alignments revealed the expected patterns for spliced transcripts indicating instances of unidentified gene models and cases where transcription exceeded known gene boundaries (**Figure 2**).

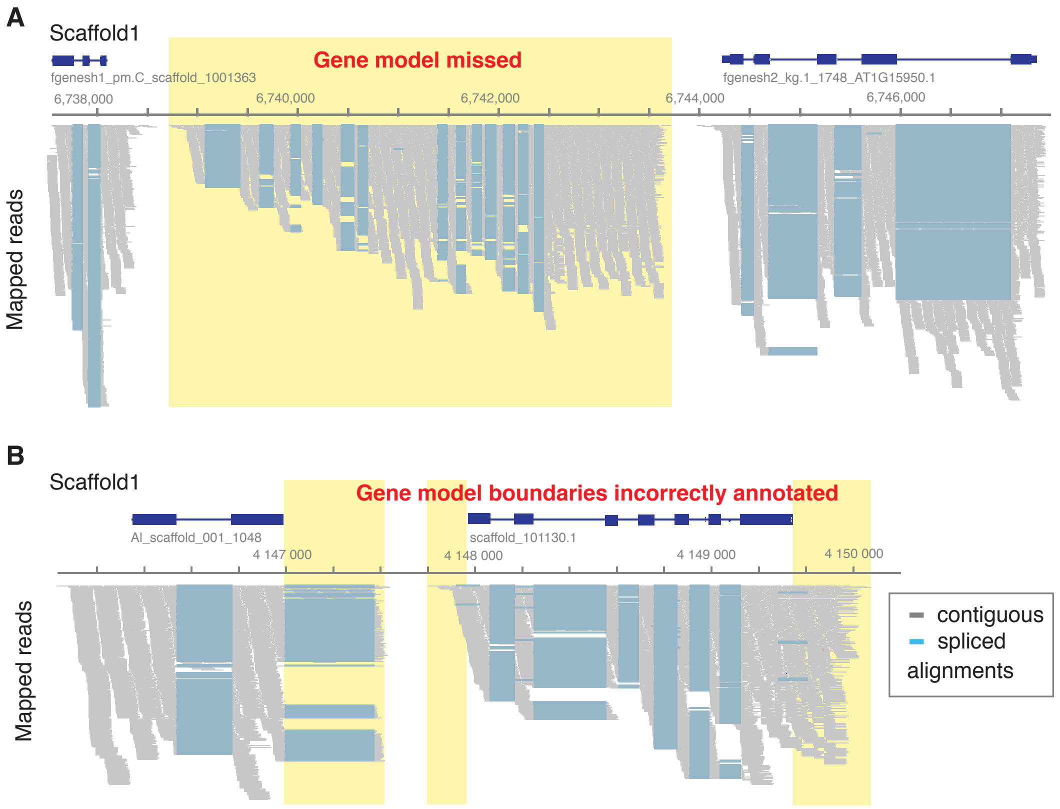


Figure 2 | Examples of incorrectly annotated gene models

(A) A gene model was entirely missing, but its locus shows clear evidence of transcription and splicing based on RNA-seq alignments. **(B)** The boundaries of two gene models do not include the full extent of the transcribed region. In the case of Al_scaffold_001_1048, an entire exon was missing in version-1.

New gene models were predicted from short read alignment data using Cufflinks 2.1.1 (Trapnell et al., 2010) independently for each tissue. In total, Cufflinks predicted 31,194 distinct gene models across all samples. An additional RNA-seq alignment-guided gene prediction using Augustus v.3.0.1 ((Stanke & Waack, 2003); (Stanke et al., 2006)) identified 40,728 gene models, including 27,830 genes, which were supported by at least five RNA-seq reads. Moreover, 30,483 and 30,837 of Augustus predicted gene models overlapped with version-1 and Cufflinks predictions, respectively (**see Materials and Methods**).

I combined 31,793 Augustus predicted gene models with evidence of transcription or with overlap with version-1 gene models to update the *A. lyrata* gene annotation (**Figure 3**). To ensure that I was not excluding any true gene models in version-1, I included 1,430 version-1 gene models with no overlap to any of the new gene models, but showed either evidence of expression or featured an ortholog in at least one of the Brassicaceae species *Arabidopsis* (The Arabidopsis Genome Initiative, 2000)*, Capsella rubella* (*C. rubella*) (Slotte et al., 2013), *Brassica rapa* (*B. rapa*) (Wang et al., 2011), *Schrenkiella parvula* (*S. parvula*) (Dassanayake et al., 2011) and *Arabis alpina* (*A. alpina)* (Willing et al., 2015) (**Figure 4A**). This step increased the number of gene models to 33,223 (**see Materials and Methods**). To identify and correct cases where incorrect gene models might have been introduced into the version-2 annotation, I utilized the very close phylogenetic relationship between *A. lyrata*, *Arabidopsis* and *C. rubella*. I compared all gene models that were considerably different between version-1 and version-2 to *Arabidopsis* and *C. rubella* orthologs (**see Materials and Methods**). If the length of the version-1 open reading frame was closer to that of the orthologs, I retained the version-1 gene model. This resulted in 548 version-2 gene models being replaced with 688 of the original version-1 gene models (**Figure 4B**). After additional step for removal of redundant gene models, I obtained a final set of 33,221 non-redundant gene models.

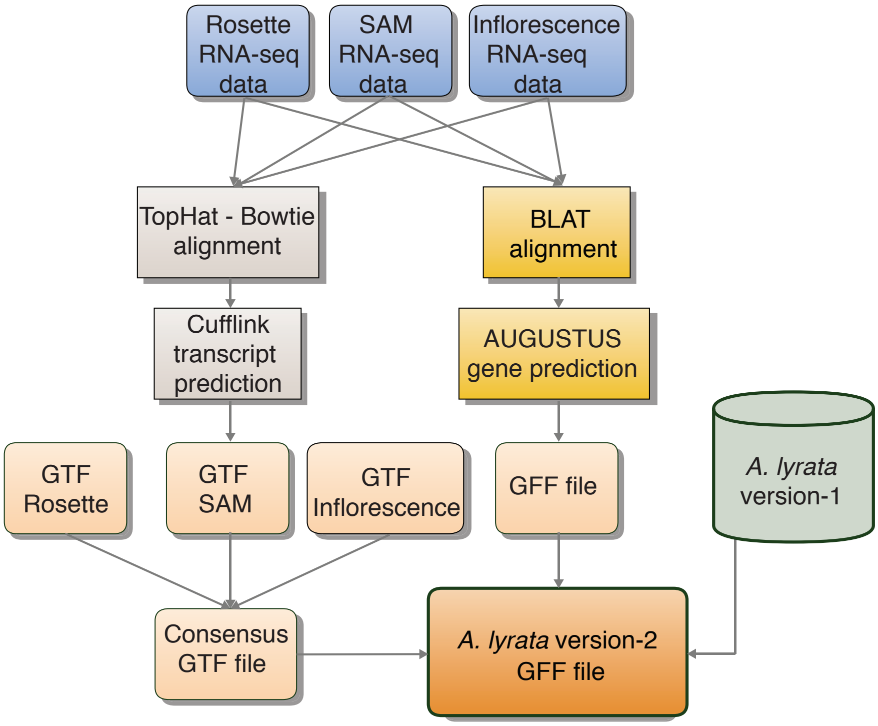


Figure 3 | Workflow for RNA-seq incorporation for version-2 annotation

Based on a recent annotation of *A. lyrata* TEs (Haudry et al., 2013) and sequence similarity to TE genes of *Arabidopsis* (Lamesch et al., 2012), I annotated 2,089 gene models as TE protein coding genes (**see Materials and Methods**). Without these, version-2 comprised of 31,132 gene models, which is ~13% more than genes count in *Arabidopsis* (Lamesch et al., 2012). Although, transfer RNA (tRNA) genes were described in the original analysis of the *A. lyrata* genome (Hu et al., 2011), version-1 lacks information regarding these loci. By rerunning tRNAScan (Lowe & Eddy, 1997), I identified 660 tRNA genes coding for all 20 amino acids. For completeness, I also incorporated 170 recently published micro RNA (miRNA) genes into the version-2 annotation file (Fahlgren et al., 2010).

Altogether, I updated the coordinates of 29,141 of the original gene models, removed 1,286 entire (mostly short) gene models, and added 1,295 new models (**Figure 4C**). Only 2,243 remained unaltered (including 688 version-1 gene models re-introduced due to their superior similarity to orthologs). The new annotation accounted for 31,132 non-TE-related gene models including 27,084 multi-exonic genes of which 2,236 featured at least one alternate isoform (**Table 1**). I also annotated 25,584 protein sequences (nearly 76% of total predicted transcripts) for known conserved protein domain using InterProScan software ((Quevillon et al., 2005); (Hunter et al., 2009)).

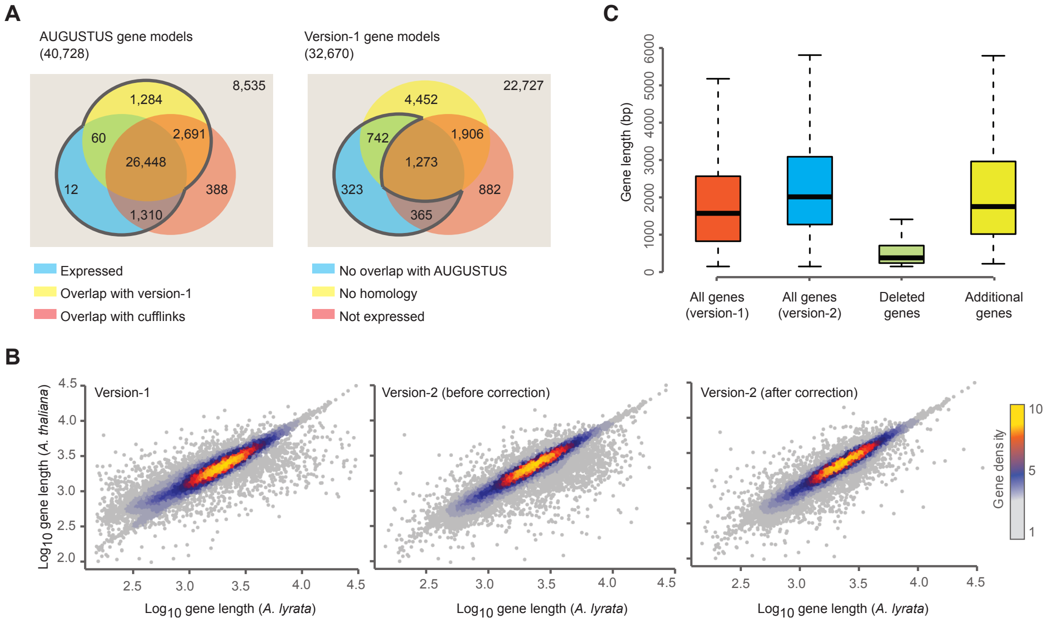


Figure 4 | Overview of gene model annotation and gene length comparison

(A) Left, version-2 gene models predicted by Augustus. Number of gene models overlapping with version-1 (yellow), genes predicted with Cufflinks (red), and genes with expression evidence (blue). Right, gene models of the version-1 annotation. Number of models without overlap to version-2 models (yellow), without orthologs in five other Brassicaceae (red), and without significant expression evidence (blue). **(B)** Correlation of the lengths of *A. lyrata* gene models with the length of their orthologous gene models in *Arabidopsis*. Left, *A. lyrata* version-1 gene models. Correlations using version-1 gene models (left), version-2 gene models before (middle) and after (right) the homology-based correction of gene models. **(C)** Length distribution of gene models including genes that were removed or newly added in the version-2.

Table 1 | Comparison of version-1 and varsion-2 annotation

	# version-1	**# version-2**
Gene models	32,670	33,221
Predicted transcripts	32,670	35,805
Protein-coding genes	32,670	31,221
TE-coding genes	-	2,089
miRNA genes	-	170
tRNA genes	-	660
Featuring ortholog (in at least one Brassicaceae)	23,996	24,146

2.2.2 Validating differences in gene model structure

Even after the above-mentioned homology-based gene length adjustments, I found cases where the corresponding gene models from the two annotations varied drastically in length. This included instances where multiple version-1 gene models were fused to form a single gene model in version-2 or vice versa (**Figure 5A and 5B**). In total, 161 version-1 genes were split (accounting for 530 genes in version-2) and 1,729 version-1 gene models were merged (accounting for 775 gene models in version-2).

I (in collaboration with Ales Pecinka and Ahmed Abdelsamad) randomly selected 14 version-1 gene models that had been split into multiple gene models in version-2 and 14 gene models that had been merged in version-2 for PCR validation (**Figure 6 and 7**). Ales Pecinka and Ahmed Abdelsamad performed PCR validation experiments with computational support in primer designing from me. For three merge cases, amplification of genomic DNA (gDNA) failed for primer validation and could not be confirmed. This was most likely due to large gDNA amplicon size (2.4 – 5 kbp) that rendered the results of these cases inconclusive. For all 24 remaining cases, PCR results fully confirmed the annotation of the new gene models.

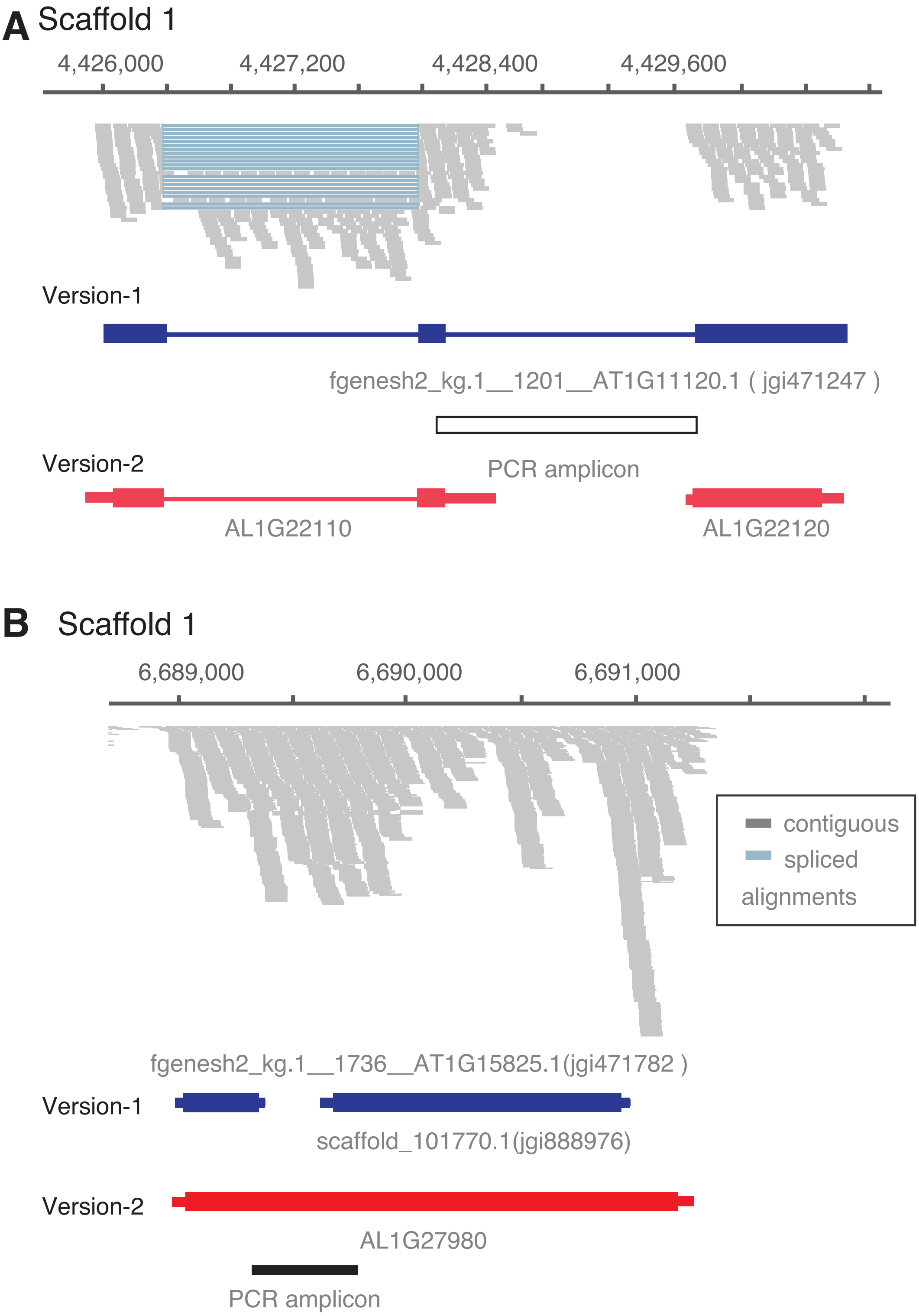


Figure 5 | Examples of wrong split and wrong merge cases in version-1 annotation

(A) Example of a gene model that was split into two gene models in version-2. Reverse transcription-PCR could not confirm the connection of both. **(B)** Example of version-1 gene models that were merged during the annotation update. Reverse transcription-PCR confirmed presence of a transcript bridging the two version-1 genes.

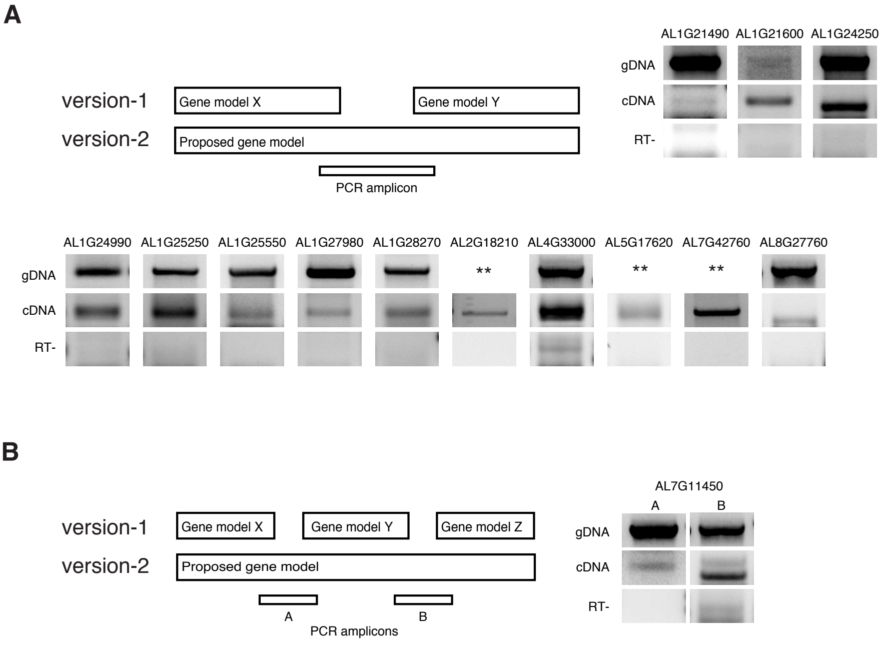


Figure 6 | Experimental validation of merged gene models in version-2 annotation

(A) Schematic drawing of two gene models in version-1 that were merged into one in version-2. PCR primers were designed to span regions predicted as intergenic in version-1. gDNA, genomic DNA; cDNA, complementary DNA, RT-, reverse transcription reaction without reverse transcriptase. ** indicate cases where gDNA reaction did not work, most likely due to large amplicon size (2.4 – 5 kb). (B) Validation of version-2 gene model combining three version-1 gene models. The principle follows the description in **(A),** except that both junctions are validated **(A and B)**.

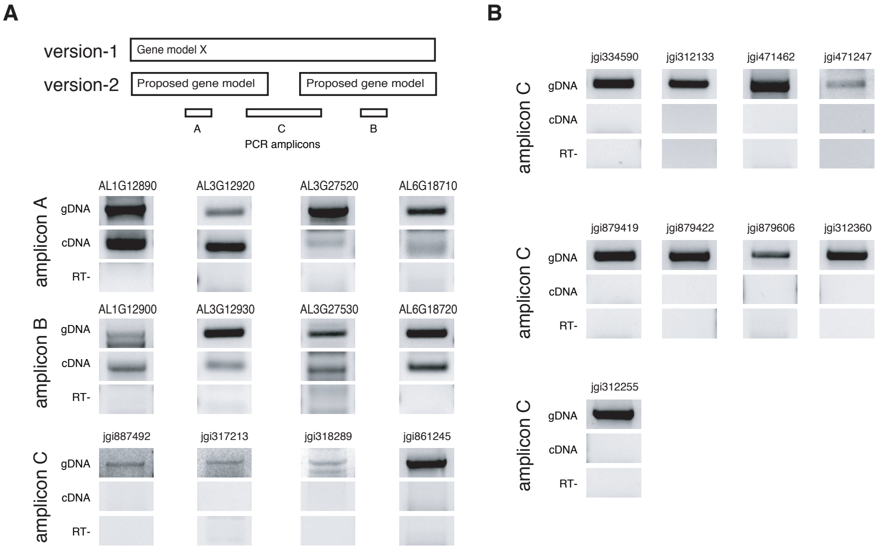


Figure 7 | Experimental validation of splitted gene models in version-2 annotation

(A) The scheme shows version-1 gene models that were split in version-2. PCR amplicons A and B were designed to target cDNA sequences common to both annotations, while amplicon C spanned a region predicted as intergenic in version-2. gDNA, genomic DNA; cDNA, complementary DNA, RT-, reverse transcription reaction without reverse transcriptase. **(B)** Additional cases tested using strategy described in **(A)** where only amplicon C was tested.

2.2.3 Comparison of version-2 annotation with other Brassicaceae

For both *A. lyrata* annotations, I predicted orthologous relationships between *A. lyrata* and five other Brassicaceae species (**see Materials and Methods**). Using version-2 gene models, 77.5% of the genes predicted with an ortholog in at least one species (24,146 out of 31,132) compared to 73% for version-1 (23,996 out of 32,670) (**Figure 8A and 8B**). The number of genes with predicted orthologs in all five Brassicaceae was also slightly higher for version-2 with 15,105 genes versus 14,850 genes with version-1. The removal of many short gene models in version-2 changed the distribution of gene model lengths (**Figure 9**). Version-1 has an excess of gene models shorter than 1 kb with a second peak around 1.5 kb, which describes a bimodal distribution that was only reflected by gene length distribution of *B. rapa*. In contrast, version-2 had only a single mode around 1.7 kb, similar to the four other species. The length distribution of predicted protein sequences in version-1 was distinct from the other Brassicaceae species, and this discrepancy largely disappeared with version-2. A third factor that contributed to the length differences between the genes of version-1 and version-2 were differences in UTR annotations (**Figure 9**). In version-1 33% of the genes were annotated without UTR information, however, in version-2 only 5% remained without 3’ and 5’ UTR annotation. The absolute and relative contributions of individual features are shown in **Figure 10**. Though, absolute increase in genomic space for all gene features was observed but CDS and UTRs benefited the most. I also observed little decrease in intronic genome space, which can be explained by introduction of splice variants previously missing from version-1 annotation.

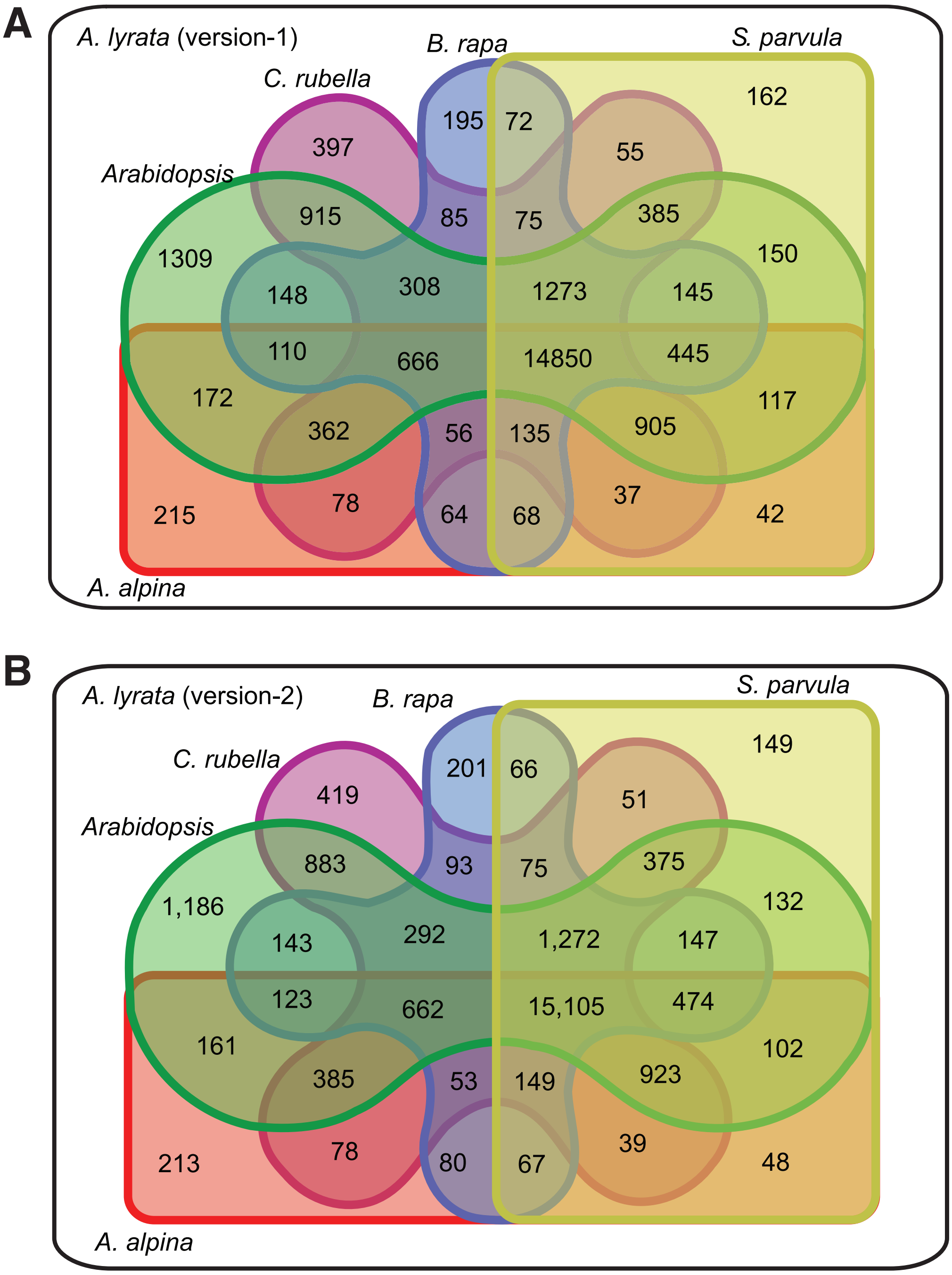


Figure 8 | Comparison of identified orthologs in five Brassicaceae

Orthologous gene models shared between **(A)** *A. lyrata* version-1, *Arabidopsis*, *A. alpina*, *B. rapa*, *C. rubella* and *S. parvula*. **(B)** Similar analysis with *A. lyrata* version-2.

The removal of many short gene models in version-2 changed the distribution of gene model lengths (**Figure 9**). Version-1 has an excess of gene models shorter than 1 kb with a second mode around 1.5 kb. *B. rapa* is the only other Brassicaceae with such a bimodal distribution. Version-2 had only a single mode around 1.7 kb, similar to the four other species. The length distribution of predicted protein sequences in version-1 had also been distinct from the other Brassicaceae species, and this discrepancy largely disappeared with version-2.

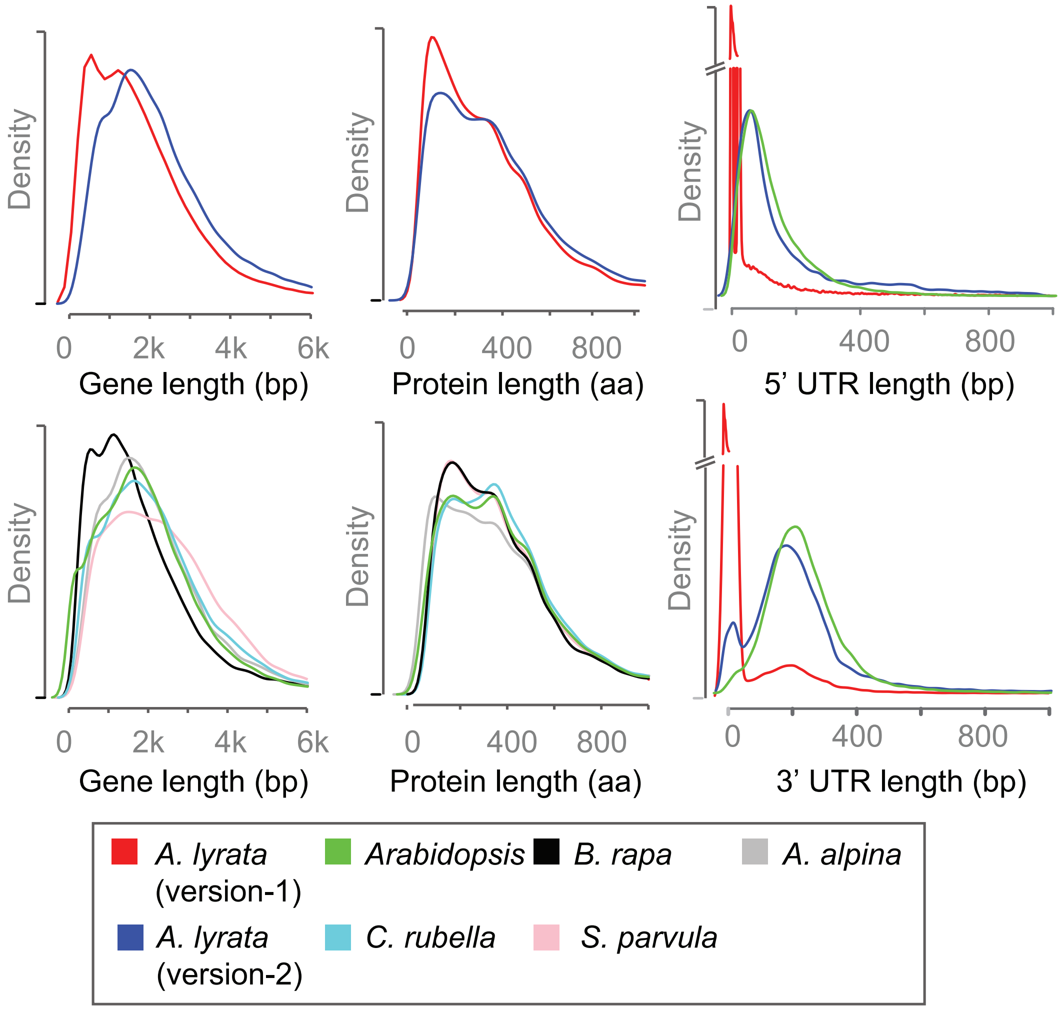


Figure 9 | Comparison of gene features within five Brassicaceae

Gene length, protein length and UTR length distributions of five Brassicaceae species including **version-1 and version-2** *A. lyrata* annotations.

Whether the bimodal distribution in *B. rapa* reflects similar ambiguity in gene annotations, or mirrors particular characteristics of *B. rapa*, including its ancient genome triplication and subsequent fractionation, is not known.

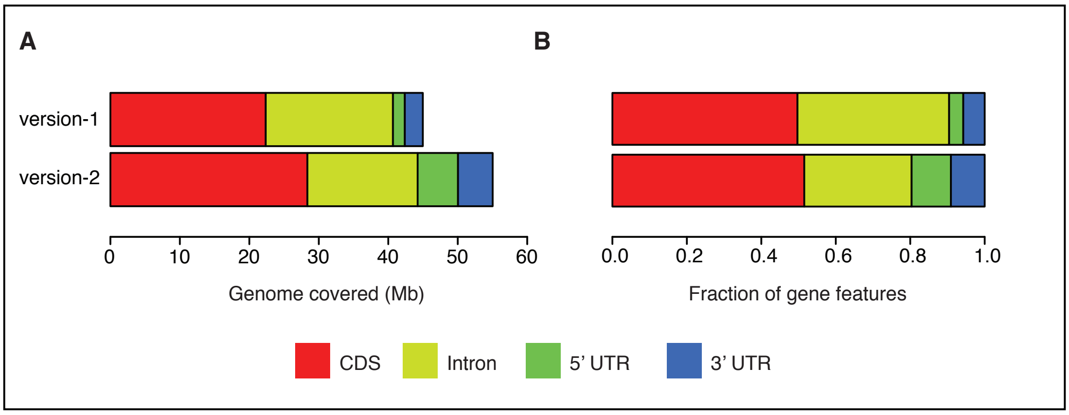


Figure 10 | Genic space comparison between version-1 and version-2 annotation

(A) Absolute genome coverage and (B) Fraction of genome coverage

2.2.4 Enhanced usability of version-2 annotation: Alternate splicing events

The availability of multiple isoforms from individual gene models in version-2 enables quantitative expression comparisons between annotated isoforms. I analyzed RNA-seq data from *A. lyrata* rosette tissues from untreated (WT), heat stressed (HS), and recovered (REC) samples in duplicate (**see Materials and Methods**). I also analyzed the data for differential gene expression using Cuffdiff v.2.0.2 (Trapnell et al., 2010). WT and REC samples resulted in 3,114 and 2,962 differentially expressed genes when compared to HS samples, whereas only 106 genes differentially expressed between WT and REC. This indicates, as expected, a strong effect of heat stress on gene expression (**see Materials and Methods**). Cuffdiff was also used to estimate differential expression between isoforms. I identified differential isoform expression of 283, 15 and 119 genes when comparing WT with HS, WT with REC, and HS with REC, respectively. In contrast, as version-1 does not include different isoforms, which is a prerequisite for isoform expression analysis as implemented in Cuffdiff, it was not possible to run this analysis using version-1.

I investigated differential splicing using a second tool, MATS v3.0.8 (Shen et al., 2012), which does not rely on prior isoform annotations and only identifies differences in individual splicing events. With version-2, MATS identified 177, 0 and 130 differential splicing events distributed over 187 distinct gene models in the three comparisons (**Figure 11**; **see Materials and Methods**). MATS reported only 99, 1 and 67 events affecting 103 gene models using version-1. The overlap of different splicing events was very high (95 common out of 103 (version-1) and 187 (version-2) gene models). Thus, almost all gene models with differential splicing events predicted based on version-1 were also predicted using version-2, however, the results based on version-2 revealed many more gene models. This was partially due to newly added genes (10 cases), but the most significant improvement came from the updates to exon-intron boundaries of existing gene models indicating that the new gene annotation improved the overall usability of this resource.

The isoform-dependent (Cuffdiff) and -independent (MATS) analyses identified only 37 common gene models. Even though Cuffdiff revealed fewer events as compared to the MATS analysis, it did identify 100 genes with differential isoform usage that were not included in the set of genes with multiple isoforms. This suggests that differential isoform expression analysis profits from prior isoform annotation. However, should not only rely on existing isoforms.

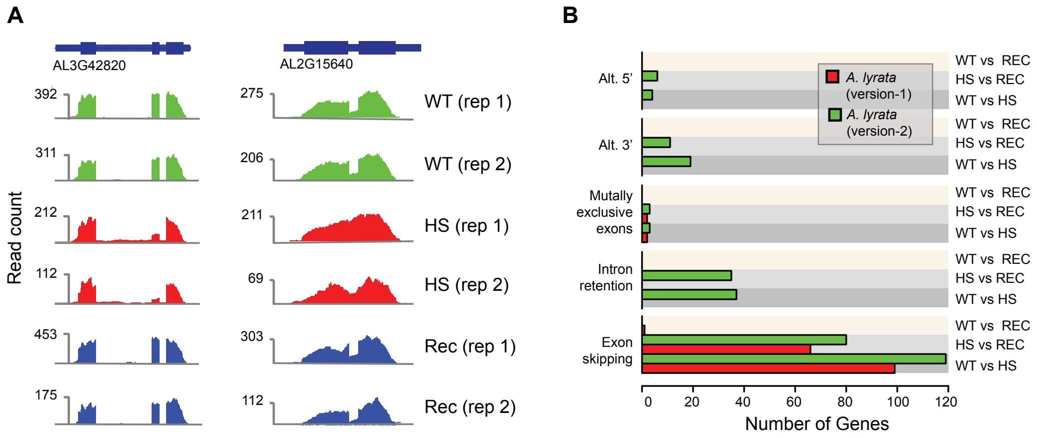


Figure 11 | Heat stress induced alternate splicing events

(A) Examples of differentially expressed isoforms in response to heat stress in *A. lyrata*. AL3G42820 expresses a second isoform that lacks the middle exon in heat-treated samples (HS). Transcripts from wild-type (WT) and recovery (REC) samples contain all three exons. AL2G15640 retains an intron in response to heat stress (HS) while wild-type (WT) and recovery (REC) samples show partial intron splicing. **(B)** A number of differential splicing events, including alternate 5' and 3' splice sites, mutually exclusive exons, intron retention, and exon skipping events identified with MATS based on version-1 and version-2 annotations.

2.3 Conclusions

The updated annotation includes 31,132 gene models with 35,805 transcripts. I also reported 1,304 gene models that were erroneously split or merged in the previous annotation. Validation of these models strongly supported our updates highlighting the importance of employing species-specific RNA-seq data for annotating genomes.

I also provided the first annotation of alternate splicing events in *A. lyrata*. Using RNA-seq samples for a heat stress experiment. This study demonstrated the improved utility of the version-2 annotation for differential isoform expression studies. This revised genome annotation advances the reference sequence of *A. lyrata* as a community resource for comparative and functional studies.

2.4 Materials and Methods

2.4.1 Plant material

A. lyrata subsp. *lyrata* MN47 plants were grown in soil under long day conditions (16 hours light, 21°C: 8 hours dark, 16°C). Vegetative rosettes and dissected shoot apices of three-week-old plants and entire inflorescences of flowering plants were harvested as mock-treated samples. For heat stress and recovery treatments, three-week old plants were incubated at 37°C for 6 hours or for an additional 48 hours at 21°C, respectively. Cold stressed samples were treated as described (Seymour et al., 2014). Laboratories of Dr. Ales Pecinka and Prof. Detlef Weigel generated plant material, used in this project.

2.4.2 Nucleic acid isolation and RNA-seq library preparation

DNA was isolated using Nucleon Phytopure kit (GE Healthcare). For total RNA isolation, samples were flash frozen in liquid nitrogen and used with Qiagen RNeasy® Plant Mini Kit, including an on-column DNase I digestion. Total RNA integrity was confirmed on the Agilent BioAnalyzer. Barcoded libraries were constructed using the Illumina TruSeq RNA kit with average of 1 μg of total RNA as starting material. The manufacturer's protocol was precisely followed with one exception in the cold-treated samples where 12 PCR cycles were used instead of the recommended 15. The library quality was monitored on a Bioanalyzer 2100 (Agilent) and the libraries were sequenced as 100-bp single end reads using Illumina sequencing. Pecinka and Weigel labs performed Nucleic acid isolation experiments.

2.4.3 RNA-seq read mapping and gene prediction

RNA-seq data was mapped to the *A. lyrata* reference genome assembly (Hu et al., 2011) using Bowtie v1.0.0 (Langmead & Salzberg, 2012) and TopHat v2.0.10 (Trapnell et al., 2009). Cufflinks v2.0.2 (Trapnell et al., 2010) was used for *de novo* transcript identification in all tissues separately. Cuffmerge (from the Cufflinks suite) was used to merge transcript annotation files obtained for three tissues separately. In addition, all short reads were aligned to the reference assembly of *A. lyrata* using BLAT v.34 (Kent, 2002) to generate an evidence file for guided gene prediction using Augustus v3.0.0. *A. lyrata* specific configuration file was generated using the version-1 annotation. To estimate agreement between Augustus and version-1 gene models, gene models with >=30% overlap (in respect to the shorter gene model) were considered. Gene models supported by five or more RNA-seq reads were considered as expressed irrespective of gene length.

To identify cases where wrong gene models were introduced in version-2, I first compared version-2 proteins (23,181 comparable proteins) with corresponding version-1 proteins. A total of 1,037 proteins were identified as outliers, where protein length difference was outside the range of +/-1 standard deviation of the distribution of length differences. For these cases version-1 and version-2 protein sequences were further compared against the proteins of their orthologs in *Arabidopsis* (Arabidopsis Initiative, 2000) and *C. rubella* (Slotte et al., 2013). If both orthologs were more similar in length to the protein of version-1, the respective version-2 gene model was replaced with version-1.

2.4.4 Ortholog identification and InterProScan annotation

Orthologous gene identification for both version-1 and version-2 was done separately at protein level using reciprocal best hits using blastall v2.2.25 (Altschul et al., 1990) and an e-value cutoff 0.001 among five Brassicaceae species. Conserved domain between proteins sequences were identified with InterProScan software (using E-value cutoff < 0.0001).

2.4.5 Identification of TE genes in version-2

Version-2 gene models harboring complete TEs (Haudry et al., 2013) within their coding regions or were entirely spanned by a TE were annotated as "TE coding genes". In addition 3,909 *Arabidopsis* TE genes (Lamesch et al., 2012) and TIGR Brassicaceae specific repeat database (Ouyang & Buell, 2004) were used to identify TE genes using blastn v2.2.25 (Altschul et al., 1990).

2.4.6 cDNA preparation and PCR

Plants were grown in soil under long day conditions until the five-leaf stage reached after approximately three weeks. cDNA samples were prepared from 1 μg total RNA of mock-treated rosettes using RevertAid First Strand cDNA Synthesis Kit with oligo d(T) primers (Thermo Scientific). Reverse transcriptase minus samples was processed in the same way without enzyme addition. PCR reactions were done in an Eppendorf thermal cycler using a standard program and the products were visualized on agarose gels stained with ethidium bromide. The PCR primer sequences can be found in Appendix II**.**

2.4.7 Differential gene expression and alternate splicing

Cufflinks (Trapnell et al., 2010) was used to calculate differential gene expression level (FPKM) with p-value < 0.01 and log2-fold change difference of more than 2. MATS (Shen et al., 2012) was used to investigate differential splicing events with over 0.01% splicing difference at a p-value < 0.01 and a false discovery rate of less than 1%. To control false positives, genes with 10,000 fold or more expression difference were excluded.

Chapter 3

Uncovering an atlas of diurnal DNA motifs (DDMs) using Phylogenetic shadowing in Brassicaceae genomes

"When all the details fit in perfectly, something is probably wrong with the story."

Charles Baxter, American novelist

Motivation and result summary

Arabidopsis and *Arabis alpina* (*Arabis* hereon) are members of Brassicaceae family, which diverged around 25 to 40 million years ago (Willing et al., 2015). With the availability of recently sequenced Brassicaceae genomes and in-house generated time series transcriptomics data, I attempted to identify and compare genes with daily cyclic (or diurnal) expression. I could also reveal an atlas of conserved TFBS enriched in the regulatory regions of co-expressed diurnal genes.

I identified in total ~50 DNA motifs (for both species), including all known and several novel DNA motifs presumably with a potential role in regulating diurnal gene expression. Moreover, most of these DNA motifs were enriched in different phases of the day. As expected, a high fraction of similar DNA motifs in *Arabidopsis* and *Arabis* were identified. Using publically available DNase I data and rescreening of motif, I could show that sites of several motifs show reduced activity of DNase I, providing hints for these motifs being true protein binding sites. Comparative enrichment analyses revealed that motifs pairs might not always be conserved in enrichment profile. As several motif pairs showed shifted enrichment profiles. Further analyses revealed several combinations of motifs being significantly enriched in promoters of diurnal genes, including previously known interacting motifs like Evening Element (EE: AAAATATCT; (Harmer et al., 2000)) and ABA response element like (ABREL: ACGTG; (Giuliano et al., 1988)); (Mikkelsen & Thomashow, 2009); (Berns et al., 2014)).

The work presented in this chapter is a collaborative effort of many people. Markus Berns, Loren Castaings, Nora Bujdoso and Julieta Mateos did RNA isolation. Eva Maria Willing analyzed RNA-seq data for identification of diurnal genes. Eva and I did comparative analysis of diurnal expression. *Cis* element identification pipeline was designed and implemented by me.

3.1 Introduction

Rhythmic (cycling) gene expression in response to day/night is referred to as diurnal expression. Primarily, light and the internal circadian clock regulate diurnal gene expression ((Dunlap, 1999); (Borland & Taybi, 2004); (McClung, 2006); (McClung, 2001)). Several studies using microarray assays in *Arabidopsis,* enabled the large-scale discovery of transcripts under control of circadian clock ((Harmer et al., 2000); (Schaffer et al., 2001); (Michael & McClung, 2002); (Blasing, 2005); (Mockler et al., 2007); (Covington & Harmer, 2007)). Diurnal control of gene expression has also been reported in several other plant species such as rice, poplar, soybean, papaya, etc. ((Filichkin et al., 2011); (Marcolino-Gomes et al., 2014); (Nose & Watanabe, 2014)). Diurnal gene expression drove through an extensive network of diurnal and clock-regulated transcription factors (TFs) and their corresponding *cis*-regulatory elements (CREs) ((Hsu & Harmer, 2014); (Greenham & Mcclung, 2015)). It is surprising despite the availability of diverse diurnal expression data sets to supplement our understanding of diurnal gene expression regulation, only a few *cis* elements are known till date.

In the year 2011, Filichkin and co-workers compared cycling transcriptome of distantly related plant species from monocot (rice) and dicot (poplar and *Arabidopsis*) and confirmed the early origin of such regulations. Only a handful *cis*-regulatory elements (CREs) are known in context of diurnal expression such as Morning Element (ME, CCACAC), G-Box (CACGTG); Evening Element (EE, AAATATCT), GATA Box (GATA), Telo Box (TBX, AAACCCT), Starch synthesis box (SBX, AAGCCC) and Protein Box (PBX, ATGCCC) and all were found conserved among these species. Although comparing distantly related species gives an idea about conserved regulatory mechanisms but reduced power for orthologous gene identification introduce an undesired level of uncertainty in such analyses. For instance, only 605 cyclic genes with expressed orthologs for *Arabidopsis*-rice-poplar gene pairs were compared (Filichkin et al., 2011). However, comparison among relatively closely related species might unravel some more insights along with a higher number of conserved *cis*-elements. Though, sensitivity can be a problem to such analysis but that can be overcome by including several species with a wide spectrum of conservation. Conducting such study in Brassicaceae has additional advantages; with several assembled and annotated genomes, well-defined phylogeny, and a higher fraction of the functional gene present for *Arabidopsis* therefor*,* specificity could be much better for such studies.

I (with other collaborators) conducted a well-designed study to compare diurnal regulation in two Brassicaceae species*.* A time-series RNA-seq data was generated for two consecutive days to identify diurnally expressed genes in both species. These genes were further analyzed to find *cis*-regulatory elements (CREs) using a powerful conservation-based approach called Phylogenetic shadowing. Phylogenetic shadowing analysis was conducted for both species along with six other Brassicaceae species to define regulatory region precisely.

The main objective of this project was to discover a larger set of CREs (Atlas of diurnal DNA motifs) for diurnal genes by employing Phylogenetic shadowing on recently available set of closely related plant species.

3.2 Results and Discussion

3.2.1 Time series transcriptomics data analysis for *Arabidopsis* and *Arabis*

Colleagues from our department, Markus Berns, Loren Castaings, Nora Bujdoso and Julieta Mateos, isolated RNA in a time-course experiment for *Arabidopsis* and *Arabis.* RNA was sampled from whole seedlings in 4 hours time-intervals for continuous 48 hours. in long day condition (**see Materials and Methods**). In total, 618 and 760 million single-end, strand unspecific 97 bp Illumina RNA-seq reads were generated for *Arabidopsis* and *Arabis*, respectively Short reads were aligned to *Arabidopsis* (Lamesch et al., 2012) and *Arabis* (Willing et al., 2015) genome assemblies using Bowtie v2.2.1 (Langmead & Salzberg, 2012) and the splice junction mapper TopHat v2.0.10 (Trapnell et al., 2009) (**see Materials and Methods**). Overall, 75% of the *Arabidopsis* and 83% of the *Arabis* short reads could be mapped to the respective reference assemblies **(Appendix III)**. High similarity in gene expression profile was observed between samples collected over two days but at the same time points for both species (**see Materials and Methods**). Allowing the use of second day data as biological replicate (**Figure 12**).

3.2.2 Identification of diurnally expressed genes

Using diurnal transcriptome data, Eva-Maria Willing, a colleague in my group, identified a set of diurnally expressed genes for both species. Genes were categorized into six different categories on the basis of expression and correlation of expression between two days **(Table 2, see Materials and Methods).** Further classification of these genes into diurnal and non-diurnal genes in low confidence and high confidence group was performed. The strict but reliable definition of diurnal expression revealed, 7,702 genes in *Arabidopsis* and 8,517 genes in *Arabis* as high confidence diurnal genes **(Table 3, see Materials and Methods)**. This estimate of diurnal genes number in *Arabidopsis*, was on boundary line of previous estimates, where 30-50% genes were estimated to show rhythmic expression using microarray and enhancer trap technologies under long day with photo cyclic conditions ((Blasing, 2005); ((Michael & McClung, 2003); (Michael et al., 2008)).

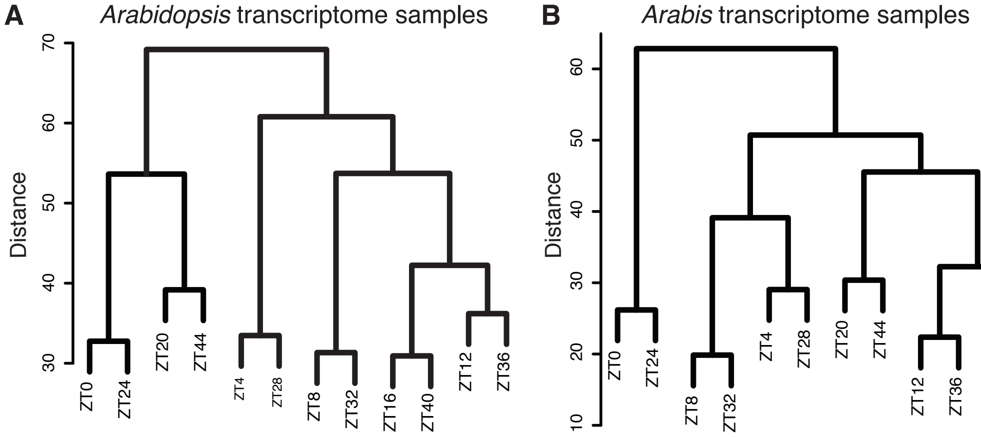


Figure 12 | Similarity analysis of samples collected at the same time over two days.

(A) Hierarchical clustering of time-series transcriptome samples for *Arabidopsis* (based on 23,169 expressed genes with Fragments Per Kilobase of transcript per Million mapped reads (FPKM >1). **(B)** Analogous analysis for 23,542 *Arabis* genes.

Table 2 | Classification of all genes in six categories

Categories	***Arabidopsis***	***Arabis***
A. Genes with no expression (FPKM = 0, for all time points)	3,305	2853
B. Genes with low expression (Max. FPKM < 3, for all time points)	995	8,455
C. Genes with low expression fold change (Max /Min < 1.5)	1,018	5,672
D. Genes with low (Pearson) correlation between expression profile of two days (r < 0.8)	14,289	5,202
E. Genes with ambiguous expression peak	907	419
F. Genes with unambiguous expression peak	6,805	8,109

Table 3 | Diurnal gene classification

Categories	***Arabidopsis***	***Arabis***
High confidence diurnal genes (r >= 0.8, p-value < 0.05)	7,702	8,517
Low confidence diurnal genes (r >= 0.5, fc > 1.1, p-value < 0.05)	1,873	2,782
High confidence non diurnal genes (r < 0.3)	6,182	7,129
Low confidence non diurnal genes (remaining expressed genes)	8,257	9,429
Low confidence diurnal (un expressed) genes (FPKM =0, all time points)	3,305	2,853
Total	**27,319**	**30,710**

3.2.3 Conservation of diurnal expression

I further checked whether the diurnal expression is conserved between the two species. Orthologous gene identification between both species was done using InParanoid ((Remm, Storm, & Sonnhammer, 2001); (Ostlund et al., 2010)). Ortholog(s) could be assigned to 5,422 genes in *Arabidopsis* (around 70% of 7,702 high confidence (hc) diurnal genes) and to 5,752 genes in *Arabis* (around 67% of 8,517 hc diurnal genes). Out of all hc diurnal genes (with assigned orthologs), 3,357 (62%) genes in *Arabidopsis* and 3,143 (55%) in *Arabis* had diurnal orthologs. This evidenced the tendency for conserved diurnal expression (p-value < 2.2e-16, Binomial test for both species); however, surprisingly a large portion of diurnal genes did not feature cyclic expression in other species. For details, 8% of *Arabidopsis* and 10% of *Arabis* diurnal genes featured orthologs with high confidence non-diurnal expression. This observation suggested that despite the trend for conservation of diurnal expression, some genes even gained/lost diurnal expression (**Figure 13; see Materials and Methods and Appendix IV**).

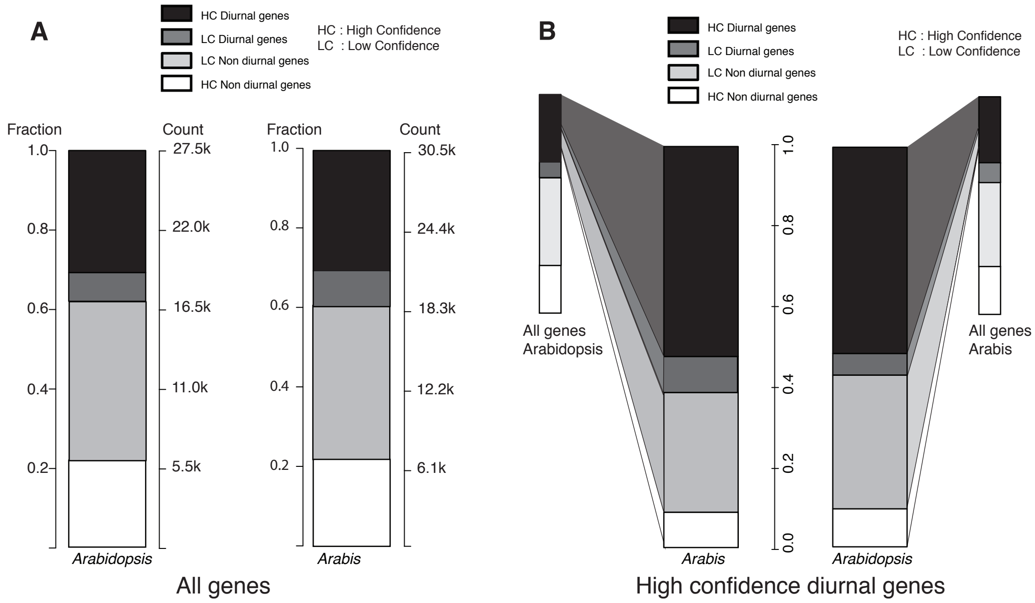


Figure 13 | Conservation of diurnal expression

(A) Bar plot showing fraction of high confidence (hc) diurnal, low confidence (lc) diurnal, lc non-diurnal, and hc non-diurnal genes in *Arabidopsis* and *Arabis*. **(B)** Fraction of genes with hc and lc conserved/non-conserved diurnal expression between species.

3.2.4 Peak expression time (phase) based clustering of diurnal genes

Cyclic (diurnal) oscillation in gene expression is shown to resemble sinus curves and can be identified by fitting sine functions to the gene expression data ((Harmer et al., 2000); (Straume, 2004); (Segal et al., 2003); (Panda et al., 2002); (Zieker et al., 2010)). Diurnal transcriptome samples were collected with 4 hours time difference. To improve the resolution, sinus curves with 1-hour phase resolution were simulated and compared against expression pattern of each gene. Correlation between simulated sinus curves and actual gene expression pattern (with six-time points) was estimated by the Pearson correlation coefficient (r) and a p-value for each curve fitting was reported. Around 90% of high confidence genes could be assigned a time point with p-value < 0.05 (**Figure 14, see Materials and Methods**). Principal component analysis (PCA) for expression data of diurnal genes revealed a non-discrete distribution of phase that was well recapitulated in the discrete one-hour resolution phase clustering.

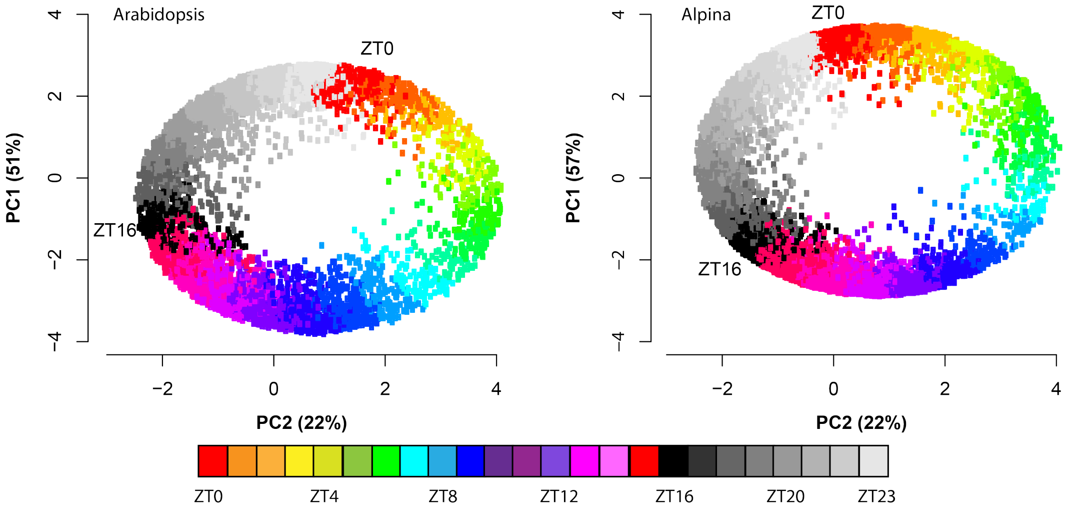


Figure 14 | Principal component analysis of diurnal genes with imposed phase assignments

(A) Distribution of diurnal genes on principal components (PC1 and PC2) and time point assignment (represented by different colors) in *Arabidopsis* **(B)** Analogous analysis was done in *Arabis*.

A very conservative strategy was used for the detection of diurnal genes because flexible modeling of periodic curves bears the danger of over-fitting. To verify the quality and robustness of phase assignment, a Model-based Periodicity Screening (MoPS) tool (Eser & Tresch, 2014) was applied additionally. Apart from mean, amplitude, and peak time, MoPS fits the shape of the curve and a periodicity score. When screening the peak time in 1hour intervals by MoPS, the estimated peak times agreed 100% with our dataset of periodic genes (hc diurnal genes) demonstrating that peak time can be estimated very accurately and consistently across methods (**Appendix V**).

Phase distribution of diurnal genes in *Arabidopsis* was bimodal. Two peaks were observed, before dusk (**Circadian Time point cluster 14 (CT14); Figure 15A**) and before dawn (**CT22; Figure 15A**). Similar observations have been reported previously for *Arabidopsis* (Michael et al., 2008), rice and popular (Filichkin et al., 2011) and recently in tomato (Müller et al., 2015). A similar bimodal distribution was also observed in *Arabis* (**Figure 15A; blue colored bars**). Despite the similarity in overall shape of the distribution, both peaks of the bimodal distribution in *Arabis* appeared to be shifted by two hours relative to *Arabidopsis* (**Figure 15A**) (Cross-Correlation (CC) > 0.5, **see Materials and Methods**). Comparing bimodal phase distribution of diurnal genes with diurnal orthologs (for both species) made this shift even more prominent with an average shift of 2 hours (CC > 0.5, **see Materials and Methods**). However, phase shift of individual genes could be drastically different for some cases, much more than the observed overall shift (**Figure 15B and 15C**). In rice and poplar a higher fraction of diurnal genes ant orthologs with similar phases were observed with phase around dusk (Filichkin et al., 2011). A similar but weaker treads was observed for *Arabidopsis* diurnal gene with but for *Arabis* diurnal genes no such tread was observed (**Figure 15B**).

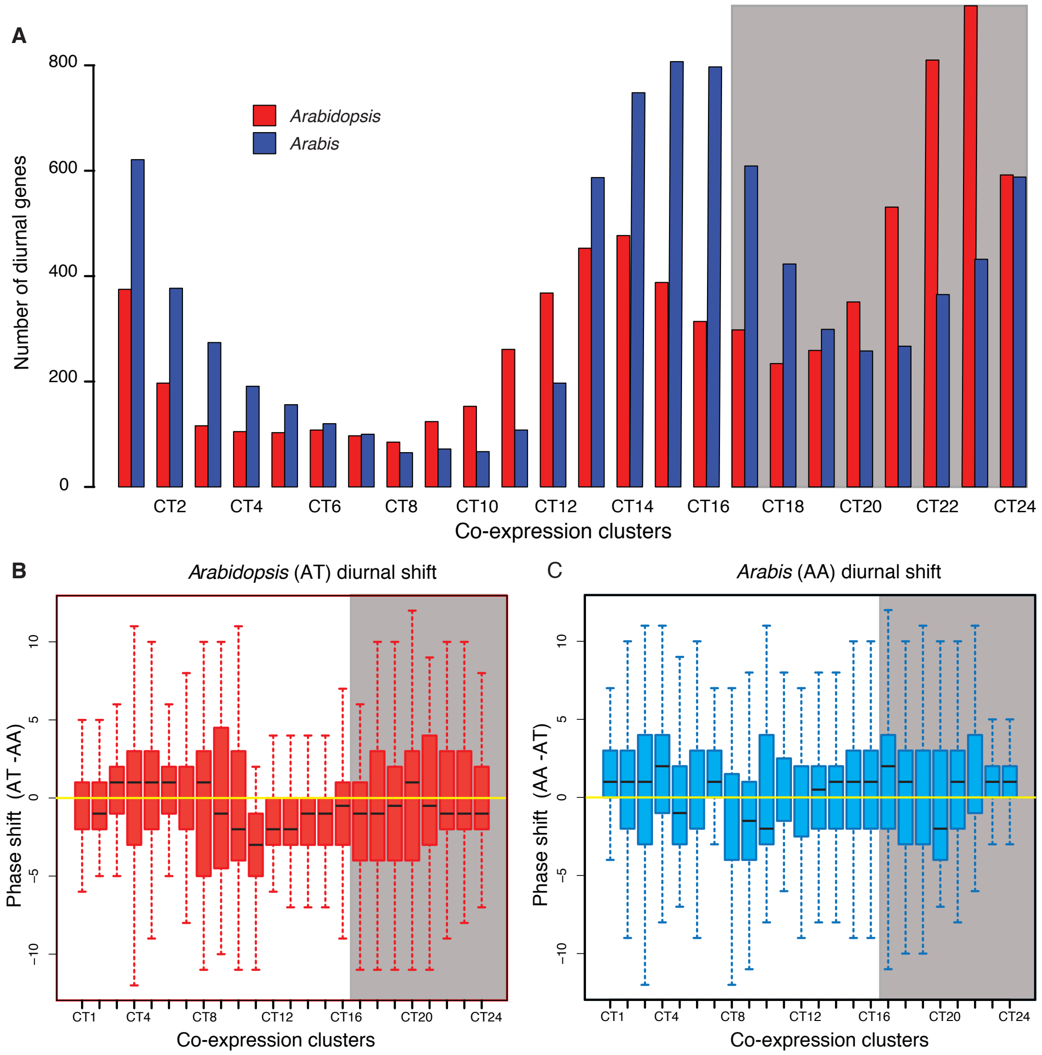


Figure 15 | Bimodal distribution of predicted phase and phase difference of orthologs of diurnal genes in *Arabidopsis* and *Arabis*

(A) Phase distribution of all high confidence diurnal genes in *Arabidopsis* (7,702) and *Arabis* (8,517**). (B)** For each co-expression cluster in *Arabidopsis*, the distribution of phase differences between diurnal genes and their orthologs are shown. **(C)** For each co-expression cluster in *Arabis*, the distribution of phase differences between diurnal genes and their orthologs are shown.

3.2.5 Comparing expression profile of clock genes

To explore if circadian clock underlies this phase shift of diurnal genes, I investigated the expression profile of key clock genes in both species. All clock genes described in this section were present in single copy gene in both species and could be reliable compared. In *Arabidopsis*, the core clock involves LATE ELONGATED HYPOCOTYL (LHY) and CIRCADIAN CLOCK ASSOCIATED 1 (CCA1), which are MYB-type transcription factors with peak expression around dawn. LHY and CCA1 inhibit expression of TIMING OF CAB expression 1 (TOC1), forming the core loop (**Figure 17**). The expression pattern of the *Arabidopsis* core loop genes and their orthologs in *Arabis* were highly similar (Avg. Phase shift = 0, **see Materials and Methods**) (**Figure 17**). *Arabidopsis* morning loop consists of PSEUDO-RESPONSE REGULATORS 5, 7, and 9 (PRR5, PRR7, PRR9). Surprisingly, I found a significant phase shift in all morning loop genes for *Arabis* as compared to *Arabidopsis* (Phase difference = 2 hours, **see Materials and Methods**; **Figure 17**). Likewise, evening loop genes, namely EARLY FLOWERING 3, 4 (ELF3, ELF4) and LUX ARRHYTHMO (LUX), I also found a significant consistent phase shift between the orthologous pairs, (Phase difference = 2 hours, **see Materials and Methods**) (**Table 4**).

Table 4 | Phase information of *Arabidopsis* clock genes and orthologs in *Arabis*

Clock gene	***Arabidopsis* gene**	**Phase**	**Orthologous gene in *Arabis***	**Phase**	**Involves in**
CCA1	AT2G46830	1	Aa_G177850	2	Core loop
LHY	AT1G01060	1	Aa_G211810	1	Core loop
PRR9	AT2G46790	5	Aa_G638260	7	Morning loop
PRR7	AT5G02810	7	Aa_G260760	9	Morning loop
PRR5	AT5G24470	10	Aa_G45380	12	Morning loop
LUX	AT3G46640	12	Aa_G656930	14	Evening loop
ELF4	AT2G40080	13	Aa_G224250	15	Evening loop
TOC1	AT5G61380	14	Aa_G448830	14	Core loop
ELF3	AT2G25930	17	Aa_G661860	20	Evening loop

3.2.6 Comparing expression profile of target genes of clock TFs

To investigate whether direct targets of clock TF genes are also shifted in *Arabis*, I analyzed publically available ChIP-seq data for some of the clock genes (CCA1, LHY, PRR5, PRR7 and TOC1) ((Nagel et al., 2015); http://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-52175/ ; (Liu,Carlsson, Takeuchi, Newton, & Farré, 2013); (Gendron et al., 2012)). The phase distribution comparison of the direct target in *Arabidopsis* with orthologs in *Arabis* also revealed phase shift in *Arabis* relative to *Arabidopsis,* for some TFs tested (**Figure 16A, 16B, 16C, 17D, and 17E**) (**see Materials and Methods**). As expected, direct targets of clock genes with non-shifted phase (CCA1 and LHY; **Figure 16A and 16B**) confirmed similar phase distribution of direct targets. The clock genes with shifted phase (PRR5 and PRR7; **Figure 16D and 16E**) displayed a clear shift of target genes in *Arabis.* This observation regarding shifted targets of clock TFs can partly explain shifted phase distribution of diurnal genes in *Arabis* as compared to *Arabidopsis*. Although gene expression pattern of TOC1 displayed a clear phase shift, but shift in phase distribution of target genes was not clearly visible (**Figure 17C**).

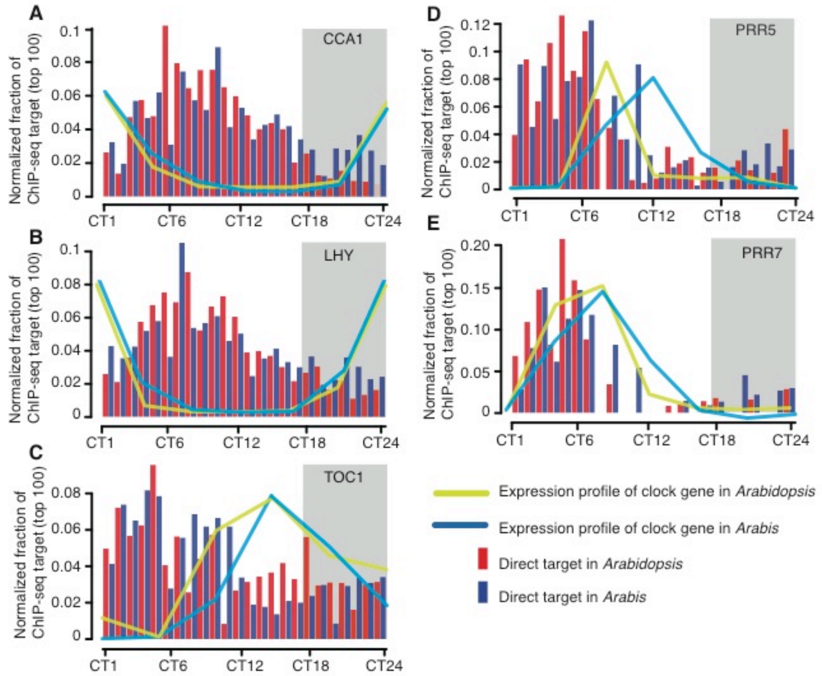


Figure 16 | Delayed expression of clock gene targets in *Arabis* compared to *Arabidopsis*

(A) Phase distribution of CCA1 targets **(B)** Phase distribution of LHY targets genes **(C)** Phase distribution of TOC1 targets genes **(D)** Phase distribution of PRR7 targets genes **(E)** Phase distribution of PRR5 targets genes

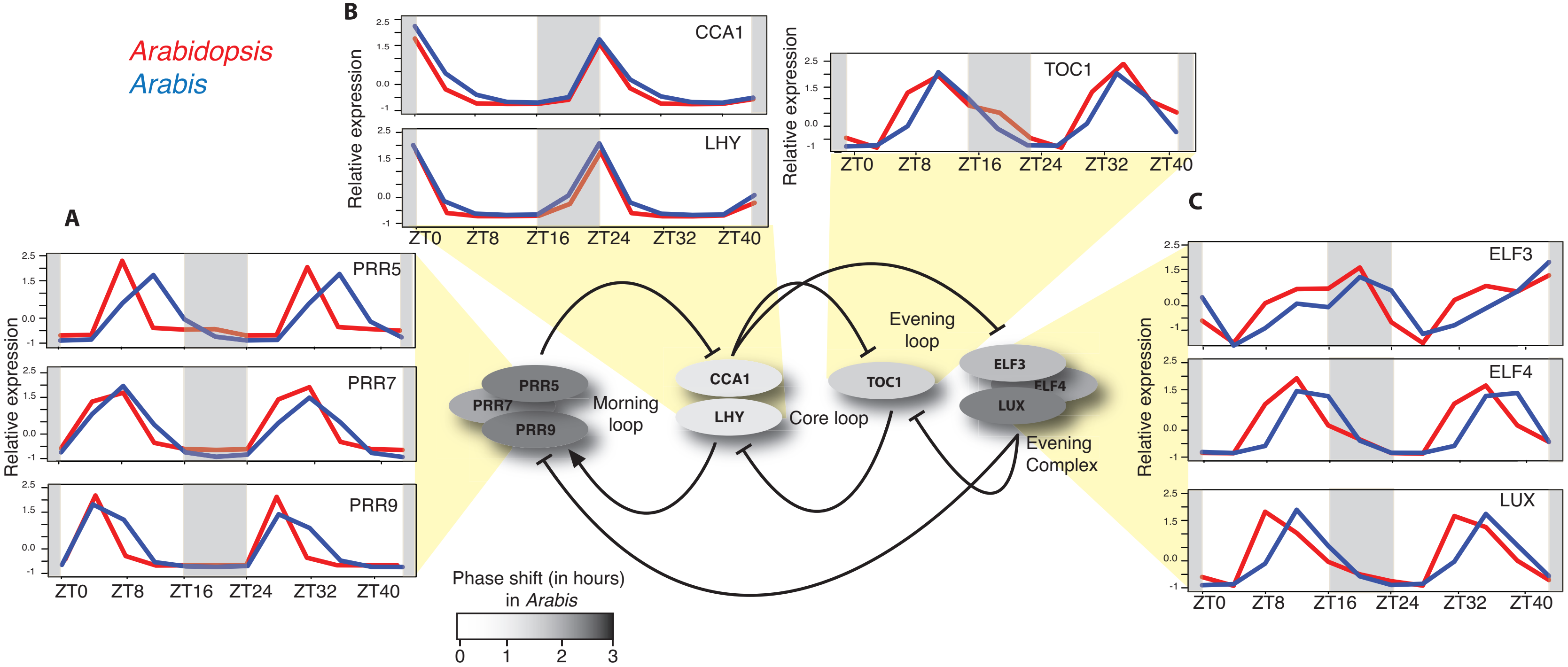


Figure 17 | Expression profile of clock genes in *Arabidopsis* and *Arabis*

(A) Morning loop, **(B)** Core loop **(C)** Evening loop **(modified from *Staiger et al, 2013*)**.

3.2.7 Leaf movement analysis

To test if the global phase shift in transcriptome could be observed on physiological level, we conducted leaf-movement-capture experiment for 40 *Arabis* and 57 *Arabidopsis* seedling. The rationale behind designing this experiment was to test functioning of *Arabis* clock as compared to *Arabidopsis* in absence of environmental cues. Leaf movement data was used as the physiological read-out of the clock, independent to transcriptional data. Clock entrainment was done for long day condition (similar condition for which transcriptional data was collected), and then leaf movement was measured in constant light condition (**see Materials and Methods**). Leaf-movement-capture experiment was performed under constant conditions to make sure that no external cue is affecting clock output, and we get real estimates (as much as possible) of the circadian clock functioning. Experiments under constant conditions are important to get an estimate of the clock functioning with no guidance from the external environment, but diurnal conditions are natural to plants with additional layers of regulations to fine-tune the clock ((Graf et al., 2010); (Haydon et al, 2013); (Fowler et al, 2005)). Leaf movement data analysis recapitulated the shift of the transcriptomic data with significant difference in phase between both species (t-test, p-value < 0.001) (**Figure 18**). The observed significant phase diffence in *Arabis* clock functioning suggested it could be possible that clock introduce these observed differences between transcriptome of two species.

3.2.8 Functional analysis of diurnal genes

To elucidate biological processes under control of diurnal expression, Gene Ontology (GO) terms were analyzed using AgriGO tool (Du et al., 2010) for all diurnal genes in *Arabidopsis* and *Arabis* separately. Due to high overlap in diurnally expressed gene*,* common GO categories (**biological process**) were found enriched (**Table 5 and Table 6**).

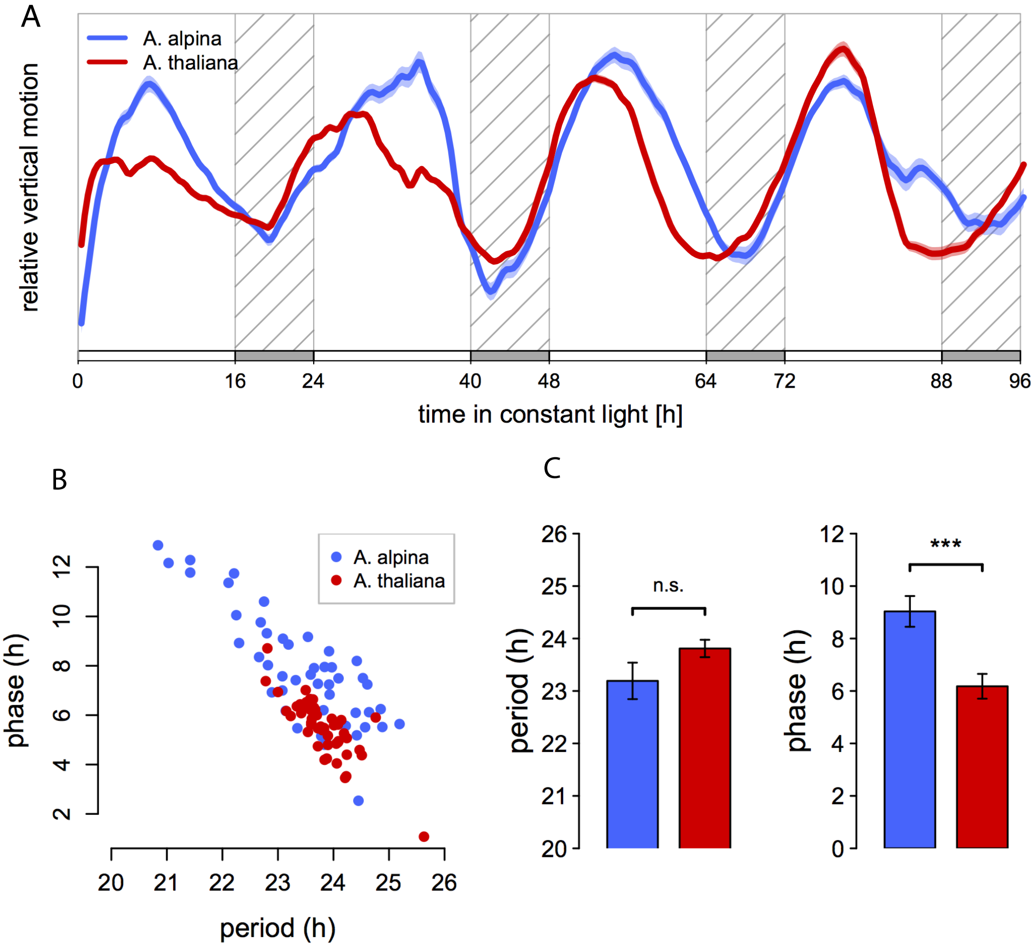


Figure 18 | Circadian leaf movement data analysis

(A) The mean relative vertical motion of 40 *Arabis* and 57 *Arabidopsis* seedlings under constant light is shown in red and blue, respectively. Colored shading shows standard error in measurement (SEM); hatched areas in the background indicate subjective nights. **(B)** Period and phase estimates of the same seedlings shown in (A). **(C)** mean period and phase estimates ± SEM (n = 40+57). Asterisks indicate significant differences (t-test, $p < 0.001$ ***, $p > 0.05$ n.s.). n.s = not significant.

Similar analysis was performed on genes, diurnal in both species (3,357 genes) that resulted in several enriched GO categories such as response to stimulus (p-value: 3.4e-28), cellular nitrogen compound metabolic process (p-value: 7.7e-23), post-embryonic development (p-value: 1.8e-16), response to light stimulus (p-value: 8.1e-16), photosynthesis (p-value: 2.7e-12), and several metabolic processes etc.

Table 5 | List of enriched functional categories (top 10) in *Arabidopsis* diurnal genes (7,551 genes)

Description of category (Biological process)	p-value	False Discovery Rate (FDR)
Response to stimulus	5.5e-55	4.6e-51
Response to abiotic stimulus	5.4e-34	2.3e-30
Response to stress	1.1e-32	3.1e-29
Response to chemical stimulus	1.4e-29	3e-26
Cellular nitrogen compound metabolic process	5.6e-27	8e-24
Post-embryonic development	7.4e-26	9.1e-23
Metabolic process	9.1e-23	9.7e-20
Response to organic substance	9.4e-20	9e-17
Cellular carbohydrate metabolic processes	3.8e-18	3.3e-15
Response to inorganic substance	9.4e-20	9e-17

Table 6 | List of enriched functional categories (top 10) in *Arabis* diurnal genes (5,332)

Description of category (Biological process)	p-value	False Discovery Rate (FDR)
Cellular nitrogen compound metabolic process	4e-27	2.8e-23
Post-embryonic development	4.9e-24	1.7e-20
Response to abiotic stimulus	5.1e-20	1.2e-16
Response to stimulus	7.3e-20	1.3e-16
Nitrogen compound metabolic process	8.7e-14	1.2e-10
Response to chemical stimulus	1.3e-13	1.5e-10
Metabolic process	5.8e-12	5.1e-09
Response to organic substance	1.8e-11	1.4e-08
Cellular ketone metabolic process	3.9e-11	2.7e-08
Establishment of localization	6e-11	3.1e-08

Further phase shift specific analyses were based on 2,751 out of 3,357 hc diurnal genes in *Arabidopsis* with phase reliably assigned in both species. Functional categories enriched in diurnal genes with similar phase (<=2hrs phase difference) and genes with the higher phase difference (> 2 hrs phase difference) were analyzed. All diurnal genes in *Arabidopsis* (with identified hc diurnal orthologs in *Arabis*) were then divided into two groups. Overall 1,324 gene pairs (diurnal gene and its ortholog in other species) were identified with similar phase and 1,427 gene pairs with higher phase difference. Genes that show a lesser phase difference were enriched for light regulation related categories (**Table 7**) whereas, most shifted genes were enriched for for functional categories such as response to stress (abiotic, chemical, inorganic stress), developmental and metabolic processes. These processes, especially metabolism related processes, are believed to be under tight control of the circadian clock (**Table 8**) ((Farré & Weise, 2012); (Haydon et al., 2013); (Greenham & Mcclung, 2015)).

Table 7 | List of enriched functional categories (top 10) in diurnal genes with similar phase (<=2hr) (1,324 genes)

Categories (Biological process)	**p-value**	**False Discovery Rate (FDR)**
Response to abiotic stimulus	3.7e-24	1e-20
Response to stimulus	4.8e-19	6.8e-16
Response to light stimulus	3.4e-18	2.7e-15
Response to radiation	4e-18	2.7e-15
Photosynthesis	1.9e-17	1.1e-14
Response to chemical stimulus	9.5e-13	3.7e-10
Cellular nitrogen compound metabolic process	1.6e-12	6.4e-10
Response to organic substance	1.6e-11	5.6e-09
Response to temperature stimulus	3.1e-11	1e-08
Response to endogenous stimulus	2e-10	5.6e-08

Table 8 | List of enriched functional categories (top 10) in diurnal genes with shifted phase (>2hr) (1,427 genes)

Categories (Biological process)	**p-value**	**False Discover Rate (FDR)**
Response to stimulus	2e-17	5.7e-14
Cellular nitrogen compound metabolic process	1.9e-15	2.7e-12
Response to chemical stimulus	4.1e-14	3.9e-11
Response to inorganic substance	1.5e-10	1e-07
Response to stress	2.7e-10	1.6e-07
Post-embryonic development	9.8e-10	4.7e-07
Response to abiotic stimulus	2.1e-09	8.7e-07
Cellular ketone metabolic process	2.7e-09	9.6e-07
Oxoacid metabolic process	4.7e-09	1.3e-06
Organic acid metabolic process	4.9e-09	1.3e-06

To gain more insight of enriched GO categories and average phase difference in genes underlies, I calculated average phase difference for all categories enriched in diurnal genes in *Arabidopsis* with diurnal orthologs in *Arabis* (p-value < 0.001, FDR < 0.05) and sorted these categories in ascending order of average phase difference (**Figure 19)**. The top five most shifted categories and least shifted functional categories were compared with published list of clock genes ((Covington & Harmer, 2007); (Michael et al., 2008); (Covington et al., 2008)). The most shifted categories showed a slightly higher fraction of clock genes (95% as compared to 88 %).

As, I already observed that clock genes and their direct targets show phase shift in *Arabis*, it was possible that a non-uniform distribution of clock genes contribute phase shift in functional categories. Therefore, I compared the average phase shift of all diurnal genes (including clock genes) with average phase shift of clock genes for each functional category. As expected average phase shift for clock genes, was more than diurnal genes for each category. This supported the hypothesis that in general, phase of clock-controlled genes are shifted whereas, genes that are also regulated by light have diluted and variable impact of the shift.

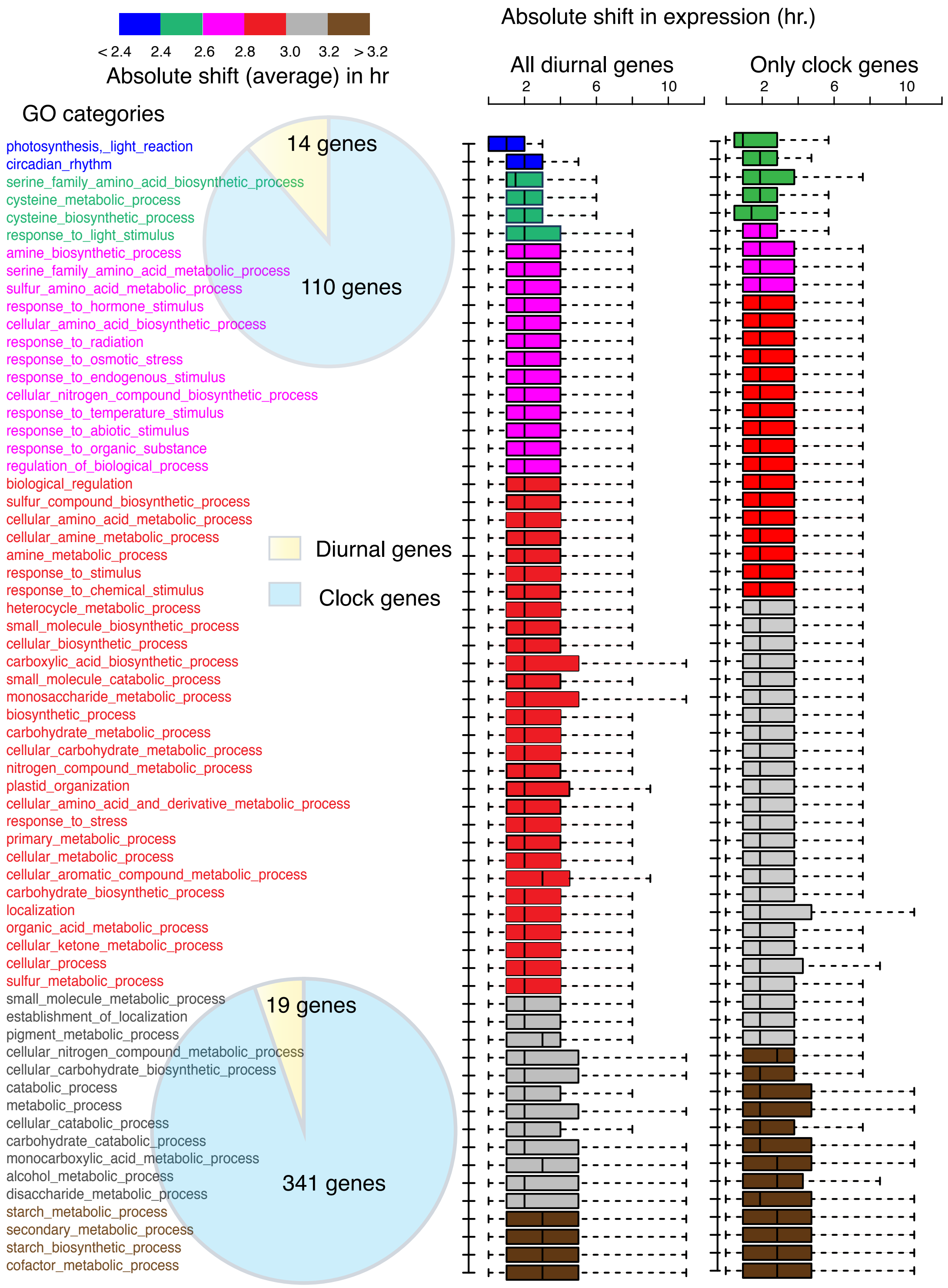


Figure 19 | GO term analysis in genes diurnal in both species.

GO categories are sorted in ascending order of average shift and boxplot for each functional category represent distribution of phase shift (with box associated plot) shown by each gene form that category. Color scale is provided for average phase difference for a functional category. Two series of boxplots represent two sets of diurnal genes, all diurnal genes and diurnal genes that are identified as clock controlled genes in constant conditions. A pie chart representing contribution of clock genes (from published data) in top five most shifted and least shifted categories.

3.2.9 Defining regulatory region for DNA motif identification

Defining promoter (or regulatory) region is the first step in identification of CREs. I decided to focus on the complete upstream intergenic region as the putative regulatory region for identification of CREs in diurnal gene. Though, CREs have been described within intron and downstream regions, the majority of CREs in *Arabidopsis* are located in upstream region of genes ((Rombauts et al., 2003); (Zhang et al., 2012); (Sieburth & Meyerowitz, 1997); (Sheldon et al., 2002); (Kooiker et al., 2005)). For DNA motif identification, this definition of putative regulatory region is reasonable compromise in gene-dense genomes like *Arabidopsis* (AT) and other seven Brassicaceae species namely, *Arabidopsis lyrata* (AL), *Capsella rubella* (CR), *Sisymbrium irio* (SI)*, Eutrema salsugineum* (ES)*, Aethionema arabicum* (AAT*)*, *Schrenkiella parvula* (SP) and *Arabis* (AA) (**Figure 20A**). Intergenic regions of orthologous genes from seven Brassicaceae species were analyzed to define conservation blocks using PHAST package (Hubisz et al., 2011) for each orthologous intergenic region in each species. In addition to unassembled region, sequences with high repeat content were masked before using RepeatMasker V3.3.0 (Smit, Hubley & Green, http://www.repeatmasker.org) (**Figure 20B, Figure 20C, see Materials and Methods**).

To summarize, after removing missing bases (N's), known repeats, and regions with low conservation, average regulatory regions were reduced to 15-85% (median) of raw intergenic length (**Figure 20D**). This huge difference in fraction of conserved intergenic region was surprising, especially among the closely related species Average intergenic length was strongly correlated (Pearson correlation coefficient; r = 0.91) with (assembled) genome size (**Figure 20A**) but average conserved intergenic region found not to be positively correlated with genome size (**Figure 20C**).

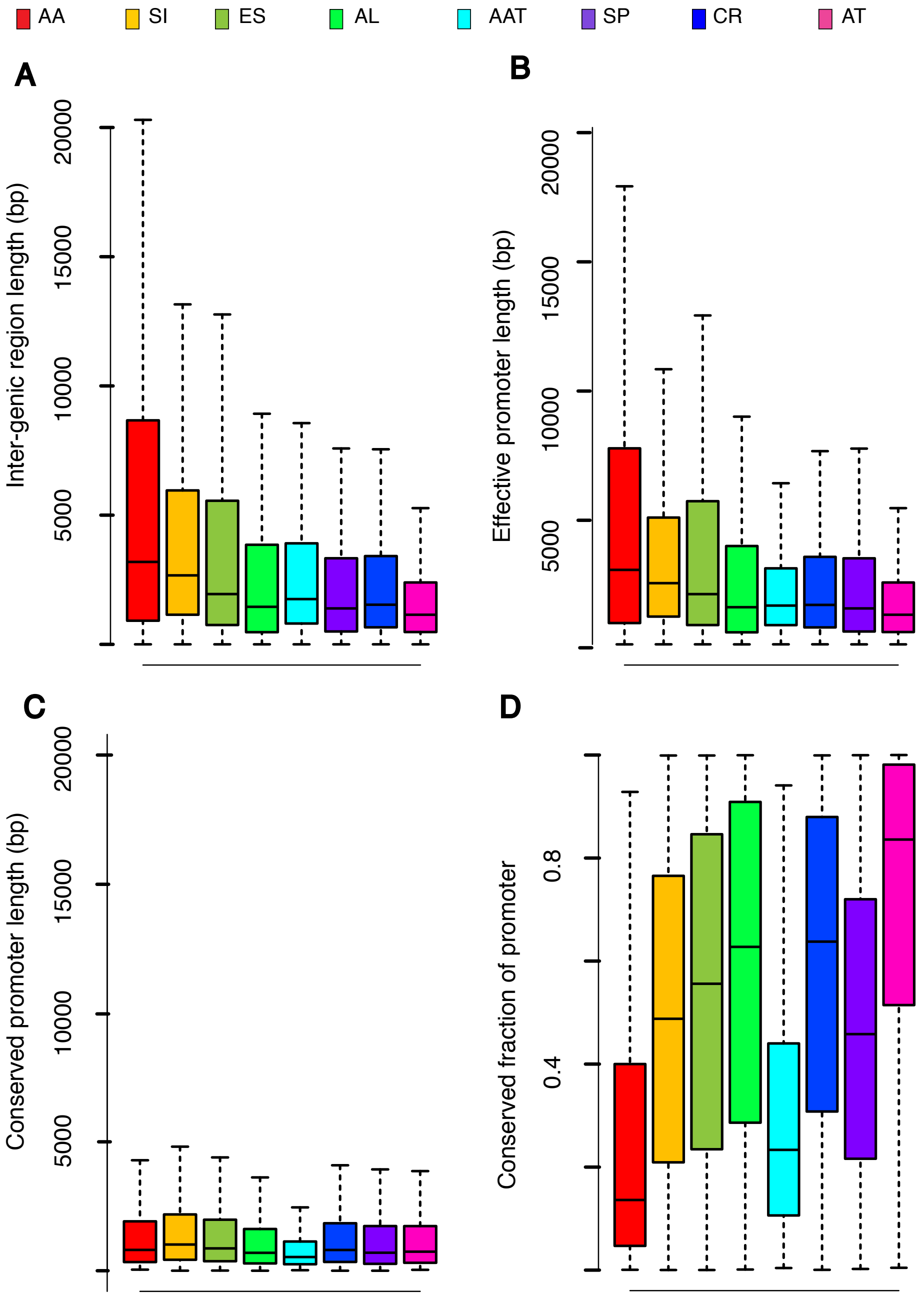


Figure 20 | Defining regulatory regions for eight Brassicaceae

***Arabis* (AA), *Sisymbrium irio* (SI), *Eutrema salsugineum* (ES), *Arabidopsis lyrata* (AL), *Aethionema arabicum* (AAT), *Schrenkiella parvula* (SP), *Capsella rubella* (CR) and *Arabidopsis* (AT). Genomes arranged in descending order of assembly size (A)** Intergenic length distribution. **(B)** Effective (unambiguous) intergenic length distribution after removing assembly gaps and known repeats. **(C)** Conserved unambiguous intergenic length distribution. **(D)** The proportion of intergenic sequences retained as the putative regulatory region.

3.2.10 Identification of CREs using Phylogenetic shadowing

Phylogenetic shadowing is a powerful approach for *de novo cis*-element identification in conserved regions of a promoter of a gene. Recently published several Brassicaceae genome assemblies encouraged us to employ Phylogenetic shadowing in diurnal genes of *Arabidopsis* and *Arabis* ((Slotte et al., 2013); (Willing et al., 2015); (Wang et al., 2011); (Dassanayake et al., 2011); (H.-J. Wu et al., 2012); (Haudry et al., 2013); (Kitashiba et al., 2014); (Lobréaux et al., 2014); (Liu et al., 2014)). Genome-wide orthologous gene prediction was performed with the "reciprocal-best-hit" approach (blastall version 2.2.26) for all above-mentioned Brassicaceae genomes in the pair-wise manner (**Table 9**) (Altschul et al., 1990).

Table 9 | Orthologs for *Arabidopsis* and *Arabis* in six other Brassicaceae

	Arabidopsis	*Arabis*
Arabidopsis		18,633
A. lyrata	22,385	17,899
C. rubella	19,991	17,502
E. salsugineum	18,975	16,944
S. parvula	18,580	17,090
S. irio	17,574	15,809
A. arabicum	14,145	12,958
Arabis	17,988	

Phylogenetic shadowing was performed for all diurnal genes to search *cis* element residing in conserved promoters regions in *Arabidopsis* and *Arabis* independently (**as described in Figure 21B**). However, as orthologs of nearly 60% of the diurnal gene in *Arabidopsis* were also identified diurnal in *Arabis,* the analysis was partially overlapping. All diurnal genes were analyzed independently with orthologous genes from seven other Brassicaceae (*Arabidopsis*/*Arabis*, *A. lyrata, C. rubella, S. parvula, E. salsugineum, A. atheanema and S. irio*) (**Figure 21B**). *De novo* identification of *cis*-elements was performed using MEME tool (Bailey & Elkan, 1994). For background correction, hidden Markov model (HMM) background file for all intergenic regions was generated and used with MEME to filter low complexity and repetitive regions. In all, 6,876 *Arabidopsis* and 8,130 *Arabis* diurnal genes (with orthologs) were analyzed with phylogenetic shadowing and resulted in 56,133 motifs with 133,076 sites (Avg. 8.16 motif per gene and 2.37 sites per motif/gene) for *Arabidopsis* and 55,820 motifs and 155,444 sites for *Arabis* (Avg. 6.83 motifs per gene and 2.78 sites per gene/motif).

Assuming that real DNA motifs will be present in several diurnal genes whereas, false positive (predicted due to high promoter sequence conservation) will not be common across many diurnal genes, I compared all identified motif to generate a similarity matrix for clustering of similar motifs using matrix comparison tool, MatAlign v-4a (http://stormo.wustl.edu/MatAlign/) (**Figure 21C**). Subsequently, Markov Cluster Algorithm (MCL) (Dongen, 2000) was used to generate clusters of similar motifs (**Figure 21C).** Motif clusters with motifs identified in more than 25 genes (but less than 10% of diurnal genes) were retained for further analysis. The rationale behind setting a lower limit (motifs identified in >25 gene) was to focus more on *cis*-elements controlling a larger set of diurnal genes and for upper limit (motifs identified in < 10% of diurnal gene) was to remove those present in hundreds of promoters and picked up mostly due to low sequence complexity. Filtering motif clusters with upper limit removed only five motif clusters, low complexity motifs such as polyA, AT repeats etc (**see Material and Methods**). Where as filtering with lower limit removed >99% motif clusters. Over 95% of the removed clusters were with <5 motifs per clusters representing mostly background noise. For resulting motif clusters (91 *Arabidopsis* and 92 *Arabis*) MEME tool was used to generate consensus motif for sequence stretches where motif was identified. For generating a consensus motif for a cluster of motifs, evidence-sequences (sequences which contributed for motif prediction) were taken only from *Arabidopsis* or *Arabis*, to avoid bias coming from other species in motif definition (**Figure 21D; see Materials and Methods**).

The final output of this pipeline, called **PhyloConCisE**, was an atlas of DNA motifs potentially important for diurnal expression. This diurnal motif atlas also includes previously identified, all seven CREs (Morning Element, GBOX, Evening Element, GATA box, Protein Box, Telo Box, Starch Box) (**Figure 22A and 22B)**.

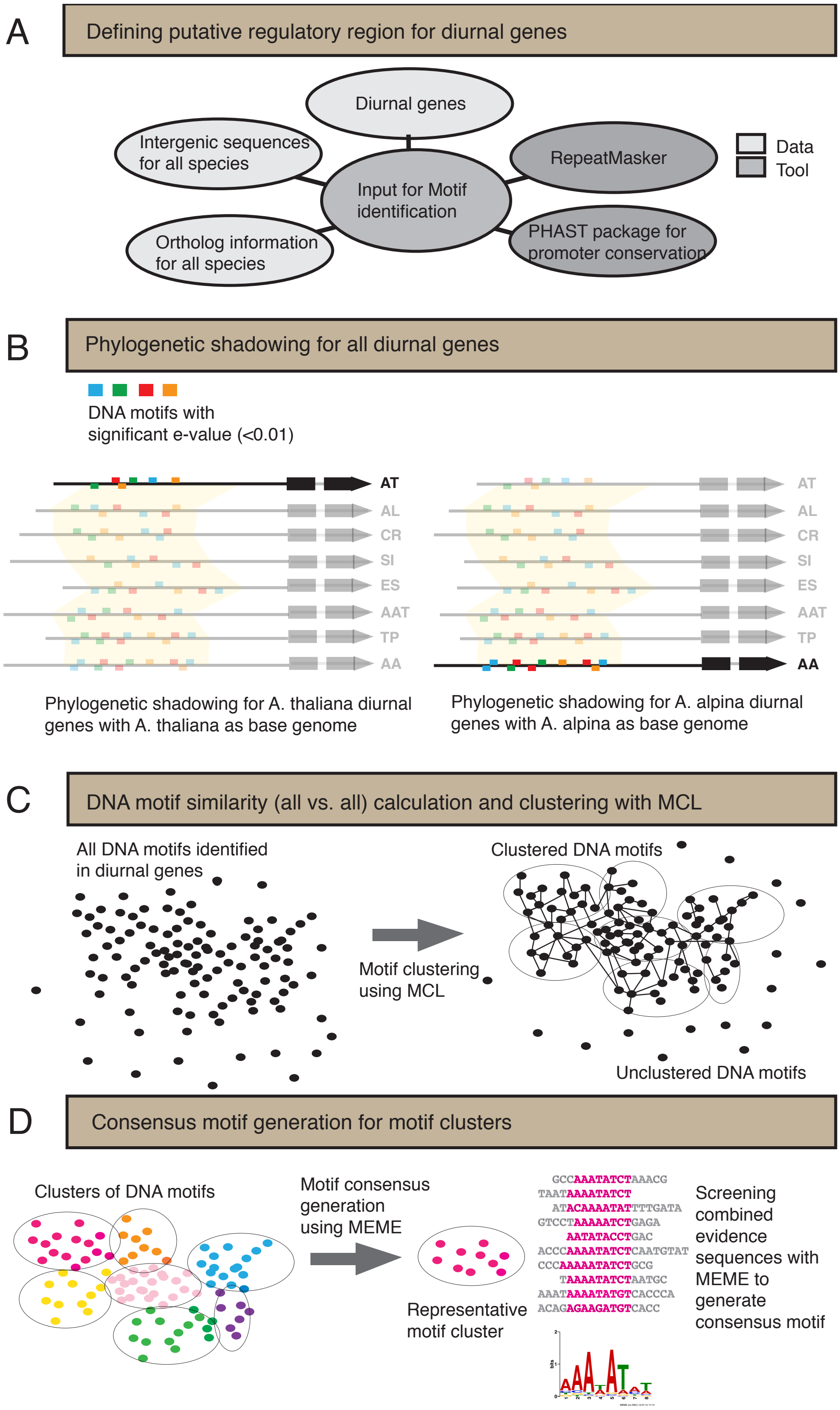

A
Defining putative regulatory region for diurnal genes
Diurnal genes
Intergenic sequences for all species
Input for Motif identification
RepeatMasker
Data
Tool
Ortholog information for all species
PHAST package for promoter conservation
B
Phylogenetic shadowing for all diurnal genes
DNA motifs with significant e-value (<0.01)
AT
AL
CR
SI
ES
AAT
TP
AA
Phylogenetic shadowing for A. thaliana diurnal genes with A. thaliana as base genome
Phylogenetic shadowing for A. alpina diurnal genes with A. alpina as base genome
C
DNA motif similarity (all vs. all) calculation and clustering with MCL
All DNA motifs identified in diurnal genes
Motif clustering using MCL
Clustered DNA motifs
Unclustered DNA motifs
D
Consensus motif generation for motif clusters
Clusters of DNA motifs
Motif consensus generation using MEME
Representative motif cluster
GCCAAATATCTAAACG
TAATAAAATATCT
ATACAAAATATTTTGATA
GTCCTAAAAATCTGAGA
AATATACCTGAC
ACCCAAAATATCTCAATGTAT
CCCAAAAATATCTGCG
TAAAATATCTAATGC
AAATAAAATATGTCACCCA
ACAGAGAAGATGTCACC
Screening combined evidence sequences with MEME to generate consensus motif
2
bits
1
0
1
2
3
4
5
6
7
8
MEME (no SSC) 14.07.14 17:14

Figure 21 | Overview of Phylogenomic shadowing pipeline

(A) Input preparation step for phylogenetic shadowing in diurnal genes **(B)** Phylogenetic shadowing for all diurnal genes **(C)** Motif clustering step **(D)** Consensus motif building step.

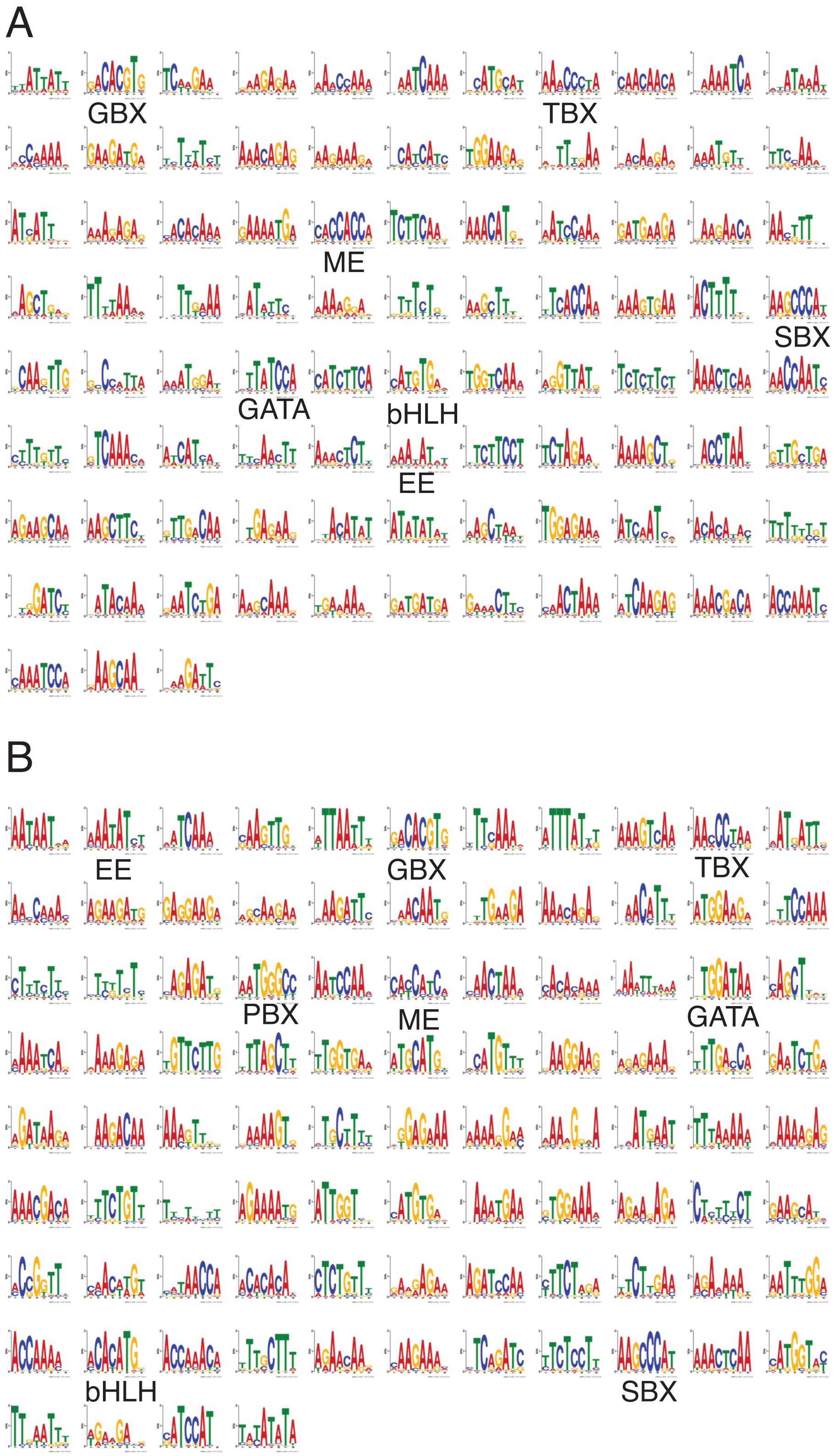


Figure 22 | Atlas of diurnal DNA motifs

(A) ***Arabidopsis*** **(B)** ***Arabis.*** EE: Evening Element, GATA: GATA box, GBX: G box, ME: Morning Element, PBX: Protein Box, SBX : Starch Box, TBX : Telo Box.

3.2.11 Filtering non-significant diurnal DNA motifs (DDMs) and generating comparative motif set

Once all DDMs (91 for *Arabidopsis* and 92 *Arabis*) were obtained, it was important to check if these motifs are showing significant enrichment for a time point cluster of diurnal genes. The rationale behind this filtering step was to remove motifs enriched in genes with different phases; such motifs should either be non-significant for phase specific expression with a ubiquitous presence in several diurnal genes or simply a false positive. In either case, it was a low priority motif for this analysis. For assessing time point specific significant enrichment, coefficient of variation (CV) >= 0.15 was selected. This filtering step resulted in 54 *Arabidopsis* and 45 *Arabis* time-specific DDMs. Next, I performed a hierarchical clustering step followed by "cutree" function (implemented in R) with appropriately estimated (elbow point estimation) number of clusters to get time-point specific DDMs (tpsDDMs) set. Clustering of tpsDDMs resulted in 44-motif clusters, out of which 29 clusters (66%) showed at least one representative from both species (**Figure 23**).

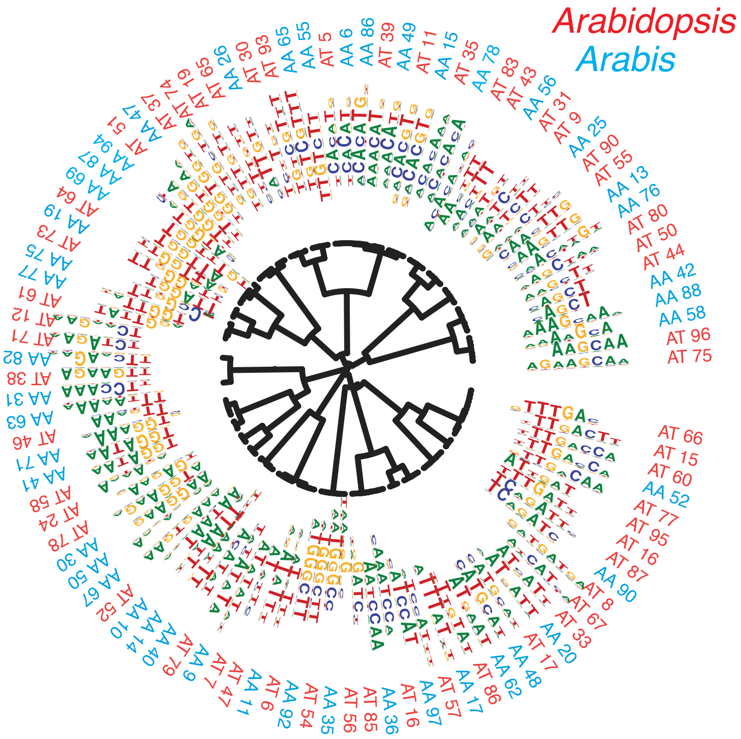


Figure 23 | Comparison of tpsDDMs

Circular dendrogram of tpsDDMs in *Arabidopsis* and *Arabis*.

3.2.12 Analysis of DNase I data for validation of tpsDDMs

To examine whether the tpsDDMs were associated with footprints in DNase-seq data set, I analyzed published DNase I hyper-sensitive sites (DHSs) recently released for *Arabidopsis* leaf tissue data (Zhang et al., 2012)). It was one of the first DHS dataset available for *Arabidopsis* and was generated to answer different biological questions. The partially digested chromatin data was collected for a single time point (mid day) for leaf tissue whereas; tpsDDMs were identified in whole seeding data and are potentially acting during different time points of a day. This might result in many tpsDDM represent unoccupied sites by TFs (with no protection from DNase I digestion) but still could provide an opportunity to validate at least some of the identified tpsDDMs. For each tpsDDM, all DHSs with a tpsDDM instances were centered with the tpsDDM instance in middle along with 50 bp flanking sequence around the instance. A clear dip(s) in the profile of DNase I activity (cuts per nucleotide) was observed at the center of the sequence alignments for 16 out of 54 tpsDDMs analyzed (30%), supporting the idea that many of these tpsDDMs are likely to be a true protein binding DNA elements. The position of the motif overlapped with a reduction of the DNase-seq read count which was used as proxy for DNase I activity (**Figure 26A, 26B, 26C, and 26D**), suggesting these sites were relatively more protected from the DNase I digestion as compared to flanking regions.

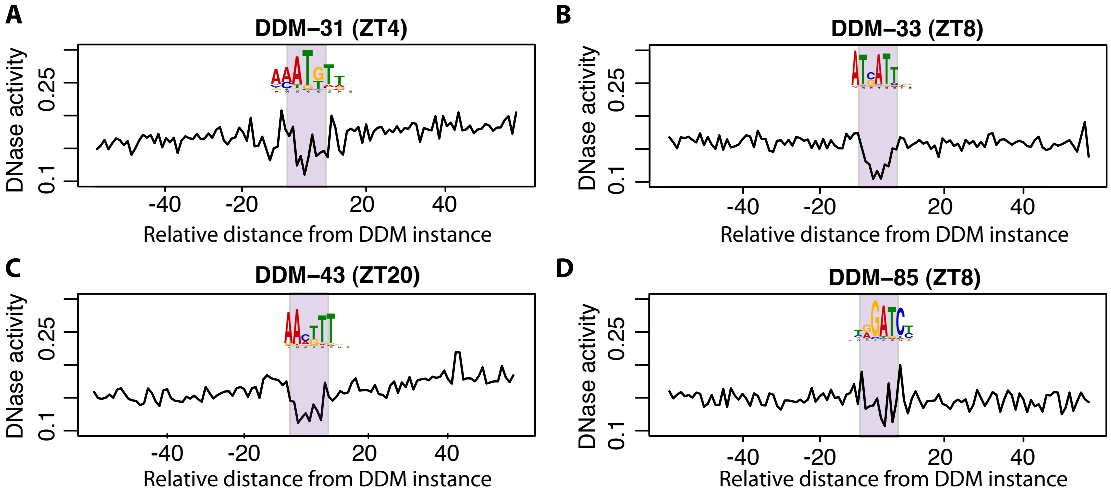


Figure 24 | Some examples showing reduced cleavage frequency of DNase I around DDM instances

(A) DDM-31 enriched for ZT4 expression cluster. **(B)** DDM-33 enriched for ZT8 expression cluster. **(C)** DDM-43 enriched for ZT20 expression cluster. **(D)** DDM-85 enriched for ZT8 expression cluster.

3.2.13 Time point specific enrichment of DDMs in diurnal genes

For 29 DDM clusters (out of 44 identified tpsDDM clusters) enrichment in six time clusters was calculated. Here, I used six time point clusters and not 24 time point clusters, to get cleaner trend, as similar analysis with 24 time point clusters resulted in enrichment pattern with unstable trend. Over half of tpsDDMs (28 out of 54 motifs in *Arabidopsis* and 26 out of 45 in *Arabis*) were significantly (p-value < 0.05, with Bonferroni correction) enriched for a specific time point. Instances of motifs were identified employing position weight matrix (PWM) based screening in 1.5kb upstream region of all genes of co-expression cluster using Motif Occurrence Detection Tool (MOODS) (Korhonen et al., 2009). To get an average number of instances per gene, total count of motif instances was divided by the number of genes present in a time point cluster. Time point with maximum sites per gene was assigned to a DDM. Only 33% of the total motifs in *Arabidopsis* (19 out of 54) were found enriched from dusk to night (ZT12, ZT16, and ZT20) suggesting a relatively lower number of transcription factors controlling dark phase expressed genes compared to light phase (36 DNA motifs out of 54). This pattern was also observed in *Arabis*, in an even more pronounced manner where only 23% (11 out of 46) motifs were enriched in dark phase gene clusters (**Figure 24A and Figure 24B, Table 10)**.

3.2.14 Shifted enrichment profile of tpsDDMs

I compared *Arabidopsis* and *Arabis* tpsDDM enrichment profiles to check if the phase shift of diurnal genes could be recapitulated with shifted enrichment profile of motifs. The normalized score (Z-score) for tpsDDM enrichment (tpsDDM instances per gene) for all six-time points was calculated for 29 motifs clusters (**Figure 25**). For clusters with more than one motif per species, average enrichment (per time point) was taken for all motifs. Though, I could analyze only 29 cases, it was enough to get an idea about the enrichment pattern of tpsDDMs. Enrichment patterns were compared with cross correlation test (implemented in R). Nearly 30% cases (**9 out of 29 cases**) similar enrichment pattern was observed for both species, in another 30% cases (**9 out of 29 cases**) motif pairs clearly displayed delayed enrichment for *Arabis*, eight cases were not found conclusive enough and three cases displayed delyed shift for *Arabidopsis* (**Figure 25A, 25B, 25C and 25D**).

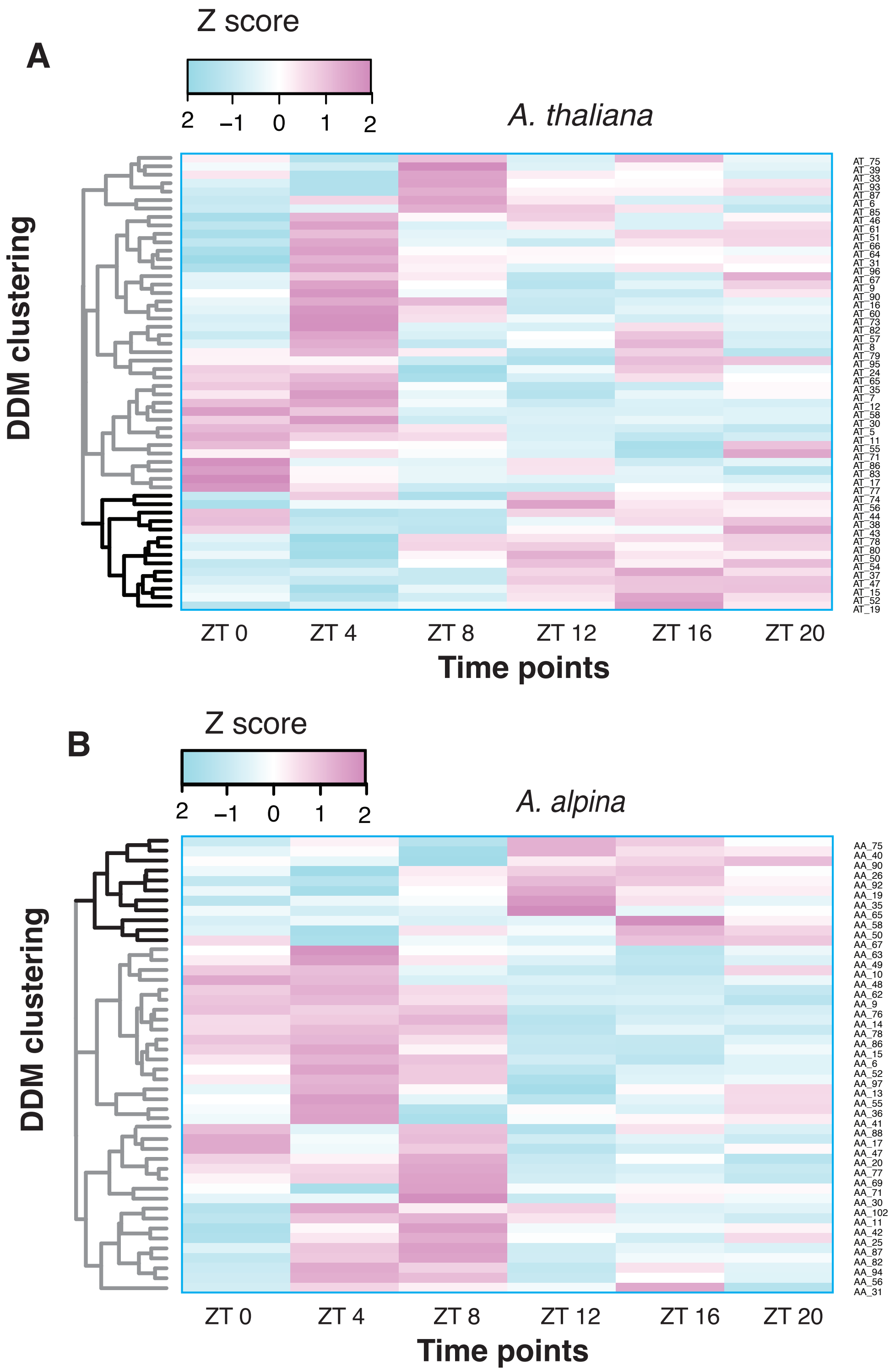


Figure 25 | Time point specific enrichment of DDMs

(A) Time point specific enrichment analysis for *Arabidopsis* tpsDDMs **(B)** Time point enrichment analysis for *Arabis* tpsDDMs.

Table 10 | Highest enrichment time point for DDM clusters (*Arabidopsis* and *Arabis*)

Motif (*Arabidopsis*)	Time point	Motif (*Arabis*)	Time point	Motif Cluster
AT_46	ZT4	AA_63	ZT4	1
AT_35, AT_86	ZT0	---------		2
AT_95	ZT16	---------		3
AT_19, AT_74	ZT16	---------		4
AT_73, AT_61	ZT4	AA_77	ZT8	5
AT_24, AT_44, AT_50, AT_30	ZT4	AA_42, AA_67	ZT8	6
AT_8, AT_67	ZT4	---------		7
AT_57	ZT4	AA_97, AA_30	ZT8	8
---------		AA_40, AA_14		9
AT_37, AT_51	ZT4	AA_47	ZT0	10
AT_7, AT_79	ZT4	AA_9	ZT4	11
AT_6	ZT8	AA_11	ZT8	12
AT_65, AT_78	ZT16	AA_50	ZT16	13
AT_64	ZT4	AA_69	ZT8	14
AT_55	ZT0	AA_13	ZT4	15
AT_33, AT_82	ZT8	AA_20, AA_62	ZT8	16
AT_93	ZT8	AA_65	ZT12	17
AT_77	ZT0	AA_55	ZT4	18
AT_83	ZT0	AA_78	ZT4	19
AT_5	ZT0	AA_6, AA_86	ZT4	20
---------		AA_17	ZT0	21
AT_17	ZT0	AA_94, AA_48	ZT4	22
---------		AA_31, AA_71	ZT16	23
AT_87	ZT8	AA_25	ZT8	24
AT_75, AT_38	ZT16	---------		25
AT_54	ZT12	AA_35, AA_92	ZT12	26
AT_85	ZT8	AA_90	ZT20	27
AT_15, AT_52	ZT16	AA_19	ZT12	28
AT_96	ZT4	AA_88	ZT0	29
AT_11	ZT0	AA_15	ZT4	30
---------		AA_87	ZT8	31
AT_31	ZT4	AA_76	ZT8	32
AT_12	ZT4	AA_82	ZT8	33
AT_90, AT_9	ZT4	---------		34
AT_60	ZT4	AA_52	ZT4	35
---------		AA_10		36
AT_56,AT_47, AT_80	ZT12	---------		37
AT_39	ZT8	AA_49	ZT4	38
---------		AA_75		39
AT_16	ZT4	AA_36	ZT4	40
---------		AA_26		41
AT_43	ZT20	AA_56	ZT4	42
AT_58	ZT0	AA_41	ZT4	43
---------		AA_58	ZT16	44

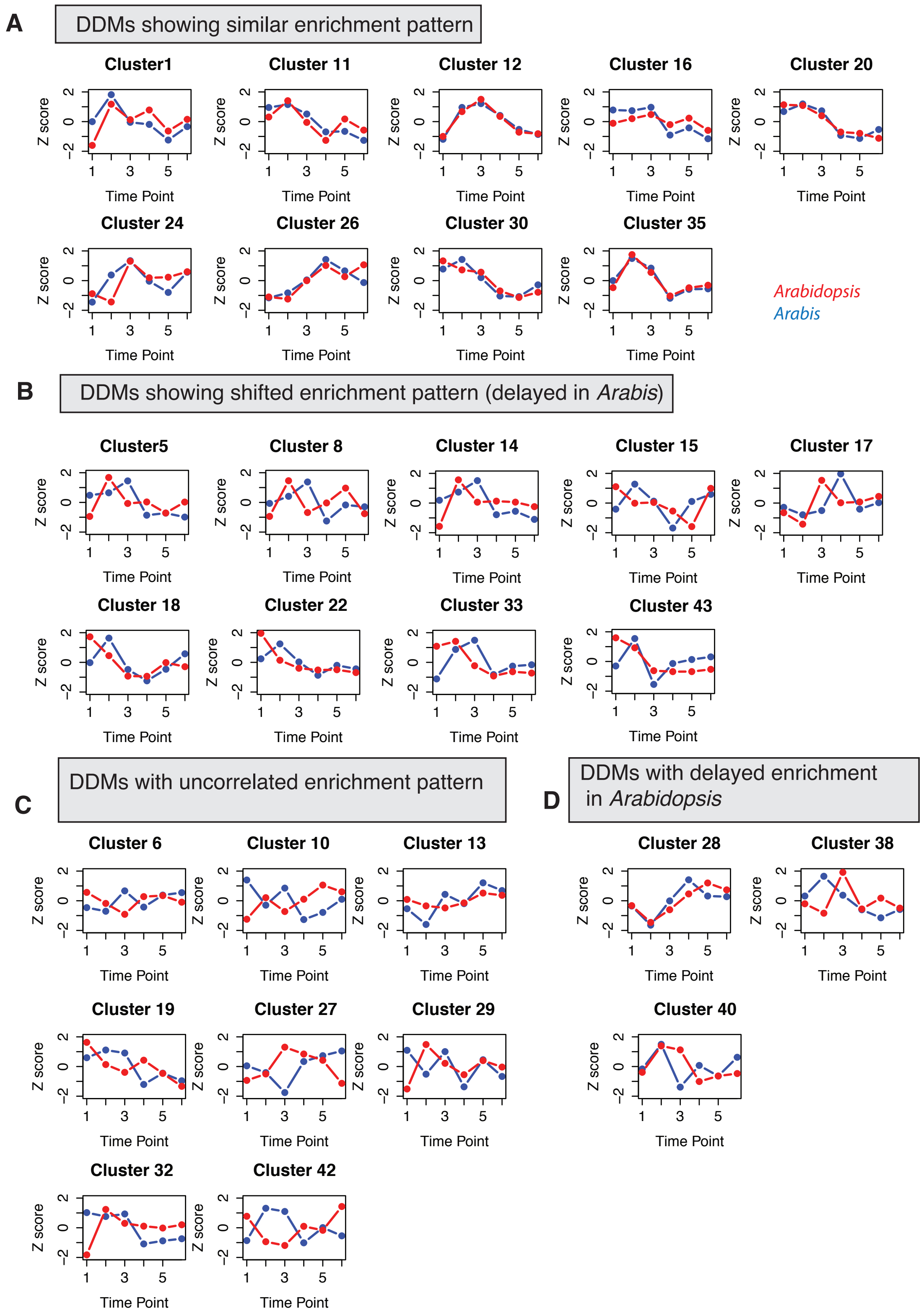


Figure 26 | Comparison of enrichment pattern for tpsDDMs in *Arabidopsis* and *Arabis*

(A) tpsDDMs with similar enrichment pattern. (B) tpsDDMs with delayed enrichment pattern for *Arabis*. (C) tpsDDMs with uncorrelated enrichment pattern (along with *Arabidopsis* tpsDDMs with delayed enrichment). (D) tpsDDMs with delayed enrichment pattern for *Arabidopsis*.

3.2.15 Identification and comparison of *cis*-regulatory modules (CRMs)

Genes are commonly regulated by several TFs, whose binding sites are clustered as separate modules rather than uniformly distributed in intergenic regions (Berman et al., 2002). These clusters of CREs are commonly referred to as *cis*-regulatory modules (CRMs). Generally, CRMs range from few hundred bps to few thousand bps in size. The set of tpsDDMs in *Arabidopsis* and *Arabis* were employed to find significantly co-occurring pairs of tpsDDMs. I tested all possible motif pairs (54X54 and 45X45) with motif spacing ≤ 50 bps for enrichment in 1,500 bp upstream to transcription start sites (TSS) of diurnal genes present in all time point clusters (**see Materials and Methods**).

To quantify enrichment, I counted the number of tpsDDM pair instances in TSS upstream region for diurnal genes present in each time point clusters (six time point clusters were considered), and then compared against a background model based on the number of instances in the union set of all diurnal genes. The significance (p-value) of enrichment was assessed using a binomial distribution, after correcting for multiple testing (**see Materials and Methods**). Motif complexes showing statistically significant enrichment ($P < 0.05$ after Bonferroni correction) were recognized as time point specific co-occurring DDM pairs (potential diurnal CRMs). Application of this approach across all motif pairs and all time points discovered 88 (64 unique pairs) for *Arabidopsis* and 91 (48 unique pairs) for *Arabis* significant tpsDDM pairs. A smaller fraction (20%) of all identified co-occurring tpsDDM pairs was commonly identified in both species and represent a set of putative conserved diurnal CRMs (**Table 11**). These conserved diurnal *cis*-regulatory modules (dCRMs) include ME, EE, GBX and several other novel CREs. Two dCRMs, identified in this study, EE-ABREL and EE-EE are already reported to show combinatorial role with close positioning (intermediate distance < 50bp) in *Arabidopsis* GIGANTEA (GI) gene (Berns et al., 2014). Our findings not only suggest that these CRMs might be important for several other *Arabidopsis* genes but also extend their importance to *Arabis* and other Brassicaceae.

Table 11 | Significant pairs of co-occurring tpsDDM enriched in different time point clusters of diurnal genes

Time point	***Arabidopsis* (significant pairs)**	***Arabis* (significant pairs)**	**Common**
ZT 0	16	16	3
ZT 4	5	6	0
ZT 8	12	7	2
ZT 12	12	29	2
ZT 16	14	16	2
ZT 20	29	17	2
	88 (64 unique pairs)	91 (48 unique pairs)	11

3.2.16 Loss and gain of diurnal expression

Around 60% of the diurnal genes identified in one species showed diurnal orthologs in the other species, highlighting conservation of diurnal expression between these two species. However, around 8.8% (479) of all diurnal genes in *Arabidopsis* and 10.6% (615) in *Arabis* were identified with hc non-diurnal orthologs. To test if these difference in diurnal expression could be explained at difference in regulatory region, I compared intergenic sequence length of three gene sets: **A**. All diurnal genes **B**. Diurnal genes with diurnal orthologs **C.** Non-diurnal genes with diurnal orthologs. Comparison of intergenic length distribution resulted with no significant differences among these sets for *Arabidopsis* and *Arabis* (Wilcox Test, p-value > 0.05) (**Figure 27A and 27B**). Though, overall intergenic length distribution revealed slightly shorter intergenic length for non-diurnal genes (with diurnal orthologs) as compared to diurnal genes with diurnal orthologs but this difference was not significant. Nevertheless, there was still a possibility that *Arabidopsis* non-diurnal gene (with diurnal ortholog) represent genes that are affected by small deletions at regulatory regions and lost crucial TFBSs for diurnal expression. As it has been shown that *Arabidopsis* genome size reduction in comparison to *Arabidopsis lyrata*, is predominantly contributed by small-scale (1-3 bp) deletions (Hu et al., 2011). There was a possibility that gene set **C** (for *Arabidopsis*) is enriched for such deletions. To get comparative estimate of regulatory motif content in intergenic region of all three set (**set A, B and C**) PWM-based search of all DDMs was performed. The presence/absence of DDM (and not time specificity) was important; I employed all previously identified DDMs (Complete set of 91 *Arabidopsis* and 92 *Arabis* DDMs). DDM instances in intergenic regions were identified and compared between all previously defined gene sets (**Figure 28A and Figure 29A**). *Arabidopsis* genes with non-diurnal expression displayed reduced content of DDM instances (median =2.1 instances of each motifs) as compared to gene with retained expression (median =2.3 instances of each motifs) (**Figure 28A**). Although, there was not any significant difference in intergenic length distribution still, I decided to normalize enrichment values with length of promoters. Length normalization reduced this difference but still genes with non-diurnal expression showed slight reduction in motifs instances per kb (**Figure 28B and 29B**).

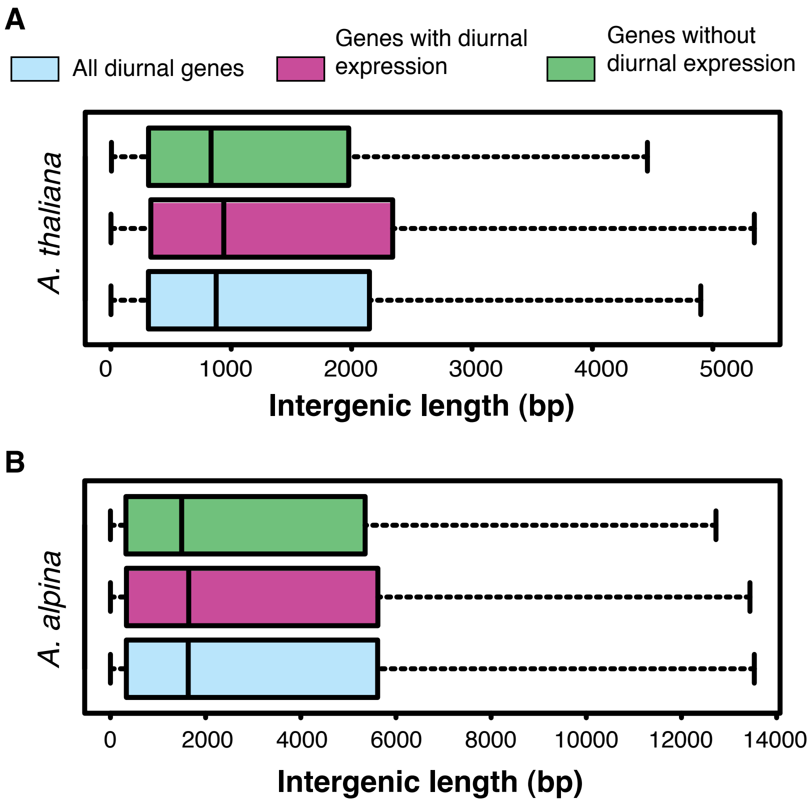


Figure 27 | Comparison of intergenic lengths of diurnal genes

(A) All *Arabidopsis* diurnal genes (light blue), *Arabidopsis* diurnal genes with diurnal ortholog in *Arabis* (pink) and non-diurnal *Arabidopsis* genes with diurnal ortholog in *Arabis* (green). (B) All *Arabis* diurnal genes (light blue), *Arabis* diurnal genes with diurnal ortholog in *Arabidopsis* (pink) and non-diurnal *Arabis* genes with diurnal ortholog in *Arabidopsis* (green).

All DDMs were identified in diurnal genes and low enrichment in non-diurnal genes might just an outcome of it, so I decided to employ known motif, Evening Element (EE) to get unbiased picture. I investigated instances of EE to get an unbiased comparison of motif enrichment. I first selected *Arabis* co-expression clusters with high enrichment of EE (ZT8: CT8, CT9, CT10, CT11; **Figure 15**) and divided orthologs of these genes (in *Arabidopsis*) into two classes First, genes with diurnal expression and second, genes without diurnal expression. Significantly (p-value < 0.05) reduced enrichment (EE instances per genes) in non-diurnal gene set for both species was observed (**Figure 28C**). Similar analysis was also done for orthologs of Arabidopsis genes with high enrichment of EE and this trend was consistent in Arabis too (**Figure 29C**). Although this reduction in enrichment doesn't not offer a causal relationship and neither it is confirming that "loss" of CRE resulted in "loss" of diurnal expression nevertheless, it provide an interesting observation and one of the many possible ways explanation of diurnal expression differences.

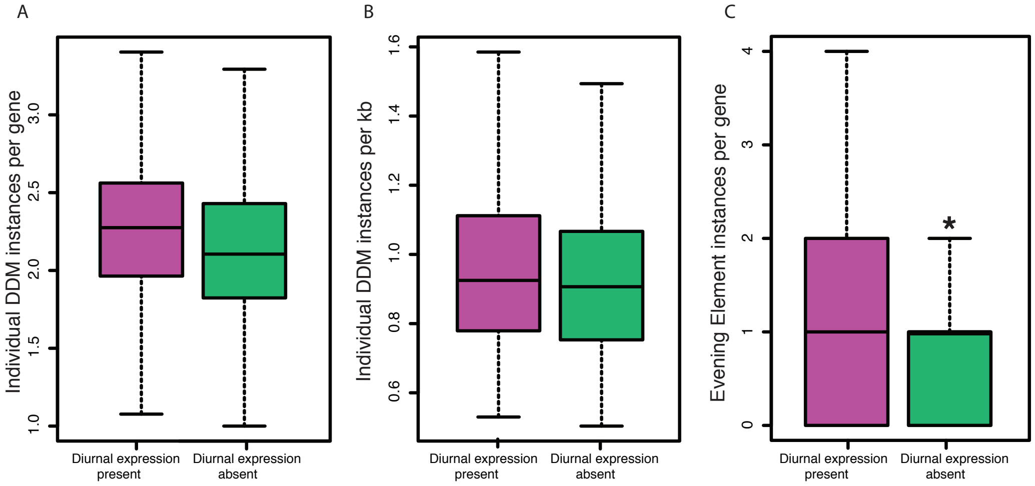


Figure 28 | DDM enrichment analysis for *Arabidopsis*

(A) Distribution of "DDM instances per gene" for *Arabidopsis* diurnal genes with diurnal ortholog (pink) and *Arabidopsis* non-diurnal genes with diurnal ortholog (green). **(B)** Distribution of DDM instances per kb of intergenic sequence for *Arabidopsis* diurnal genes with diurnal ortholog and *Arabidopsis* non-diurnal genes with diurnal ortholog **(C)** Orthologs of *Arabis* diurnal genes (with peak expression at CT8 -CT11; **Figure 15A**) with diurnal expression in *Arabidopsis* (pink), Orthologs of *Arabis* diurnal genes with non-diurnal expression in *Arabidopsis*. * Significantly different (Wilcox test, p-value < 0.05)

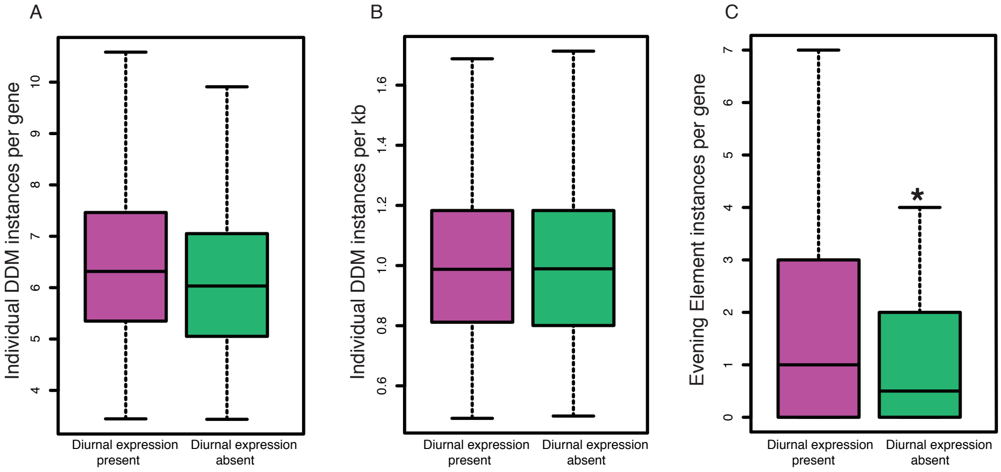


Figure 29 | DDM enrichment analysis for *Arabis*

(A) Distribution of "DDM instances per gene" for *Arabis* diurnal genes with diurnal ortholog (pink) and non-diurnal genes with diurnal ortholog (green). **(B)** Distribution of DDM instances per kb of intergenic sequence for *Arabis* diurnal genes with diurnal ortholog and *Arabis* non-diurnal genes with diurnal ortholog **(C)** Orthologs of *Arabidopsis* diurnal genes (with peak expression at CT8 -CT11; **Figure 15A**) with diurnal expression in *Arabis* (pink), Orthologs of *Arabidopsis* diurnal genes with non-diurnal expression in *Arabis*. * Significantly different (Wilcox test, p-value < 0.05)

3.3 Materials and Methods

3.3.1 RNA-seq read mapping

Single-end Illumina RNA-seq read data was mapped to the *Arabidopsis* (TAIR10) and *Arabis* (Version 4) reference genome assembly ((Lamesch et al., 2012); (Willing et al., 2015)) using Bowtie v2.2.1 (Langmead & Salzberg, 2012) and TopHat v2.0.10 (Trapnell et al., 2009). A maximum of two mismatches and minimal anchor length of 10 bp was used for mapping. Cufflinks v2.2.1 (Trapnell et al., 2010) was used on uniquely aligned reads to extract normalized read counts, Fragments Per Kilobase of transcript per Million mapped reads (FPKM) for each gene in each time point sample.

3.3.2 Identification of diurnal genes

Genes with FPKM > 3, for at least one of the 12-time points and a fold change (fc) of larger than 1.5 between the minimum and maximum expression values were used for further analysis. The expressed genes were classified as diurnal, if the expression values of the two consecutive days showed a significant (p-value < 0.05) Pearson correlation coefficient (r) of larger than (or equal to) 0.8. A model based pattern-matching approach (similar to Michael et al., 2008) was performed to assign a time point at one-hour resulution (as transcriptome samples were taken in 4 hours resolution). Sinus curves representing the 24 phases for a day were fit to gene expression data using Pearson correlation. A “best fit” time point was assigned to a gene at a significant (p-value < 0.05) and correlation cutoff of 0.8 with the corresponding sinus pattern.

High confidence diurnal genes

All expressed genes (FPKM > 3 for least one time point) with at least 150% change in minimum and maximum expression (maximum fold change > 1.5) and with high correlation for two day’s expression (r >= 0.8, p-value <0.05) were classified as high confidence diurnal genes.

* r = Pearson correlation coefficient

Low confidence diurnal genes

All expressed genes (FPKM > 0 for least one time point) with sufficient correlation for two day's expression (r >= 0.5, p-value <0.05) were classified as low confidence diurnal genes. This list excludes high confidence diurnal genes.

High confidence non-diurnal genes

All expressed genes (FPKM > 0 for least one time point) with extremely low correlation for two day's expression (r < 0.3, p-value <0.05) were classified as high confidence non- diurnal genes

Low confidence non-diurnal genes

All genes not included in any other category.

3.3.3 Leaf movement analysis

Seeds of the *Arabidopsis* reference lab strain (Col-0) and the *Arabis* accession Pajares were imbibed and stratified for four days in dark at 4°C. Seeds were then sown on standard soil using a completely randomized design. Seedlings were entrained in a controlled environment chamber (Elbanton, Kerkdriel, Netherlands) for two to four days under cool white fluorescent tubes (~100 µmol m^{-2} s^{-1}) in 16:8 hours light/dark and 20:18°C temperature cycles. At the dark : light transition, we transferred the seedlings to an identical chamber set to constant light and temperature (~100 µmol m^{-2} s^{-1} and 25°C) and started the image capture. Pictures of seedlings were taken at an interval of 20 minutes during four days of constant conditions using Pentax Optio WG-1 digital cameras triggered by their internal intervalometers. Estimates for the vertical movement of the seedlings were obtained by employing the automated leaf movement analysis program TRiP (Greenham, Lou, Remsen, Farid, & McClung, 2015). Estimates for the circadian period, phase and relative amplitude error (RAE) per plant were obtained via fast Fourier transform nonlinear least-squares analysis using the biological rhythms analysis software system BRASS (available on http://millar.bio.ed.ac.uk). Following the common practice, first 24 hours were excluded from the analysis to remove potential noise caused by the transfer from the entrainment chamber to the imaging chamber. The RAE, estimated by BRASS, is a measure of the robustness of a rhythm and can theoretically have values between 0 and 1, where a value of 0 indicates a perfect rhythm and a value of 1 a rhythm that is not statistically significant2. Only seedlings with RAE values below 0.25 were further analyzed. Additionally, period outliers were removed defined as seedling with period values smaller than 5 % or greater than 95 % quantile.

3.3.4 Defining regulatory regions for eight Brassicaceae genomes

Repeat masking for intergenic sequences was performed using RepeatMasker V3.3.0 (Smit AFA, Hubley, R & Green, P., http://www.repeatmasker.org) for masking *Arabidopsis*-specific repeats.

3.3.5 Identification of conservation block and DNA motifs

Orthologous gene identification for all species in the pairwise manner was done separately at protein level using reciprocal best hits using blastall v2.2.25 (Altschul et al., 1990). *Arabidopsis* and *Arabis* were taken as base genome while identifying orthologs in other seven Brassicaceae. Conserved blocks were generated comparing intergenic sequences of orthologous gene of eight Brassicaceae species with PHAST package ((Siepel et al., 2005)). All conservation blocks with length > 10bp and PhastCons score > 0 were used for all genes in all species. DNA motifs were identified using de novo motif identification tool, MEME (Bailey & Elkan, 1994) with following parameter: model: anr (any number of repeats), motif size: 5-15, e-value: < 0.01, reverse compliment =True, number of motifs: 10 motifs (maximum). For motif identification, always a base genome was selected (*Arabidopsis* or *Arabis*), and motifs with no instances in base genome were discarded. To define final set of motifs, all motif clusters with several motifs efficiently filtered for false positives.

3.3.6 Comparison of DNA motifs for both species

Complete set of DDMs (91 for *Arabidopsis* and 92 *Arabis*) was further filtered to get a set of tpsDDM (54 for *Arabidopsis* and 45 for *Arabis*, on the basis of enrichment variability in time point clusters (cov >= 0.15)). For comparability, similar tpsDDMs in *Arabidopsis* and *Arabis*, all tpsDDMs were clustered using MatAlign v-4a (http://stormo.wustl.edu/MatAlign/) followed by hierarchal clustering using hclust function implemented in R.

3.3.7 Motif co-occurrence analysis

All tpsDDMs were used as models of TF-binding specificity. I screened all, motif – motif pairs for co-occurring instances within a region length <= 50 bps. I allowed partially overlapping motif dimers (allowed offset >= 5 for motif length 7 – 10 bp). I considered only the co-occurring motifs within up to 50 bps allowed spacing between the two motifs. For each time point gene cluster, motif pair (*M1, M2*) significance was calculated as follows. First, matches to individual motifs were identified within 1.5kb upstream region for genes. Pairs of motif matches that fit within the specified length (50bp*)* were taken as instances of the motif pair.

Let *O12* and *o12* be the number of observed motif pair occurrences (within 50 bp) in a given time point gene cluster (foreground) and in the background set of all diurnal genes from other time point clusters. Also, let *A12* and *a12* be the number of all possible motif pair occurrences in the foreground and the background, respectively. Possible occurrence of the motif pair means, any occurrence such that the whole pair resides within the 1.5kb region. Then *f12 = O12/A12* is the probability of observing in the foreground the pair of motifs (*M1, M2*). Likewise, *b12 = o12/a12* is the probability of observing in the background the same pair of motifs *(M1, M2)*. I defined the null hypothesis, as the foreground probability *f12* and the background probability *b12* are the same. Consequently, the p-value of observing in the foreground at least *O12* occurrences of the motif pair was calculated as the probability of observing at least *O12* successes in *A12* trials.

3.3.8 Analysis of DNase I data

To generate DNase I activity patterns around motif instances, I identified all instances of these tpsDDMs using MOODS tool (Korhonen et al., 2009) with p-value <0.001 in footprint regions in leaf tissue (Zhang et al., 2012) and aligned the sequences using the motif as the center and included 50 bp of flanking sequence both sides for each tpsDDM. I counted the numbers of DNase I cut (number of reads starting from a particular position) for each position using the DNase-seq read data using a custom shell script.

Chapter 4

Discussion

4.1 Annotating coding regions of genome

Despite over two decades of continuous research efforts, accurate annotation of protein coding genes is still a challenging task. Manually annotated genes and availability of some experimental data (ESTs, proteins) along with homology searches have helped improving gene prediction accuracy. With advancements in high-throughput RNA sequencing and its applicability to genome annotation helps improving gene annotation even further.

4.1.1 Improving the annotation of *A. lyrata*

Incomplete and fragmented assemblies can lead to under and over-predication (in the case of genes split to different contigs) of gene number. Improved sequencing technology enables better assemblies and thus better annotation. Novel sequencing strategies, like single-molecule real-time (SMRT) sequencing developed by Pacific Biosciences with more than 10kb of read length, seem promising for updating draft assemblies and better genome annotations (Eid et al., 2009). However, perhaps the availability of expression evidence is more important than the quality of assembly. One requirement for this is to generate enough species-specific expression data. Shotgun sequencing of entire transcriptomes was enabled by the invention of RNA-seq a couple of years ago (Mortazavi et al., 2008). Also, recent sequencing technologies such as PacBio's Isoform sequencing (Iso-seq) developed full-length transcript sequencing possibilities. Full-length transcript information facilitates elucidation of complex gene structures with improved resolution to generate accurate gene models ((Sharon et al., 2013); (Thomas, Underwood, Tseng, & Holloway, 2014); (Minoche et al., 2015)). This can partly overcome the limitation of current prediction tools for predicting gene with short exons, several exons, long introns or many isoforms with better accuracies, which are challenging to address even with shotgun RNA sequencing data.

The improver annotation (version-2) of *A. lyrata* includes updates on ~90% of the gene models, added TE-related genes, updates the list of non-coding RNAs, and over 2,000 genes with alternate transcripts. In version-1, 33% of all predicted genes were annotated without UTRs information, however, in version-2 only 5% remained without annotated UTR. Moreover, the version-2 introduced splice variants, which were previously missing in the version-1 annotation. In all, the updated annotation includes 31,132 gene models (35,805 transcripts), which is little lower than previous estimates One of the major features of version-2 annotation is the identification and correction of several hundreds of wrongly-split or wrongly-merged gene models.

4.2 Annotating non-coding regions of genome

Unlike the coding regions of genomes where experimental evidence of expression can be utilized for annotation, annotation of non-coding regions is not straightforward. Moreover, the space of non-coding regions is typically much larger compared to the coding regions and lacks clear signature of functional sequences.

Generally, computational approaches for the identification of CREs include identification of over-enriched sequences in promoters of co-expressed or homologous genes from closely related species ((Mikkelsen & Thomashow, 2009); (Covington et al., 2008); (Mockler et al., 2007); (Filichkin et al., 2010); (Filichkin et al., 2011); (Berns et al., 2014)). Using co-expressed genes, finding overrepresented instances of DNA motifs is complicated by the fact that not all co-expressed genes are co-regulated. Second, due to combinatorial nature of TFs, same DNA motif sites may be present in genes that are not co expressed, or co–regulated. This can affect motif identification using discriminative approaches. On the contrary using sets of orthologs might suffer from false ortholog identification and matching of evolutionary unrelated promoter regions.

However, even more challenging than the identification is the validation/confirmation of computationally predicted motifs. While experimental approaches, which target individual motifs are rather tedious, we tried to utilize high-throughput sequencing methods revealing global binding of proteins to DNA. If regulatory proteins bind to DNA, the bound region will be protected from DNase I cleavage, which nonspecifically digests unbound DNA (Galas & Schmitz, 1978). Protected DNA can be identified genome-wide by quantitative high-throughput sequencing of partially digested DNase I chromatin ((Hesselberth et al., 2009); (Boyle et al., 2011)).

I have re-analyzed publicly available DNase-seq data following state-of-the-art analysis pipelines in order to confirm some of the predicated binding motifs. Although, there are several tools that can be used for DNA footprints such as FootprintMixture (Yardımcı et al., 2014), CENTIPEDE (Pique-Regi et al., 2011) and DNaseR (Madrigal, 2013)) but nucleotide bias in DNase I cutting frequencies is not addressed at all. This bias complicate the analysis and interpretation of the putative binding patterns ((He et al., 2014); (Koohy et al., 2013); (Raj & McVicker, 2014); (Sung et al., 2014); (Rusk, 2014); (Madrigal, 2015)). Development of computational tool(s) that can correct for sequence-specific bias by improved training of predictive models will still be required to make full use of these data.

4.3 Diurnal expression and its regulation

Analysis of diurnal expression in two Brassicaceae revealed a phase shift in many key clock genes and hundreds of their targets. It has been shown in circadian clock mutants of several crop plants that genetic determinant underlies the change in circadian clock functioning at least in case of flowering time and photoperiod sensitivity ((Turner, 2005); (Murphy et al., 2011); (Pin et al., 2012); (Zakhrabekova et al., 2012)). This compelled us to speculate that shifted phase of the circadian clock in *Arabis* might have a genetic basis. Although, external cue (light and temperature) reset the clock during the start of each day but shifted expression of clock genes already affect the expression of hundreds of downstream genes before getting reset (Chapter 3).

A Higher proportion of genes associated with light regulation in the subset of diurnal genes with least phase shift also supported our hypothesis that light might play a role in diluting the phase shift shown by *Arabis* diurnal genes. It has been shown that naturally occurring variations can also affect circadian clock functioning to provide fitness benefit in flies (Joshi & Gore, 1999), in *Arabidopsis* (Michael et al., 2003)) and to more recently in Tomato (Müller et al., 2015). Though, I do not show any data on fitness benefit to *Arabis* with shifted clock; this might be an interesting aspect to explore in future studies if there is any (fitness or other) benefit of this phase shift.

The high fraction (around 60% for both species) of common diurnal genes confirmed high expression conservation between *Arabidopsis* and *Arabis*. CRE identification in diurnal genes revealed a much larger set of diurnal CREs (Atlas of DDM) compared to what was known previously. Even with genome-wide data analysis previous studies resulted in only a handful CREs ((Mockler et al., 2007); (Michael et al., 2008); (Covington et al., 2008); (Filichkin et al., 2011)). Although, a similar number of diurnal genes identified in several previous studies but the way CRE identification was performed might underlies this difference in output. CRE identification was done within single species for co-expressed genes and promoter region close to TSS (generally 500 bp upstream) was explored for CRE identification. These two differences, using promoter conservation information (through Phylogenetic shadowing) and exploring complete intergenic region might help us to find more CREs.

The substantial number of common CREs (around 65%) between both species indicated similar regulation of diurnal expression in both species. Several conserved co-occurring DDMs also suggested the presence of conserved diurnal *cis*-regulatory modules (CRMs). These conserved diurnal *cis*-regulatory modules (dCRMs) include previously identified CREs such as ME, EE, GBX and several other novels CREs. Two conserved CRMs predicted in this study (EE-ABREL and EE-EE) have already identified to regulated expression of a GI gene in *Arabidopsis* (Berns et al., 2014). These findings not only suggest that these CRMs might be important for several other *Arabidopsis* genes but also extend their importance to *Arabis* and other Brassicaceae.

Despite conservative selection pressure on functional TFBS ((Dawid, 2006); (Schmidt et al., 2010); (Q. He et al., 2011); (Haudry et al., 2013)), TFBS can be gained and lost rapidly leading to CRE pattern differences among closely related species even within short timescale ((Borneman et al., 2007); (Doniger & Fay, 2007); (Dowell, 2010); (Villar et al., 2014); (Schaefke et al., 2015)). Orthologs of over 60% of the genes with diurnal expression in *Arabidopsis* and *Arabis* were also diurnally expressed but around 10% of these genes had orthologs with the clear absence of diurnal expression. The promoters of these genes featured an underrepresentation of DDMs as compared to diurnal genes (with diurnal orthologs). This underrepresentation does not necessarily imply causality but offers an interesting possibility of a connection between these two observations.

Appendix

Appendix I: Short read mapping information (read numbers in millions)

This RNA-seq data was used for re-annotation of *A. lyrata* genome to generate version-2 annotation.

Tissue	**Sample**	**Read length (bp)**	**Raw reads**	**Reads aligned uniquely**	**Reads aligned uniquely (spliced alignments)**
Rosette (WT)	Rep 1	96	16.0	8.3	3.0
Rosette (WT)	Rep 2	96	12.0	6.0	2.7
Rosette (Heat stressed)	Rep 1	96	17.6	9.7	3.4
Rosette (Heat stressed)	Rep 2	96	5.9	3.1	1.2
Rosette (Recovered)	Rep 1	96	15.6	5.6	4.1
Rosette (Recovered)	Rep 2	96	5.8	2.8	1.3
Shoot apical meristem (WT)	Rep 1	101	14.5	7.5	3.6
Rosette (WT)	Rep 1	101	19.6	9.3	5.6
Rosette (WT)	Rep 2	101	18.1	9.3	4.8
Inflorescence (WT)	Rep 1	75	32.0	12.9	9.3
Inflorescence (WT)	Rep 2	75	32.0	12.6	9.0
Rosette (WT and cold stressed)	Rep 1	75-100	102.7	59.7	24.6
			291.8	146.8	72.6

Appendix II: Primer information

Information about primers used in this study to validate split and merged gene is online available at **S1 Dataset** cited in paper:

Improving the annotation of *Arabidopsis lyrata* using RNA-Seq Data. *Plos One*, *10*(9), e0137391. doi:10.1371/journal.pone.0137391

(http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137391#pone.0137391.s001)

Appendix III:

This time-series RNA-seq data was used for identification of diurnal genes in *Arabidopsis* and *Arabis*

Table: Short read mapping information

#	Species	Time point	Read length	Total reads (Millions)	Mapped reads (%)	Uniquely mapped reads (%)	Unmapped reads (%)
1	At	ZT0	97	60.1	72.3	64.3	27.7
2	At	ZT4	97	54.6	76.5	65.5	23.5
3	At	ZT8	97	46.6	76.9	66.4	23.1
4	At	ZT12	97	48.8	75.4	67.4	24.6
5	At	ZT16	97	50.7	69.3	62.1	30.7
6	At	ZT20	97	51.1	75.4	67.7	24.6
7	At	ZT24	97	49.5	76.7	67.9	23.3
8	At	ZT28	97	47.8	79.2	68.3	20.8
9	At	ZT32	97	44.0	78.7	68.2	21.3
10	At	ZT36	97	52.8	74.7	66.4	25.3

11	At	ZT40	97	59.1	74.4	66.3	25.6
12	At	ZT44	97	53.4	76.9	68.6	23.1
13	Aa	ZT0	97	43.7	76.4	64.5	23.6
14	Aa	ZT4	97	47.2	76.0	63.9	24.0
15	Aa	ZT8	97	79.2	72.1	60.7	27.9
16	Aa	ZT12	97	84.0	85.9	73.5	14.1
17	Aa	ZT16	97	53.0	88.4	75.9	11.6
18	Aa	ZT20	97	65.8	86.4	74.3	13.6
19	Aa	ZT24	97	59.3	89.1	75.7	10.9
20	Aa	ZT28	97	50.7	88.3	74.4	11.7
21	Aa	ZT32	97	71.4	80.2	68.0	19.8
22	Aa	ZT36	97	66.0	80.2	68.2	19.8
23	Aa	ZT40	97	67.6	89.6	76.9	10.4
24	Aa	ZT44	97	71.6	89.8	77.2	10.2

Aa: *Arabis* samples

At: *Arabidopsis* samples

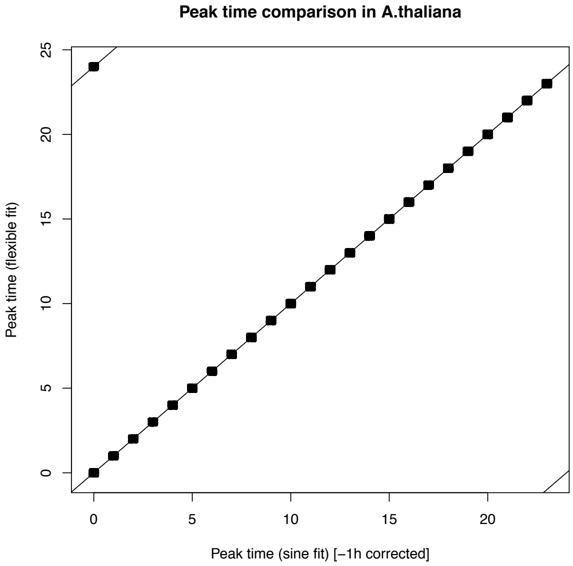
Appendix IV:

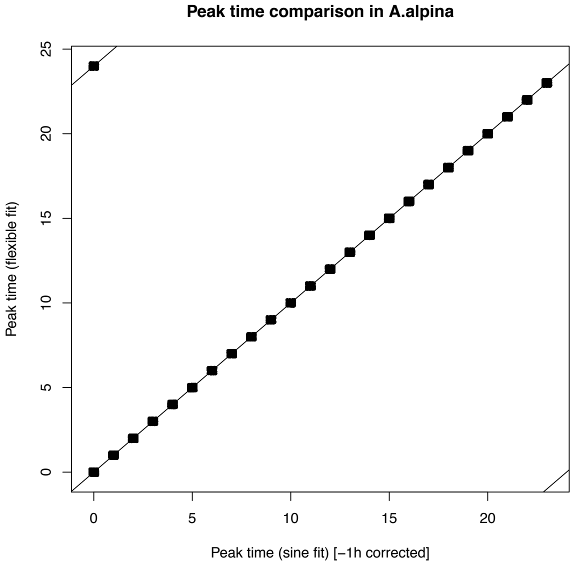
Ortholog information for diurnal genes in *Arabidopsis* and *Arabis*

	Arabidopsis	***Arabis***
Diurnal genes (hc)	7,702	8,517
Diurnal genes (lc)	1,873	2,782
Non-diurnal genes (hc)	6,182	7,129
Non-diurnal genes (lc)	11,562	12,282
Diurnal gene (hc) with orththolo	5,422	5,752
Diurnal gene (hc) with diurnal (hc) orthologs	2,751	2,835
Diurnal gene (hc) with diunal (lc) orththolo	606	308
Diurnal gene (hc) with non-diunal orththolo (hc)	479	615

Appendix V:

High agreement between **flexible fit** method and **MoPS** method (Model-based Periodicity Screening) for time point assignment in *Arabidopsis* and *Arabis* high confidence diurnal genes.





Bibliography

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., et al. (2000). The genome sequence of *Drosophila melanogaster*. ***Science***, ***287*****(5461), 2185–2195.** doi:10.1126/science.287.5461.2185

Aerts, S., Van Loo, P., Moreau, Y., & De Moor, B. (2004). A genetic algorithm for the detection of new *cis*-regulatory modules in sets of coregulated genes. ***Bioinformatics***, ***20*****(12), 1974–1976.** doi:10.1093/bioinformatics/bth179

Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., & De Moor, B. (2003). Computational detection of *cis* -regulatory modules. ***Bioinformatics***, ***19 (suppl 2)***, **ii5–ii14**. doi:10.1093/bioinformatics/btg1052

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. ***J Mol Biol.*****, 215(3), 403–410.** doi:10.1006/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. ***Nucleic Acids Res.***, ***25*****(17), 3389–3402.** doi:10.1093/nar/25.17.3389

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000). Gene Ontology: tool for the unification of biology. ***Nat Genet.*****, 25(1), 25–29.** doi:10.1038/75556

Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. ***Proc Int Conf Intell Syst Mol Biol.,*** **2, 28–36.**

Beckstette, M., Homann, R., Giegerich, R., & Kurtz, S. (2006). Fast index based algorithms and software for matching position specific scoring matrices. ***BMC Bioinformatics***, ***7*****, 389.** doi:10.1186/1471-2105-7-389

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S., et al. (2004). Ultraconserved elements in the human genome. ***Science***, ***304*****(5675), 1321–1325**. doi:10.1126/science.1098119

Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., et al. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. ***Proc Natl Acad Sci U S A.***, ***99*****(2), 757–762**. doi:10.1073/pnas.231608898

Berns, M. C., Nordström, K., Cremer, F., Toth, R., Hartke, M., Simon, S., et al. (2014). Evening expression of arabidopsis GIGANTEA is controlled by combinatorial interactions among evolutionarily conserved regulatory motifs. ***Plant Cell.*****, 26(10), 3999–4018.** doi:10.1105/tpc.114.129437

Bläsing, O.E., Gibon, Y.,Günther, M., Höhne, M., Morcuende, R., Osuna, D., et al. (2005). Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in *Arabidopsis***. *Plant Cell.*, 17 (12), 3257-3281.** doi:10.1105/tpc.105.035261.1

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K.D., Ovcharenko, I., Pachter, L., et al. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. ***Science*, 299(5611): 1391-1394.** doi: 10.1126/science.1081331

Borland, A. M., & Taybi, T. (2004). Synchronization of metabolic processes in plants with Crassulacean acid metabolism. ***J Exp Bot.*, *55*(400), 1255–1265.** doi:10.1093/jxb/erh105

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., et al. (2007). Divergence of transcription factor binding sites across related yeast species. ***Science*, 317(5839), 815–819.** doi:10.1126/science.1140748

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., et al. (2008). High-resolution mapping and characterization of open chromatin across the genome. ***Cell*, 132(2), 311–322.** doi:10.1016/j.cell.2007.12.014

Boyle, C. A., Boulet, S., Schieve, L. A., Cohen, R. A., Blumberg, S. J., Yeargin-Allsopp, M., et al. (2011). Trends in the prevalence of developmental disabilities in US children, 1997-2008. ***Pediatrics*, 127(6), 1034–1042.** doi:10.1542/peds.2010-2989

Brent, M. R. (2005). Genome annotation past, present, and future : how to define an ORF at each locus. ***Genome Res*. 15(12), 1777–1786**. doi:10.1101/gr.3866105

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. ***J Mol Biol.*, *268*(1), 78–94.** doi:10.1006/jmbi.1997.0951

Cao, J., Schneeberger, K., Ossowski, S., Günther, T., Bender, S., Fitz, J., et al. (2011). Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. ***Nat Genet.*, *43*(10), 956–963.** doi:10.1038/ng.911

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., et al. (2003). Finding functional features in Saccharomyces genomes by phylogenetic footprinting. ***Science*, *301*(5629), 71–76.** doi:10.1126/science.1084337

Covington, M. F., & Harmer, S. L. (2007). The circadian clock regulates auxin signaling and responses in *Arabidopsis*. ***PLoS Biol.*, *5*(8), e222**. doi:10.1371/journal.pbio.0050222

Covington, M. F., Maloof, J. N., Straume, M., Kay, S. A, & Harmer, S. L. (2008). Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. ***Genome Biol.*, *9*(8), R130.** doi:10.1186/gb-2008-9-8-r130

Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., Blakesley, R., Bouffard, G., et al. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. ***Proc Natl Acad Sci U S A.,* 101(4), 992–997.** doi:10.1073/pnas.0307540100

D'Hont, A., Denoeud, F., Aury, J.-M., Baurens, F.-C., Carreel, F., Garsmeur, O., et al. (2012). The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. ***Nature*, *488*(7410), 213–217.** doi:10.1038/nature11241

Darwish, O., Shahan, R., Liu, Z., Slovin, J. P., & Alkharouf, N. W. (2015). Re-annotation of the woodland strawberry (*Fragaria vesca*) genome. ***BMC Genomics*, *16*, 29.** doi:10.1186/s12864-015-1221-1

Dassanayake, M., Oh, D.H., Haas, J. S., Hernandez, A., Hong, H., Ali, S., et al. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. ***Nat Genet.*, *43*(9), 913–918.** doi:10.1038/ng.889

Dawid, I. B. (2006). The regulatory genome, by Eric H. Davidson (2006), Academic Press. ***The FASEB Journal*, *20*(13), 2190–2191**. doi:10.1096/fj.06-1103ufm

Deng, W., Ying, H., Helliwell, C. A., Taylor, J. M., Peacock, W. J., & Dennis, E. S. (2011). FLOWERING LOCUS C (FLC) regulates development pathways throughout the life cycle of *Arabidopsis*. ***Proc Natl Acad Sci U S A.*, *108*(16), 6680–6685.** doi:10.1073/pnas.1103175108

Doniger, S. W., & Fay, J. C. (2007). Frequent gain and loss of functional transcription factor binding sites. ***PLoS Comput Biol.*, *3*(5), e99.** doi:10.1371/journal.pcbi.0030099

Dowell, R. D. (2010). Transcription factor binding variation in the evolution of gene regulation. ***Trends Genet.*, *26*(11), 468–475.** doi:10.1016/j.tig.2010.08.005

Du, Z., Zhou, X., Ling, Y., Zhang, Z., & Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. ***Nucleic Acids Res.*, *38*(Web Server issue), W64–W70.** doi:10.1093/nar/gkq310

Dunlap, J. C. (1999). Molecular bases for circadian clocks. ***Cell*, *96*(2), 271–290.** doi:10.1016/S0092-8674(00)80566-8

Earl, D., Bradnam, K., St. John, J., Darling, A., Lin, D., Fass, J., et al. (2011). Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. ***Genome Res.*, *21*(12), 2224–2241**. doi:10.1101/gr.126599.111

Eckalbar, W. L., Hutchins, E. D., Markov, G. J., Allen, A. N., Corneveaux, J. J., Lindblad-Toh, K., et al. (2013). Genome reannotation of the lizard *Anolis carolinensis* based on 14 adult and embryonic deep transcriptomes. ***BMC Genomics*, *14*, 49.** doi:10.1186/1471-2164-14-49

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. ***Science* , *323*(5910), 133–138.** doi:10.1126/science.1162986

Eser, P., & Tresch, A. (2014). MoPS: MoPS - Model-based Periodicity Screening. **R package version 1.5.0.**

Eversole, K., Feuillet, C., Mayer, K.F.X., Rogers, J. (2014). Slicing the wheat genome. ***Science*, 345(6194), 285-287.** doi:10.1126/science.1257983

Fahlgren, N., Jogdeo, S., Kasschau, K. D., Sullivan, C. M., Chapman, E. J., Laubinger, S., et al. (2010). MicroRNA gene evolution in *Arabidopsis lyrata* and *Arabidopsis thaliana*. ***Plant Cell.*, *22*(4), 1074–1089**. doi:10.1105/tpc.110.073999

Farré, E. M., & Weise, S. E. (2012). The interactions between the circadian clock and primary metabolism. ***Curr Opin Plant Biol.*, *15*(3), 293–300.** doi:10.1016/j.pbi.2012.01.013

Filichkin, S. a, Breton, G., Priest, H. D., Dharmawardhana, P., Jaiswal, P., Fox, S. E., et al. (2011). Global profiling of rice and poplar transcriptomes highlights key conserved circadian-controlled pathways and *cis*-regulatory modules. ***PloS One*, *6*(6), e16907.** doi:10.1371/journal.pone.0016907

Filichkin, S. A, Priest, H. D., Givan, S. A, Shen, R., Bryant, D. W., Fox, S. E., et al. (2010). Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. ***Genome Res.*, 20(1), 45–58.** doi:10.1101/gr.093302.109

Finn, R. D., Mistry, J., Schuster-böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., et al. (2006). Pfam : clans, web tools and services. ***Nucleic Acids Res.*, *34(Database issue)*, D247–251.** doi:10.1093/nar/gkj149

Fonseca, J. P., Menossi, M., Thibaud-Nissen, F., & Town, C. D. (2010). Functional analysis of a TGA factor-binding site located in the promoter region controlling salicylic acid-induced NIMIN-1 expression in *Arabidopsis*. ***Genet Mol Res.*, *9*(1), 167–175.** doi:10.4238/vol9-1gmr704

F**owler, S. G., Cook, D., & Thomashow, M. F.** (2005). Low temperature induction of *Arabidopsis CBF1, 2,* and *3* is gated by the circadian clock. ***Plant Physiol.*, *137*(3), 961–968.** doi: 10.1104/pp.104.058354

Frith, M. C., Fu, Y., Yu, L., Chen, J.F., Hansen, U., Weng, Z. (2004). Detection of functional DNA motifs via statistical over-representation. ***Nucleic Acids Res.*, *32*(4), 1372–1381.** doi:10.1093/nar/gkh299

Frith, M. C., Hansen, U., & Weng, Z. (2002). A gibbs sampling algorithm to detect clustered *cis*-elements , ***Regulatory Genomic Sequences*, BGRS 2002, 61–63.**

Frith, M. C., Li, M. C., & Weng, Z. (2003). Cluster-Buster: Finding dense clusters of motifs in DNA sequences. ***Nucleic Acids Res.*, *31*(13), 3666–3668.** doi:10.1093/nar/gkg540

Galas, D. J., & Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. ***Nucleic Acids Res.*****,** ***5*****(9), 3157–3170.** doi:10.1093/nar/5.9.3157

Gan, X., Stegle, O., Behr, J., Steffen, J. G., Drewe, P., Hildebrand, K. L., et al. (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. ***Nature*****,** ***477*****(7365), 419–423.** doi:10.1038/nature10414

Gao, Z., Zhao, R., & Ruan, J. (2013). A genome-wide *cis*-regulatory element discovery method based on promoter sequences and gene co-expression networks. ***BMC Genomics*****,** ***14*** **(Suppl 1), S4.** doi:10.1186/1471-2164-14-S1-S4

Gendron, J. M., Pruneda-Paz, J. L., Doherty, C. J., Gross, A. M., Kang, S. E., & Kay, S. A. (2012). *Arabidopsis* circadian clock protein, TOC1, is a DNA-binding transcription factor. ***Proc Natl Acad Sci U S A.*****,** ***109*****(8), 3167–3172.** doi:10.1073/pnas.1200355109

Genome 10K Community of Scientists. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. ***J Hered.*****,** ***100*****(6), 659–674.** doi:10.1093/jhered/esp086

Giuliano, G., Pichersky, E., Malik, V. S., Timko, M. P., Scolnik, P. A., & Cashmore, A. R. (1988). An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. ***Proc Natl Acad Sci U S A.*****,** ***85*****(19), 7089–7093.** doi:10.1073/pnas.85.19.7089

Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M., Brockman, W., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. ***Nat Biotechnol.*****,** ***27*****(2), 182–189.** doi:10.1038/nbt.1523

Goff, S. A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. *Japonica*). ***Science*****,** ***296*****(5565), 92–100.** doi:10.1126/science.1068275

Graf, A., Schlereth, A., Stitt, M., & Smith, A. M. (2010). Circadian control of carbohydrate availability for growth in *Arabidopsis* plants at night. ***Proc Natl Acad Sci U S A., 107(20),*** **9458-9463.**doi:10.1073/pnas.0914299107

Greenham, K., & Mcclung, C. R. (2015). Integrating circadian dynamics with physiological processes in plants. ***Nat Rev Genet.*****,** ***16*****(10), 598–610.** doi:10.1038/nrg3976

Greenham, K., Lou, P., Remsen, S. E., Farid, H., & McClung, C. R. (2015). TRiP: Tracking rhythms in plants, an automated leaf movement analysis program for circadian period estimation. ***Plant Methods*****,** ***11*****(1), 33.** doi:10.1186/s13007-015-0075-5

Gregis, V., Andrés, F., Sessa, A., Guerra, R. F., Simonini, S., Mateos, J. L., et al. (2013). Identification of pathways directly regulated by SHORT VEGETATIVE PHASE during vegetative and reproductive development in *Arabidopsis*. ***Genome Biol.*****,** ***14*****(6), R56.** doi:10.1186/gb-2013-14-6-r56

Guigó, R., Flicek, P., Abril, J. F., Reymond, A., Lagarde, J., Denoeud, F., et al. (2006). EGASP: the human ENCODE genome annotation assessment project. ***Genome Biol.***, *7* **(Suppl 1), S2.** doi:10.1186/gb-2006-7-s1-s2

Ha, N., Polychronidou, M., & Lohmann, I. (2012). COPS: detecting co-occurrence and spatial arrangement of transcription factor binding motifs in genome-wide datasets. ***PloS One***, *7***(12), e52055.** doi:10.1371/journal.pone.0052055

Harmer, S. L. (2009). The circadian system in higher plants. ***Annu Rev Plant Biol.***, ***60*(1), 357–377.** doi:10.1146/annurev.arplant.043008.092054

Harmer, S. L., Hogenesch, J. B., Straume, M., Chang, H. S., Han, B., Zhu, T., et al. (2000). Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock. ***Science***, ***290*(5499), 2110–2113.** doi:10.1126/science.290.5499.2110

Haudry, A., Platts, A. E., Vello, E., Hoen, D. R., Leclercq, M., Williamson, R. J., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. ***Nat Genet.***, ***45*(8), 891–898.** doi:10.1038/ng.2684

Haydon, M. J., Mielczarek, O., Robertson, F. C., Hubbard, K. E., & Webb, A. A. R. (2013). Photosynthetic entrainment of the *Arabidopsis thaliana* circadian clock. ***Nature***, ***502*(7473), 689–692.** doi:10.1038/nature12603

He, H. H., Meyer, C. A., Hu, S. S., Chen, M.W., Zang, C., Liu, Y., et al. (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. ***Nat Methods.***, ***11*(1), 73-78.** doi:10.1038/nmeth.2762

He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., et al. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. ***Nat Genet.***, ***43*(5), 414–420.** doi:10.1038/ng.808

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., et al. (2009). Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. ***Nat Methods.***, ***6*(4), 283–289.** doi:10.1038/nmeth.1313

Hong, R. L., Hamaguchi, L., Busch, M. A., & Weigel, D. (2003). Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. ***Plant Cell.***, ***15*(6), 1296–1309.** doi: 10.1105/tpc.009548

Hsu, P. Y., & Harmer, S. L. (2014). Wheels within wheels: the plant circadian system. ***Trends Plant Sci.***, ***19*(4), 240–249.** doi:10.1016/j.tplants.2013.11.007

Hu, T. T., Pattyn, P., Bakker, E. G., Cao, J., Cheng, J.F., Clark, R. M., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. ***Nat Genet.***, ***43*(5), 476–481.** doi:10.1038/ng.807

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., et al. (2009). High-throughput genotyping by whole-genome resequencing. ***Genome Res., 19,*** **1068–1076.** doi:10.1101/gr.089516.108.

Hubisz, M. J., Pollard, K.S., Siepel, A. (2011). PHAST and RPHAST: phylogenetic analysis with space/time models. ***Brief Bioinform.,*** **12(1), 41-51.** doi: 10.1093/bib/bbq072

Hudson, M. E., & Quail, P. H. (2003). Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. ***Plant Physiol.***, ***133*****(4), 1605–1616.** doi: 10.1104/pp.103. 030437

Hunkapiller, T., Kaiser, R.J., Koop, B.F., Hood, L. (1991). Large-scale and automated DNA sequence determination. ***Science***, **254(5028), 59-67.** doi: 10.1126/science.1925562

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. ***Nucleic Acids Res.***, ***37*****(Database issue), D211–D215.** doi:10.1093/nar/gkn785

Hupalo, D., & Kern, A. D. (2013). Conservation and functional element discovery in 20 angiosperm plant genomes. ***Mol Biol Evol.***, ***30*****(7), 1729–1744.** doi:10.1093/molbev/mst082

International Barley Genome Sequencing Consortium, Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P. (2012). A physical, genetic and functional sequence assembly of the barley genome. ***Nature***, ***491*****(7426), 711–716.** doi:10.1038/nature11543

International Rice Genome Sequencing Project (2005). The map-based sequence of the rice genome. ***Nature***, ***436*****(7052), 793–800.** doi:10.1038/nature03895

Iqbal, Z., Caccamo, M., Turner, I., Flicek, P., & McVean, G. (2012). *De novo* assembly and genotyping of variants using colored de Bruijn graphs. ***Nat Genet.***, ***44*****(2), 226–232.** doi:10.1038/ng.1028

Jiang, H., & Wong, W. H. (2008). SeqMap: Mapping massive amount of oligonucleotides to the genome. ***Bioinformatics***, ***24*****(20), 2395–2396.** doi:10.1093/bioinformatics/btn429

Johnson, D. S., Mortazavi, A., Myers, R. M., Wold, R. (2007). Genome-wide mapping of *in vivo* protein-DNA Interactions. ***Science, 316*****(*****5830*****), 1497–1502.** doi: 10.1126/science.1141319

Johnston, J. S., Pepper, A. E., Hall, A. E., Chen, Z. J., Hodnett, G., Drabek, J., et al. (2005). Evolution of genome size in Brassicaceae. ***Ann Bot.***, ***95*****(1), 229–235.** doi:10.1093/aob/mci016

Joshi, D. S., & Gore, A. P. (1999). Latitudinal variation in eclosion rhythm among strains of *Drosophila ananassae*. ***Indian J Exp Biol.***, ***37*****(7), 718–724.**

Kaufmann, K., Muiño, J. M., Jauregui, R., Airoldi, C. A., Smaczniak, C., Krajewski, P., & Angenent, G. C. (2009). Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the *Arabidopsis* flower. ***PLoS Biol.***, ***7*(4), e1000090.** doi:10.1371/journal.pbio.1000090

Kaufmann, K., Pajoro, A., & Angenent, G. C. (2010). Regulation of transcription in plants: mechanisms controlling developmental switches. ***Nat Rev Genet.***, ***11*(12), 830–842.** doi:10.1038/nrg2885

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. ***Nature***, ***423*(6937), 241–54.** doi:10.1038/nature01644

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. ***Genome Res.***, **12(4), 656–664.** doi:10.1101/gr.229202.

Kitashiba, H., Li, F., Hirakawa, H., Kawanabe, T., Zou, Z., Hasegawa, Y. (2014). Draft sequences of the Radish (*Raphanus sativus* L.) genome. ***DNA Res., 21(5)*, 481-490.** doi: 10.1093/dnares/dsu014

Kohany, O., Gentles, A. J., Hankus, L., & Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC* ***Bioinformatics***, *7*, **474.** doi:10.1186/1471-2105-7-474

Koohy, H., Down, T. A., & Hubbard, T. J. (2013). Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. ***PLoS ONE***, ***8*(7), e69853.** doi:10.1371/journal.pone.0069853

Kooiker, M., Airoldi, C. A., Losa, A., Manzotti, P. S., Finzi, L., Kater, M. M., & Colombo, L. (2005). BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic *Arabidopsis* gene SEEDSTICK. ***Plant Cell.***, ***17*(3), 722–729.** doi: 10.1105/tpc.104.030130

Koornneef, M., & Meinke, D. (2010). The development of *Arabidopsis* as a model plant. ***Plant J.***, ***61*(6), 909–921.** doi:10.1111/j.1365-313X.2009.04086.x

Korf, I. (2004). Gene finding in novel genomes. ***BMC Bioinformatics***, ***5*, 59.** doi:10.1186/1471-2105-5-59

Korhonen, J., Martinmäki, P., Pizzi, C., Rastas, P., & Ukkonen, E. (2009). MOODS: fast search for position weight matrix matches in DNA sequences. ***Bioinformatics***, ***25*(23), 3181–3182.** doi:10.1093/bioinformatics/btp554

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. ***Nucleic Acids Res.***, ***40*(Database issue), D1202–D1210.** doi:10.1093/nar/gkr1090

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. ***Nat Methods.***, ***9*****(4), 357–359.** doi:10.1038/nmeth.1923

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. ***Science,*** **262(5131), 208-214.** doi:10.1126/science.8211139

Lee, J., He, K., Stolc, V., Lee, H., Figueroa, P., Gao, Y., et al. (2007). Analysis of transcription factor HY5 genomic binding sites revealed its hierarchical role in light regulation of development. ***Plant Cell.***, ***19*****(3), 731–749.** doi: 10.1105/tpc.106.047688

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. ***Genome Res.***, ***18*****(11), 1851–1858.** doi:10.1101/gr.078212.108

Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. ***Bioinformatics***, ***24*****(5), 713–714.** doi:10.1093/bioinformatics/btn025

Li, Z., Zhang, Z., Yan, P., Huang, S., Fei, Z., & Lin, K. (2011). RNA-Seq improves annotation of protein-coding genes in the cucumber genome. ***BMC Genomics***, ***12*****(1), 540.** doi:10.1186/1471-2164-12-540

Liefooghe, A., Touzet, H., & Varre, J.S. (2009). Self-overlapping occurrences and Knuth-Morris-Pratt algorithm for weighted matching. ***In Proceedings of Third International Conference on Language and Automata Theory and Applications (LATA)***, **5457, 481–492**

Lin, H., Zhang, Z., Zhang, M. Q., Ma, B., & Li, M. (2008). ZOOM! Zillions of oligos mapped. ***Bioinformatics***, ***24*****(21), 2431–2437.** doi:10.1093/bioinformatics/btn416

Liu, L., White, M. J., & MacRae, T. H. (1999). Transcription factors and their genes in higher plants. ***Eur J Biochem.***, **262(2), 247–257.** doi: 10.1046/j.1432-1327.1999.00349.x

Liu, S., Lorenzen, E. D., Fumagalli, M., Li, B., Harris, K., Xiong, Z., et al. (2014). Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. ***Cell***, ***157*****(4), 785–794.** doi:10.1016/j.cell.2014.03.054

Liu, T., Carlsson, J., Takeuchi, T., Newton, L., & Farré, E. M. (2013). Direct regulation of abiotic responses by the *Arabidopsis* circadian clock component PRR7. ***Plant J.***, ***76*****(1), 101–114.** doi:10.1111/tpj.12276

Lobréaux, S., Manel, S., & Melodelima, C. (2014). Development of an *Arabis alpina* genomic contig sequence data set and application to single nucleotide polymorphisms discovery. ***Mol Ecol Resour.***, ***14*****(2), 411–418.** doi:10.1111/1755-0998.12189

Long, Q., Rabanal, F. A., Meng, D., Huber, C. D., Farlow, A., Platzer, A., et al. (2013). Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. ***Nat Genet.***, ***45*****(8), 884–890.** doi:10.1038/ng.2678

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. ***Nucleic Acids Res.***, ***25*****(5), 955–964.** doi: 10.1093/nar/25.5.0955

Lukashin, A. V., & Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. ***Nucleic Acids Res.***, ***26*****(4), 1107–1115.** doi:10.1093/nar/26.4.1107

Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., et al. (2008). Finding and comparing syntenic regions among *Arabidopsis* and the outgroups papaya, poplar, and grape: CoGe with Rosids. ***Plant Physiol.***, ***148*****(4), 1772–1781.** doi:10.1104/pp.108.124867

Lysak, M. A., Koch, M. A., Beaulieu, J. M., Meister, A., & Leitch, I. J. (2009). The dynamic ups and downs of genome size evolution in Brassicaceae. ***Mol Biol Evol.***, ***26*****(1), 85–98.** doi:10.1093/molbev/msn223

Madrigal, P. (2013). DNaseR: DNase I footprinting analysis of DNase-seq data. **R package.**

Madrigal, P. (2015). On accounting for sequence-specific bias in genome-wide chromatin accessibility wxperiments: recent advances and contradictions. ***Front Bioeng Biotechnol.***, **3,144.** doi:10.3389/fbioe.2015.00144

Mamanova, L., Coffey, A. J., Scott, C. E., Kozarewa, I., Turner, E. H., Kumar, A., et al. (2010). Target-enrichment strategies for next- generation sequencing. ***Nat Methods.***, ***7*****(2), 111–118.** doi:10.1038/nmeth.1419

Marcolino-Gomes, J., Rodrigues, F. A., Fuganti-Pagliarini, R., Bendix, C., Nakayama, T. J., Celaya, B., et al. (2014). Diurnal oscillations of soybean circadian clock and drought responsive genes. ***PLoS ONE***, ***9*****(1), e86402.** doi:10.1371/journal.pone.0086402

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. ***Nature***, ***437*****(7057), 376–380.** doi:10.1038/nature03959

Mateos, J. L., Madrigal, P., Tsuda, K., Rawat, V., Richter, R., Romera-Branchat, M., et al. (2015). Combinatorial activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C define distinct modes of flowering regulation in *Arabidopsis*. ***Genome Biol.***, ***16*****(1), 31.** doi:10.1186/s13059-015-0597-1

McClung, C. R. (2006). Plant circadian rhythms. ***Plant Cell.***, ***18*****(4), 792–803.** doi:10.1105/tpc.106.040980

McClung, C. R. (2011). The genetics of plant clocks. *Advances in Genetics* (1st ed., Vol. 74). **Elsevier Inc.** doi:10.1016/B978-0-12-387690-4.00004-0

Mercier, E., Droit, A., Li, L., Robertson, G., Zhang, X., & Gottardo, R. (2011). An integrated pipeline for the genome-wide analysis of transcription factor binding sites from ChIP-Seq. ***PloS One*, *6*(2), e16432.** doi:10.1371/journal.pone.0016432

Meshi, T., & Iwabuchi, M. (1995). Plant transcription factors. ***Plant Cell Physiol.*, *36*(8), 1405–1420.** doi:10.1007/978-1-61779-154-3

Michael, T. P., & McClung, C. R. (2002). Phase-specific circadian clock regulatory elements in *Arabidopsis*. ***Plant Physiol.*, *130*(2), 627–638.** doi:10.1104/pp.004929

Michael, T. P., & McClung, C. R. (2003). Enhancer trapping reveals widespread circadian clock transcriptional control in *Arabidopsis*. ***Plant Physiol.*, *132*(2), 629–639.** doi:10.1104/pp.021006

Michael, T. P., Mockler, T. C., Breton, G., McEntee, C., Byer, A., Trout, J. D., et al. (2008). Network discovery pipeline elucidates conserved time-of-day-specific *cis*-regulatory modules. ***PLoS Genetics*, *4*(2), e14.** doi:10.1371/journal.pgen.0040014

Michael, T.P., Salomé, P.A., Yu, H.J., Spencer, T.R., Sharp, E.L., McPeek, M.A., et al. (2003). Enhanced fitness conferred by naturally occurring variation in the circadian clock. ***Science*, *302*(5647), 1049–1053.** doi:10.1126/science.1082971

Mikkelsen, M. D., & Thomashow, M. F. (2009). A role for circadian evening elements in cold-regulated gene expression in *Arabidopsis*. ***Plant J.*, *60*(2), 328–339.** doi:10.1111/j.1365-313X.2009.03957.x

Minoche, A. E., Dohm, J. C., Schneider, J., Holtgräwe, D., Viehöver, P., Montfort, M., et al. (2015). Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. ***Genome Biol.*, 16, 184.** doi:10.1186/s13059-015-0729-7

Mockler, T. C., Michael, T. P., & Priest, H. D. (2007). The DIURNAL Project : DIURNAL and circadian expression profiling , model-based pattern matching , and promoter analysis. ***Cold Spring Harb Symp Quant Biol.*, 72, 353–363.** doi:10.1101/sqb.2007.72.006

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. ***Nat Methods.*, *5*(7), 621–628.** doi:10.1038/nmeth.1226

Müller, N. A., Wijnen, C. L., Srinivasan, A., Ryngajllo, M., Ofner, I., Lin, T., et al. (2015). Domestication selected for deceleration of the circadian clock in cultivated tomato. ***Nat Genet.*. doi:10.1038/ng.3447**

Murphy, R. L., Klein, R. R., Morishige, D. T., Brady, J. A., Rooney, W. L., Miller, F. R., et al. (2011). Coincident light and clock regulation of pseudoresponse regulator protein 37 (PRR37) controls photoperiodic flowering in sorghum. ***Proc Natl Acad Sci U S A.*, 108(39), 16469–16474.** doi:10.1073/pnas.1106212108

Nagel, D. H., Doherty, C. J., Pruneda-Paz, J. L., Schmitz, R. J., Ecker, J. R., & Kay, S. A. (2015). Genome-wide identification of CCA1 targets uncovers an expanded clock network in *Arabidopsis*. ***Proc Natl Acad Sci U S A.***, ***112*****(34), E4802–E4810.** doi:10.1073/pnas.1513609112

Nakamichi, N., Kiba, T., Kamioka, M., Suzuki, T., Yamashino, T., Higashiyama, T., et al. (2012). Transcriptional repressor PRR5 directly regulates clock-output pathways. ***Proc Natl Acad Sci U S A.***, ***109*****(42), 17123–17128.** doi:10.1073/pnas.1205156109

Nordström, K. J. V., Albani, M. C., James, G. V., Gutjahr, C., Hartwig, B., Turck, F., et al. (2013). Mutation identification by direct comparison of whole-genome sequencing data from mutant and wild-type individuals using k-mers. ***Nat Biotechnol.***, ***31*****(4), 325–330.** doi:10.1038/nbt.2515

Nose, M., & Watanabe, A. (2014). Clock genes and diurnal transcriptome dynamics in summer and winter in the gymnosperm Japanese cedar (*Cryptomeria japonica* (L.f.) D.Don). ***BMC Plant Biol.***, ***14*****(1), 308.** doi:10.1186/s12870-014-0308-1

Ossowski, S., Schneeberger, K., Clark, R. M., Lanz, C., Warthmann, N., & Weigel, D. (2008). Sequencing of natural strains of *Arabidopsis thaliana* with short reads. ***Genome Res.***, ***18*****(12), 2024–2033.** doi:10.1101/gr.080200.108

Østergaard, L., & Yanofsky, M. F. (2004). Establishing gene function by mutagenesis in *Arabidopsis thaliana*. ***Plant J.***, ***39*****(5), 682–696.** doi:10.1111/j.1365-313X.2004.02149.x

Ouyang, S., & Buell, C. R. (2004). The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. ***Nucleic Acids Res.***, ***32*****(Database issue), D360–363.** doi:10.1093/nar/gkh099

Page, D. R., & Grossniklaus, U. (2002). The art and design of genetic screens: *Arabidopsis thaliana*. ***Nat Rev Genet.***, ***3*****(2), 124–136.** doi:10.1038/nrg730

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V, & Grotewold, E. (2006). AGRIS and AtRegNet. a platform to link *cis*-regulatory elements and transcription factors into regulatory networks. ***Plant Physiol.***, ***140*****(3), 818–829.** doi:10.1104/pp.105.072280

Panda, S., Antoch, M. P., Miller, B. H., Su, A. I., Schook, A. B., Straume, M., et al. (2002). Coordinated transcription of key pathways in the mouse by the circadian clock. ***Cell***, ***109*****(3), 307–320.** doi:10.1016/S0092-8674(02)00722-5

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. ***Nat Rev Genet.***, ***10*****(10), 669–680.** doi:10.1038/nrg2641

Pavesi, G., Mauri, G., & Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. ***Bioinformatics***, ***17 Suppl 1***, **S207–214.** doi:10.1093/bioinformatics/17.suppl_1.S207

Pin, P. A., Zhang, W., Vogt, S. H., Dally, N., Büttner, B., Schulze-Buxloh, G., et al. (2012). The role of a pseudo-response regulator gene in life cycle adaptation and domestication of beet. ***Curr Biol.***, ***22*****(12), 1095–1101.** doi:10.1016/j.cub.2012.04.007

Pique-Regi, R., Degner, J. F., Pai, A. A., Gaffney, D. J., Gilad, Y., & Pritchard, J. K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. ***Genome Res.***, ***21*****(3), 447–455.** doi:10.1101/gr.112623.110

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). *De novo* identification of repeat families in large genomes, ***Bioinformatics***, ***21(suppl 1)***, **i351–i358.** doi:10.1093/bioinformatics/bti1018

Priest, H. D., Filichkin, S. A., & Mockler, T. C. (2009). *cis*-regulatory elements in plant cell signaling. ***Curr Opin Plant Biol.***, ***12(5)***, **643–649.** doi:10.1016/j.pbi.2009.07.016

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. ***Nucleic Acids Res.***, ***40*****(Database issue), D130–D135.** doi:10.1093/nar/gkr1079

Qu, L.J., & Zhu, Y.X. (2006). Transcription factor families in *Arabidopsis*: major progress and outstanding issues for future research. ***Curr Opin Plant Biol.***, ***9*****(5), 544–549.** doi:10.1016/j.pbi.2006.07.005

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. ***Nucleic Acids Res.***, ***33*****(Web Server issue), W116–120.** doi:10.1093/nar/gki442

Raj, A., & McVicker, G. (2014). The genome shows its sensitive side. ***Nat Methods.***, ***11*****(1), 39–40.** doi:10.1038/nmeth.2770

Ratan, A., Zhang, Y., Hayes, V. M., Schuster, S. C., Miller, W. (2010). Calling SNPs without a reference sequence. ***BMC Bioinformatics***, ***11***, **130.** doi:10.1186/1471-2105-11-130

Rawat, V., Abdelsamad, A., Pietzenuk, B., Seymour, D. K., Koenig, D., Weigel, D., et al. (2015). Improving the annotation of *Arabidopsis lyrata* using RNA-Seq Data. ***Plos One***, ***10*****(9), e0137391.** doi:10.1371/journal.pone.0137391

Remm, M., Storm, C. E. V., & Sonnhammer, E. L. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. ***J Mol Biol.***, **314(5), 1041–1052.** doi:10.1006/jmbi.2000.5197

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., et al. (2000). Genome-wide location and function of DNA binding proteins. ***Science***, ***290*****(5500), 2306–2309.** doi: 10.1126/science.290.5500.2306

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., et al. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. ***Science* , *290*(5499), 2105–2110.** doi:10.1126/science.290.5499.2105

Robertson, G., Schein, J., Chiu, R., Corbett, R., Field, M., Jackman, S. D., et al. (2010). *De novo* assembly and analysis of RNA-seq data. ***Nat Methods.*, *7*(11), 909–912.** doi:10.1038/nmeth.1517

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P., & van de Peer, Y. (2003). Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. ***Plant Physiol.*, *132*(3), 1162–1176.** doi:10.1104/pp.102.017715

Roth, F.P., Hughes, J.D., Estep, P.W., Church, G.M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantification. ***Nat Biotechnol.,* 16(10), 939-945.** doi:10.1038/nbt1098-939

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. ***Nature*, *475*(7356), 348–352.** doi:10.1038/nature10242

Ruparel, H., Bi, L., Li, Z., Bai, X., Kim, D. H., Turro, N. J., & Ju, J. (2005). Design and synthesis of a 3'-O-allyl photocleavable fluorescent nucleotide as a reversible terminator for DNA sequencing by synthesis. ***Proc Natl Acad Sci U S A.*, *102*(17), 5932–5937.** doi:10.1073/pnas.0501962102

Rusk, N. (2014). Transcription factors without footprints. ***Nat Methods.*, *11*(10), 988–989.** doi:10.1038/nmeth.3128

Salamov, A. A., & Solovyev, V. V. (2000). *Ab initio* gene finding in *Drosophila* genomic DNA. ***Genome Res.*, 10(4), 516-522. , 516–522.** doi:10.1101/gr.10.4.516

Salmela, L., & Tarhio, J. (2007). Algorithms for weighted matching, in Proc. SPIRE 2007, LNCS 4726, ***Springer-Verlag*, 276–286.**

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., et al. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. ***Genome Res.*, *22*(3), 557–567.** doi:10.1101/gr.131383.111

Salzberg, S.L. (2007). Genome re-annotation: a wiki solution? ***Genome Biol.,* 8(1), 102.** doi: 10.1186/gb-2007-8-1-102

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. ***Proc Natl Acad Sci U S A.*, *74*(12), 5463–5467.** doi:10.1073/pnas.74.12.5463

Sarkar, C., & Maitra, A. (2008). Deciphering the *cis*-regulatory elements of co-expressed genes in PCOS by *in silico* analysis. ***Gene*, *408*(1-2), 72–84.** doi:10.1016/j.gene.2007.10.026

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012). The tomato genome sequence provides insights into fleshy fruit evolution. ***Nature***, ***485*(7400), 635–641.** doi:10.1038/nature11119

Schaefke, B., Wang, T.Y., Wang, C.Y., & Li, W.H. (2015). Gains and losses of transcription factor binding sites in *Saccharomyces cerevisiae* and *Saccharomyces paradoxus*. ***Genome Biol Evol.***, ***7*(8), 2245–2257.** doi:10.1093/gbe/evv138

Schaffer, R., Landgraf, J., Accerbi, M., Simon, V., Larson, M., & Wisman, E. (2001). Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. ***Plant Cell.***, ***13*(1), 113–123.** doi:10.1105/tpc.13.1.113

Schatz, M. C. (2009). CloudBurst: highly sensitive read mapping with MapReduce. ***Bioinformatics***, ***25*(11), 1363–1369.** doi:10.1093/bioinformatics/btp236

Schatz, M. C., Delcher, A. L., & Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. ***Genome Res.***, **20(9), 1165–1173.** doi:10.1101/gr.101360.109.20

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010). Genome sequence of the palaeopolyploid soybean. ***Nature***, ***463*(7278), 178–183.** doi:10.1038/nature08670

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009). The B73 maize genome: complexity, diversity, and dynamics. ***Science***, ***326*(5956), 1112–1115.** doi:10.1126/science.1178534

Schneeberger, K., & Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. ***Trends Plant Sci.***, ***16*(5), 282–288.** doi:10.1016/j.tplants.2011.02.006

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A. H., Nielsen, K. L., et al. (2009). SHOREmap: simultaneous mapping and mutation identification by deep sequencing. ***Nat Methods.***, ***6*(8), 550–551.** doi:10.1038/nmeth0809-550

Schneeberger, K., Ossowski, S., Ott, F., Klein, J. D., Wang, X., Lanz, C., et al. (2011). Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. ***Proc Natl Acad Sci U S A.***, ***108*(25), 10249–10254.** doi:10.1073/pnas.1107739108

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. ***Nat Genet.***, ***34*(2), 166–176.**

Seymour, D. K., Koenig, D., Hagmann, J., Becker, C., & Weigel, D. (2014). Evolution of DNA methylation patterns in the Brassicaceae is driven by differences in genome organization. ***PLoS Genetics***, ***10*(11), e1004785.** doi:10.1371/journal.pgen.1004785

Sharon, D., Tilgner, H., Grubert, F., & Snyder, M. (2013). A single-molecule long-read survey of the human transcriptome. ***Nat Biotechnol.***, ***31*(11), 1009–1014.** doi:10.1038/nbt.2705

Sharov, A. A., & Ko, M. S. H. (2009). Exhaustive search for over-represented DNA sequence motifs with CisFinder. ***DNA Res.***, ***16*(5), 261–273.** doi:10.1093/dnares/dsp014

Sheldon, C. C., Conn, A. B., Dennis, E. S., & Peacock, W. J. (2002). Different regulatory regions are required for the vernalization-induced repression of FLOWERING LOCUS C and for the epigenetic maintenance of repression. ***Plant Cell.***, ***14*(10), 2527–2537.** doi:10.1105/tpc.004564

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z.X., Zhou, Q., et al. (2012). MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. ***Nucleic Acids Res.***, ***40*(8), e61.** doi:10.1093/nar/gkr1291

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. ***Nat Biotechnol.***, ***26*(10), 1135–1145.** doi:10.1038/nbt1486

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. ***Science***, ***309*(5741), 1728–1732.** doi:10.1126/science.1117389

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (*Fragaria vesca*). ***Nat Genet.***, ***43*(2), 109–116.** doi:10.1038/ng.740

Sieburth, L. E., & Meyerowitz, E. M. (1997). Molecular dissection of the AGAMOUS control region shows that *cis* elements for spatial regulation are located intragenically. ***Plant Cell.***, ***9*(3), 355–365.** doi:10.1105/tpc.9.3.355

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. ***Genome Res.***, ***15*(8), 1034–1050.** doi:10.1101/gr.3715005

Singh, K. B. (1998). Transcriptional regulation in plants: the importance of combinatorial control. ***Plant Physiol.***, ***118*(4), 1111–1120.** doi:10.1104/pp.118.4.1111

Slotte, T., Hazzouri, K. M., Ågren, J. A., Koenig, D., Maumus, F., Guo, Y.L., et al. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. ***Nat Genet.***, ***45*(7), 831–835.** doi:10.1038/ng.2669

Staiger, D., Shin, J., Johansson, M., & Davis, S. J. (2013). The circadian clock goes genomic. ***Genome Biol.***, ***14*(6), 208.** doi:10.1186/gb-2013-14-6-208

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. ***Bioinformatics***, ***19*(Suppl 2), ii215–ii225.** doi:10.1093/bioinformatics/btg1080

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. ***BMC Bioinformatics***, ***7*****, 62.** doi:10.1186/1471-2105-7-62

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., et al. (2007). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. ***Nature***, ***450*****(7167), 219–232.** doi:10.1038/nature06340

Stormo, G. D. (2000). DNA binding sites: representation and discovery. ***Bioinformatics,*** ***16*****(1), 16–23.** doi:10.1093/bioinformatics/16.1.16

Stormo, G. D., Schneider, T.D., Gold, L., Ehrenfeucht, A. (1982). Use of the "Perceptron" algorithm to distinguish translational initiation sites in *E. coli*. ***Nucleic Acids Res.,*** ***10*****(9), 2997–3011.** doi: 10.1093/nar/10.9.2997

Straume, M. (2004). DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. ***Methods Enzymol.***, ***383*****(2001), 149–166.** doi:10.1016/S0076-6879(04)83007-6

Sun, H., De Bie, T., Storms, V., Fu, Q., Dhollander, T., Lemmens, K., et al. (2009). ModuleDigger: an itemset mining framework for the detection of *cis*-regulatory modules. ***BMC Bioinformatics***, ***10 Suppl 1*****, S30.** doi:10.1186/1471-2105-10-S1-S30

Sung, M.H., Guertin, M. J., Baek, S., & Hager, G.L. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. ***Mol Cell***, ***56*****(2), 275–285.** doi:10.1016/j.molcel.2014.08.016

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., & Jones, R. T. (1988). Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. ***J Mol Biol.***, ***203*****(2), 439–455.** doi:10.1016/0022-2836(88)90011-3

The Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the Flowering plant *Arabidopsis thaliana*. **Nature, *408, 796-815.*** doi: 10.1038/35048692

The C. elegans Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: A platform for investigating biology. ***Science, 282*****(5396), 2012–2018.** doi:10.1126/science.282.5396.2012

Thomas-Chollier, M., Sand, O., Turatsinze, J.V., Janky, R., Defrance, M., Vervisch, E., et al. (2008). RSAT: regulatory sequence analysis tools. ***Nucleic Acids Res.***, ***36*****(Web Server issue), W119–W127.** doi:10.1093/nar/gkn304

Thomas, S., Underwood, J. G., Tseng, E., & Holloway, A. K., Bench To Basinet CvDC Informatics Subcommittee. (2014). Long-read sequencing of chicken transcripts and identification of new transcript isoforms. ***PLoS ONE***, ***9*****(4), e94650.** doi:10.1371/journal.pone.0094650

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. ***Bioinformatics*, *25*(9), 1105–1111.** doi:10.1093/bioinformatics/btp120

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. ***Nat Biotechnol.*, *28*(5), 511–515.** doi:10.1038/nbt.1621

Tsompana, M., & Buck, M. J. (2014). Chromatin accessibility: a window into the genome. ***Epigenetics Chromatin*, *7*(1), 33.** doi:10.1186/1756-8935-7-33

Turner, A., Beales, J., Faure, S., Dunford, R.P., Laurie, D.A. (2005). The pseudo-response regulator Ppd-H1 provides adaptation to photoperiod in barley. ***Science*, *310*(5750), 1031–1034.** doi:10.1126/science.1117619

Van Dongen, S. (2000). Graph clustering. Graph Stimulation by Flow Clustering, ***PhD thesis*, University of Utrecht.** doi:10.1016/j.cosrev.2007.05.001

Van Helden, J., André, B., & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. ***J Mol Biol.*, *281*(5), 827–842.** doi:10.1006/jmbi.1998.1947

Varshney, R. K., Song, C., Saxena, R. K., Azam, S., Yu, S., Sharpe, A. G., et al. (2013). Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. ***Nat Biotechnol.*, *31*(3), 240–246.** doi:10.1038/nbt.2491

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (*Malus × domestica* Borkh.). ***Nat Genet.*, *42*(10), 833–839.** doi:10.1038/ng.654

Villar, D., Flicek, P., & Odom, D. T. (2014). Evolution of transcription factor binding in metazoans - mechanisms and functional implications. ***Nat Rev Genet.*, *15*(4), 221–233.** doi:10.1038/nrg3481

Wang, X., Wang, H., Wang, J., Sun, R., Wu, J., Liu, S., et al. (2011). The genome of the mesopolyploid crop species *Brassica rapa*. ***Nat Genet.*, *43*(10), 1035–1039.** doi:10.1038/ng.919

Weigel, D., & Mott, R. (2009). The 1001 genomes project for *Arabidopsis thaliana*. ***Genome Biol.*, *10*(5), 107.** doi:10.1186/gb-2009-10-5-107

Willing, E.M., Rawat, V., Mandáková, T., Maumus, F., James, G. V., Nordström, K. J. V., et al. (2015). Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. ***Nature Plants*, 1(2), 14023.** doi:10.1038/nplants.2014.23

Wray, G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V. (2003). The evolution of transcriptional regulation in eukaryotes. ***Mol Biol Evol.*, *20*(9), 1377–1419.** doi:10.1093/molbev/msg140

Wrzodek, C., Schröder, A., Dräger, A., Wanke, D., Berendzen, K. W., Kronfeld, M., et al. (2010). ModuleMaster: A new tool to decipher transcriptional regulatory networks. ***Biosystems.*, *99*(1), 79–81.** doi:10.1016/j.biosystems.2009.09.005

Wu, C. H., Apweiler, R., Bairoch, A., Natale, D. A., Barker, W. C., Boeckmann, B., et al. (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. ***Nucleic Acids Res.*, *34(Database issue)*, D187–D191.** doi:10.1093/nar/gkj161

Wu, H.J., Zhang, Z., Wang, J.Y., Oh, D.H., Dassanayake, M., Liu, B., et al. (2012). Insights into salt tolerance from the genome of *Thellungiella salsuginea*. ***Proc Natl Acad Sci U S A.*, *109*(30), 12219–12224.** doi:10.1073/pnas.1209954109

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. ***Nucleic Acids Res.*, *35(Web Server issue)*, W265–W268.** doi:10.1093/nar/gkm286

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. ***Nat Rev Genet.*, *13*(5), 329–342.** doi:10.1038/nrg3174

Yardımcı, G. G., Frank, C. L., Crawford, G. E., & Ohler, U. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. ***Nucleic Acids Res.*, *42*(19), 11865–11878.** doi:10.1093/nar/gku810

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., et al. (2002). A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). ***Science*, 296(5565), 79-92.**

Zakhrabekova, S., Gough, S. P., Braumann, I., Müller, A.H., Lundqvist, J., Ahmann, K., et al. (2012). Induced mutations in circadian clock regulator Mat-a facilitated short-season adaptation and range extension in cultivated barley. ***Proc Natl Acad Sci U S A.*, *109*(11), 4326–4331.** doi:10.1073/pnas.1113009109

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. ***Genome Res.*, *18*(5), 821–829.** doi:10.1101/gr.074492.107

Zhang, W., Zhang, T., Wu, Y., & Jiang, J. (2012). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in *Arabidopsis*. ***Plant Cell.*, *24*(7), 2719–2731.** doi:10.1105/tpc.112.098061

Zhou, Q., & Wong, W. H. (2004). CisModule: *De no*vo discovery of cis-regulatory modules by hierarchical mixture modeling. ***Proc Natl Acad Sci U S A.*, *101*(33), 12114–12119.** doi:10.1073/pnas.0402858101

Acknowledgement

First and foremost, I want to say a special thanks to my supervisor Dr. Korbinian Schneeberger. You mentored me in a true sense and your passion toward Science was always a source of motivation for me. Thanks for not only making this Ph.D. possible but also for introducing me to a "better" way of doing science. Every scientific discussion was a new lesson learnt and a step towards being a better scientist.

I would also like to thank Prof. George Coupland for his valuable comments throughout my Ph.D.

I am also thankful to Prof. Thomas Wiehe and Prof. Martin Hülskamp being a part of my Ph.D. thesis committee.

During my Ph.D., many collaborators (difficult to name all of them) also helped me with their expert knowledge. Theo, Yue, Ahmed, Björn, Niels, Julieta, Maida, Markus, Vid, Geo ….., I am thankful to all of you.

All this would not have been possible without my wonderful colleague. A special thanks to Eva, Jonas, Ben, Vipul, Mathieu, Geo, Hequan, Karl, Wen-Biao and Maria for your contribution in creating a friendly and scientifically stimulating environment in Group_Schneeberger. "Can I ask a question if you are not busy?" is the sentence that helped me learning many new concepts and thanks guys for answering that question always with a "yes".

I also want to thank my Indian friends, who provided a friendly and homely environment away from home.

My thanks are too small to acknowledge how much support I got from my lovely wife, Ranu. Thanks for everything you did to make this Ph.D. possible for me. I am also thankful for my family for their support and trust in me.

Erklärung

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen , die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie abgesehen von unten angegbenen Teilpublikationen nocgh nicht veröffentlicht worden ist sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen dieser Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. George Coupland und Dr. Korbinian Schneeberger betreut worden.

Date:

Koln: Vimal Rawat