Journal of Experimental Botany Advance Access published March 22, 2016

Journal of Experimental Botany
doi:10.1093/jxb/erw119

FLOWERING NEWSLETTER REVIEW

The analysis of Gene Regulatory Networks in plant evo-devo

Aurélie C. M. Vialette-Guiraud[1], Amélie Andres-Robin[1], Pierre Chambrier[1], Raquel Tavares[2] and Charles P. Scutt[1,*]

[1] Laboratoire de Reproduction et Développement des Plantes (UMR 5667 – CNRS/INRA/ENS-Lyon/université Lyon 1/université de Lyon), Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France
[2] Laboratoire de Biométrie et Biologie Évolutive (UMR 5558 – CNRS/université Lyon 1/université de Lyon), Bâtiment Gregor Mendel, 43 bd du 11 novembre 1918, 69622 Villeurbanne Cedex, France

* Correspondence: charlie.scutt@ens-lyon.fr

Received 28 January 2016; Accepted 29 February 2016

Editor: Lars Hennig, Swedish University of Agricultural Science

Abstract

We provide an overview of methods and workflows that can be used to investigate the topologies of Gene Regulatory Networks (GRNs) in the context of plant evolutionary-developmental (evo-devo) biology. Many of the species that occupy key positions in plant phylogeny are poorly adapted as laboratory models and so we focus here on techniques that can be efficiently applied to both model and non-model species of interest to plant evo-devo. We outline methods that can be used to describe gene expression patterns and also to elucidate the transcriptional, post-transcriptional, and epigenetic regulatory mechanisms underlying these patterns, in any plant species with a sequenced genome. We furthermore describe how the technique of Protein Resurrection can be used to confirm inferences on ancestral GRNs and also to provide otherwise-inaccessible points of reference in evolutionary histories by exploiting paralogues generated in gene and whole genome duplication events. Finally, we argue for the better integration of molecular data with information from paleobotanical, paleoecological, and paleogeographical studies to provide the fullest possible picture of the processes that have shaped the evolution of plant development.

Key words: Ancestral Sequence Reconstruction, evolutionary-developmental biology, evo-devo, epigenetic regulation, Gene Regulatory Network, plant, post-transcriptional regulation, Protein Resurrection, transcriptional regulation.

What's the problem?

The ambition of evolutionary-developmental biology (evo-devo) is to explain how evolutionary processes have shaped organismal development and the emergence of new biological forms. Molecular-genetic studies of model organisms indicate that developmental processes are typically controlled by Gene Regulatory Networks (GRNs) whose components are linked together by several different types of regulatory interaction (Fig. 1A). However, evo-devo studies must be performed on species chosen for their phylogenetic positions relative to evolutionary transitions of interest and, thus, in many cases, will include non-model species that are not ideally suited to molecular-genetic analyses. The objective of this article is to review methods and workflows that can be used to compare GRNs between model and non-model plants and thus identify cases of conservation or non-conservation in regulatory components and interactions since the most recent common ancestor of the species under consideration (Fig. 1B). By comparing network topologies in appropriate taxa, it should be possible to infer the ancestral states of GRNs at various different levels in an organismal phylogeny and thereby correlate changes in network topology with evolutionary transitions at the morphological level. Furthermore, the technique of Protein Resurrection may be used to reconstruct ancestral regulatory molecules which can then be subjected to a range

© The Author 2016. Published by Oxford University Press on behalf of the Society for Experimental Biology. All rights reserved.
For permissions, please email: journals.permissions@oup.com

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

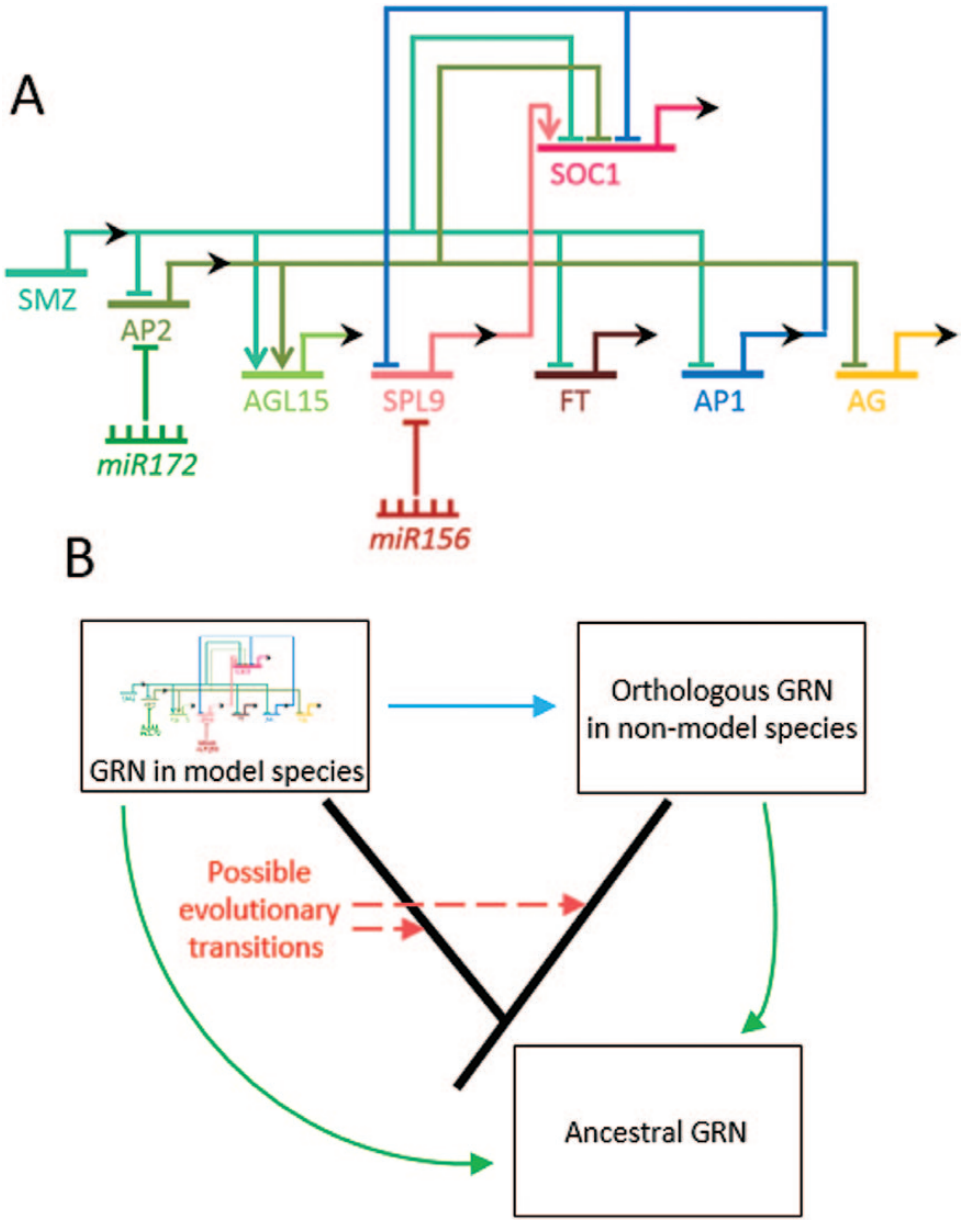

Fig. 1. Gene Regulatory Networks. (A) Part of a Gene Regulatory Network (GRN) controlling the induction of flowering in the model plant *Arabidopsis thaliana*, based on Kaufmann *et al.* (2010). *Cis*-regulatory regions of genes encoding network components are shown by thick horizontal lines. Transcription is shown by black arrowheads. Positive and negative transcriptional regulation is shown by coloured arrows and bars (respectively) above the *cis*-regulatory regions. Two components of the GRN are regulated post-transcriptionally by *miR156* and *miR172*. (B) An evo-devo analysis of the molecular causes of an evolutionary transition (red arrows) in one of two lineages leading to a model and a non-model species, respectively. The GRN controlling the developmental character or process of interest (e.g. flowering) is first analysed in the model species. The methods of analysis described in this article can then be used to assemble the orthologous GRN in the non-model species (blue arrow). The network topologies in the model and non-model species can be compared (green arrows) to infer the ancestral state of the GRN in their most recent common ancestor.

of studies in the laboratory, thus providing direct insights into the ancestral states of GRNs.

Mapping developmental character states onto organismal phylogenies

Evo-devo analyses invariably start from an established organismal phylogeny or, if necessary, from two or more possible alternative phylogenies. Establishing and refining such phylogenies in all branches of the tree of life has become a major goal of the biological sciences. Current consensus phylogenies within the angiosperms are summarized on the Angiosperm Phylogeny Group website (http://www.mobot.org/MOBOT/research/APweb/; Bremer *et al.*, 2009), while more detailed phylogenies within individual angiosperm groups are covered in a large number of separate publications. Recent molecular phylogenies that summarize the evolutionary relationships between the major groups of embryophytes (land plants) include Wickett *et al.* (2014), while the relationship between the land plants and the aquatic groups from which they emerge have been investigated by Finet *et al.* (2010*b*).

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

A first step in evo-devo studies is to combine morphological and developmental data with organismal phylogenies to infer the positions and types of morphological/developmental transitions that have taken place. Such analyses can conveniently be performed using maximum parsimony or maximum-likelihood procedures in programs such as MacClade (Vazquez, 2004) and Mesquite (Maddison and Maddison, 2015). The accurate coding of character states is of key importance in character state reconstructions and care should be taken to discriminate between superficially similar structures or developmental processes of likely independent origin.

The reconstruction of ancestral character states allows the choice of species for detailed investigations of the molecular basis underlying developmental transitions. Such studies may focus on the acquisition or loss of a simple character, the progressive acquisition or loss of a complex character, or on a series of acquisitions and losses (transitions and reversions). Species should be chosen from evolutionary branches or clades representing each character state to be compared. These species should include at least one model in which the involvement of specific regulatory molecules in the developmental process of interest has been demonstrated and, if possible, model species representing both character states in a given evolutionary transition should be investigated. Examples of such pairs of model species that have proven useful for evo-devo analyses include the relatively closely related Brassicaceae species *Arabidopsis thaliana* and *Cardamine hirsuta*, which differ in leaf morphology and fruit structure, among other traits (Hay *et al.*, 2014). Similarly, an in-depth comparison of direct transcriptional regulation involving the transcription factor SEPALLATA3 has recently been made between the closely related models *A. thaliana* and *A. lyrata* (Muino *et al.*, 2016). However, in many cases, non-model species, for which direct functional genetic studies cannot readily be performed, must be incorporated into evo-devo analyses. If possible, such analyses should include two or more taxa situated on either side of a developmental transition of interest, as this experimental design is likely to provide better support for correlations between the developmental transition and underlying molecular changes.

Requirements in non-models for the efficient investigation of Gene Regulatory Networks

A first requirement for evo-devo studies, and one which is sometimes not so easily fulfilled, is the availability of appropriate biological material. In some cases, plant species of interest may be endemic to areas that are difficult to access, endangered and thus subject to specific regulations for collection and export, or difficult to grow in cultivation. In addition, reproductive stages (e.g. flowers, cones, etc) may be physically difficult to access or produced sporadically or over a long time-period making it difficult to sample all relevant developmental stages during fieldwork. A second practical problem for evo-devo studies concerns the application to non-model species of basic techniques of molecular biology such as the extraction of nucleic acids or of more sophisticated methods such as RNA *in situ* hybridization.

Once appropriate biological material has been obtained and methods for nucleic acid extraction successfully applied, the presence or absence of individual regulatory genes, by comparison to well-studied model organisms, can be investigated in practically any species using standard procedures for molecular cloning and/or PCR amplification. However, the study of complex GRNs in non-model organisms is greatly facilitated by the availability of complete genome sequences. A draft genome assembly, in which most coding sequences are associated on scaffolds with their potential *cis*-acting regulatory sequences (e.g. promoters and intronic and 3′-regulatory sequences), is a tremendous advantage for the investigation of gene network relationships. A fully sequenced genome also confers the possibility, via phylogenetic analyses, of reaching firm conclusions on gene orthology between species and on the positions within the organismal phylogeny of events such as gene duplications and losses. In addition to complete genome sequences, transcriptomic data is of tremendous use to the evo-devo analysis of GRNs, as described in the section on Transcriptomic Data.

An exhaustive list of genomic and transcriptomic data-sets in plants would be very long and, at the current rate of progress, out of date by the time this article appeared in press. We therefore give, in Table 1, a list of web resources for a sample of species across the phylogenetic tree of the land plants and their aquatic relatives, concentrating on species whose genome sequences are available. Many more genomes are currently available within the eudicots and Poaceae (grasses) than in other groups of angiosperms, while all other major groups of land plants are, for the moment, represented by a small number of genome sequences or, in some cases (e.g. ferns, liverworts, and hornworts), contain no species with sequenced genomes. However, the rate of progress in sequencing technology is such that, within a few years, the limited availability of whole genome sequences may cease to be a significant impediment to plant evo-devo research.

What aspects of Gene Regulatory Networks are accessible to study in non-model plants?

Model plants, on which most molecular-developmental analyses are based, typically show a range of characteristics that lend themselves to laboratory studies including reasonably short generation times, manageable physical sizes, ease of cultivation, and favourable characteristics for the storage and manipulation of germplasm. Critically, model plants are highly amenable to forward and/or reverse functional genetic analyses and, in most cases, are readily transformable, facilitating mis-expression and over-expression studies, the use of RNA interference (RNAi) technology, and the latest developments in genome editing (Doudna and Charpentier, 2014) which allow precise genomic changes to be introduced at any existing locus. By contrast, the non-model plants that

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

Table 1. *A sample across the land plants and their aquatic relatives of key taxa for evo-devo studies that possess sequenced genomes or extensive transcriptomic resources*

Clade	Species	Genome	Transcriptomic data
Green algae	*Chlamydomonas reinhardtii*	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Creinhardtii	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Creinhardtii
	Volvox carteri	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vcarteri	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Vcarteri
	Ostreococcus lucimarinus	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Olucimarinus	
Bryophyta	*Physcomitrella patens*	http://www.cosmoss.org/	http://www.cosmoss.org/
	Sphagnum fallax	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Sfallax_er	
Lycopodiophyta	*Selaginella moellendorffii*	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Smoellendorffii	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Smoellendorffii
Monilophyta			Onekp.com (74 species)
Ginkgoales	*Ginkgo biloba*		Onekp.com (1species)
Cycadales			Onekp.com (4 species)
Gnetales			Onekp.com (3 species)
Conifers	*Picea abies*	http://congenie.org/	http://congenie.org/
	Pinus taeda	http://congenie.org/	http://congenie.org/
Amborellaceae	*Amborella trichopoda*	http://www.amborella.org/	http://www.amborella.org/
Monocots	*Phalaenopsis equestris*	http://orchidbase.itps.ncku.edu.tw/est/home2012.aspx	http://orchidbase.itps.ncku.edu.tw/est/home2012.aspx
	Phoenix dactylifera	http://qatar-weill.cornell.edu/research/datepalmGenome/	
	Musa acuminata	http://www.nature.com/nature/journal/v488/n7410/full/nature11241.html	http://www.nature.com/nature/journal/v488/n7410/full/nature11241.html
	Oryza sativa	http://rice.plantbiology.msu.edu/	http://rice.plantbiology.msu.edu/
Basal eudicots	*Aquilegia caerulea*	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Acoerulea	http://www.ncbi.nlm.nih.gov/bioproject/270946
	Nelumbo nucifera	http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AQOG01	http://www.ncbi.nlm.nih.gov/bioproject/196884
Rosids	*Vitis vinifera*	http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/	http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-691
	Arabidopsis thaliana	http://www.arabidopsis.org/	http://signal.salk.edu/cgi-bin/atta
	Populus trichocarpa	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Ptrichocarpa	https://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Ptrichocarpa
	Fragaria vesca	https://www.rosaceae.org	https://www.rosaceae.org
	Medicago truncatula	http://www.nature.com/nature/journal/v480/n7378/full/nature10625.html	http://www.nature.com/nature/journal/v480/n7378/full/nature10625.html
Asterids	*Beta vulgaris*	http://bvseq.molgen.mpg.de/index.shtml	http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-13-99
	Actinidia chinensis	http://bioinfo.bti.cornell.edu/cgi-bin/kiwi/home.cgi	http://www.ncbi.nlm.nih.gov/sra?term=SRA065642
	Coffea canephora	http://coffee-genome.org/	http://coffee-genome.org/
	Solanum tuberosum	http://solanaceae.plantbiology.msu.edu/index.shtml	http://solanaceae.plantbiology.msu.edu/index.shtml

frequently occupy key positions in plant phylogeny, and whose study is therefore inescapable in many evo-devo analyses, may have a very large physical size, a long life cycle, or require growth conditions that are difficult to provide. These species may also be refractory to genetic transformation and present practical problems for the use and storage of germplasm.

Despite such experimental limitations, numerous techniques can be used to study GRNs in non-model plants, particularly compared with models in which functional experiments have been performed on orthologous network components. These methods include DNA sequence and gene expression analyses, *in vitro* analyses to investigate protein–DNA and protein–protein interactions, and the *in silico* scanning of genomic and transcriptomic datasets to uncover transcriptional and post-transcriptional regulatory sites. A major source of genome regulation derives from epigenetic modifications to DNA and chromatin (Engelhorn *et al.*, 2014), and these mechanisms too are open to experimental investigation in non-model species with sequenced genomes. Methods for the analysis of all of the above aspects of GRNs in non-model species are outlined in the following sections, while a further section describes methods for the reconstruction and study of ancestral components of such networks. A final section argues for the integration of

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

data from molecular and non-molecular studies that will be required to understand more fully the many factors that have shaped the evolution of plant morphology.

The structure of gene families and coding sequences

An investigation of the structure of the gene families potentially involved in evolutionary transitions of interest forms an early step in most molecular-based evo-devo analyses. The gene families under investigation, typically identified in model systems, may include transcriptional regulators and their direct target genes, as well as small RNAs and their transcribed targets. The presence of fully sequenced genomes in both the model and non-model species included in evo-devo analyses is an advantage for phylogenetic reconstructions. Care should be taken with taxonomic sampling and the choice of evolutionary models to minimize the phenomenon of long-branch attraction which tends to distort gene phylogenies (Boussau *et al.*, 2014). Methods used for phylogenetic reconstruction are described in numerous websites, textbooks, and technical articles such as Lemay *et al.* (2009) and can be performed locally or on web-based platforms, typically using Maximum Likelihood or Bayesian methods.

The reconstruction of GRNs, sometimes including many hundreds of co-regulated genes, may require methods for the identification of gene orthologues on a large scale. In these cases, the manual construction of phylogenies is very time-consuming and it is simpler and more efficient to use an automated workflow for the detection of orthologues. Early work for the large-scale identification of orthologues exploited reciprocal best BLAST-hits (Altschul *et al.*, 1990) between sequenced genomes or transcriptomes as a guide to probable orthology. However, these simple methods have been superseded by the use of data pipelines that can be used to identify clusters of related sequences (Miele *et al.*, 2011) from the results of automated BLAST searches, perform multiple alignments (Edgar, 2004), and phylogenies on these data (Guindon and Gascuel, 2003; Price *et al.*, 2009) and then comprehensively identify gene orthology relationships (Dufayard *et al.*, 2005; Bigot *et al.*, 2013).

The identification of gene orthology relationships also indicates the positions within the organismal phylogeny of the gene duplication and loss events that frequently underlie such evolutionary processes as neo-, sub-, and non-functionalization (Taylor and Raes, 2004). Within this phylogenetic context, the coding sequences of individual genes can then be analysed for any modular evolution events that might have taken place such as domain losses or rearrangements (Kersting *et al.*, 2012). Such changes in the domain structure of regulatory proteins represent potentially important evolutionary mechanisms and have been found in numerous families of plant transcriptional regulators (Finet *et al.*, 2010*a*; Reymond *et al.*, 2012).

Transcriptomic data

Transcriptomic data are now available for over 1 000 plant species, covering all major clades of embryophytes and their algal relatives (Matasci *et al.*, 2014) and representing a major resource for plant-evo-devo. Basic transcriptomic data, providing an extensive list of transcribed genes, can be used to detect possible occurrences of gene loss or duplication in species for which complete genome sequences are not yet available. More detailed transcriptomic data, covering different tissues and/or stages, can also be used to deduce changes in gene expression patterns that correlate with evolutionary transitions at the morphological level.

Transcriptomic data can now be obtained from picogram quantities of RNA, such as can be prepared from tissues isolated by Laser Capture Microdissection (LCM) and other methods (Crosetto *et al.*, 2015). The combination of LCM and RNA-sequencing (RNA-seq) enables very precise global correlations to be made between gene expression patterns in specific tissues or even within individual cell types. Positive and negative correlations in expression between groups of genes can then be used to generate hypotheses to explain their interactions and relationships. In this way, groups of co-regulated genes can be defined, as can possible positive or negative regulatory interactions (e.g. transcription factors and their positively or negatively regulated targets). Such data can be used directly to form hypotheses for the evolution of regulatory interactions but can also be used to correlate with data concerning transcriptional, post-transcriptional or epigenetic control mechanisms, thus contributing to more robust hypotheses for the topology and evolution of GRNs. A detailed discussion of the uses and interpretation of transcriptomic data in evo-devo research is provided by Pantalacci and Semon (2015).

A further technique of interest to plant evo-devo is small RNA-seq, which can be used to identify complete sets of small RNA regulators in species occupying key phylogenetic positions. This method is also compatible with LCM techniques, as discussed above, for the precise localization of small RNA regulators within organs and tissues. Taylor *et al.* (2014) provide a protocol for the designation of novel small RNAs, starting from small RNA-seq data, as well as a useful survey of the origins and occurrences of small RNA regulators across the major groups of land plants.

Transcriptional regulation

Current thinking suggests that many of the developmental novelties that characterize morphological evolution arose not from changes to the coding sequences of developmental regulators but from changes in the interactions that link these components together in GRNs (Carroll, 2008). One of the most important of these regulatory mechanisms occurs through the control of transcription. Protein-coding nuclear genes in plants, as well as most small RNA regulators, are transcribed by RNA polymerase II (Pol II), which initiates transcription by binding as part of a Pre-Initiation Complex (PIC) to the proximal promoter element of its target genes, as reviewed by Engelhorn *et al.* (2014). In addition to Pol II, the PIC contains six general transcription factors that play universal roles in the initiation of transcription. Transcription is regulated by gene-specific enhancer or repressor elements that are frequently situated in proximal or distal promoter regions

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

(upstream of the transcribed region) although these may also occur in introns and 3′-flanking sequences. One important mechanism for transcriptional regulation involves the physical interaction of gene-specific transcription factors with the Mediator Complex, which also interacts physically with the PIC and may play a general role in transcriptional regulation, as reviewed by Samanta and Thakur (2015). Other mechanisms through which gene-specific transcription factors act on transcription include interactions with enzymes that act on various aspects of chromatin structure and biochemistry (as detailed later in the section on Epigenetic Regulation), thus regulating access of the transcriptional machinery to DNA.

Transcription factors may bind to specific target sites (Transcription Factor Binding Sites, TFBSs) in enhancer or repressor elements situated in *cis*-regulatory regions, and may do so individually, as dimers, or as part of a higher order complex of factors. Such higher order complexes may bind simultaneously to two or more DNA sites which may be brought into close juxtaposition by looping of the DNA. Transcription factor dimers typically interact with TFBSs of around 5–20 bp. In many cases, however, the core binding motif (containing the most precisely defined nucleotide positions) falls within the shorter end of this size range (5–10 bp). The presence of TFBSs close to a transcribed region of interest may be taken to indicate the possibility of a direct transcriptional regulatory interaction.

In model plants, direct transcriptional regulation can be investigated using a range of *in vivo* methods. For example, two variants of chromatin immunoprecipitation (ChIP) termed ChIP-seq and ChIP-exo (Rhee and Pugh, 2011) can be used to provide a map of *in vivo* transcription-factor binding to an entire genome. Interestingly, ChIP-seq data may indicate physical interactions between a given transcription factor and a very large number of downstream target genes, such as the 3 475 predicted targets for the *A. thaliana* transcription factor SEPALLATA3 (Kaufmann *et al.*, 2009) and it is possible that not all the targets identified in such cases are of biological significance. To identify targets of biological significance specifically, ChIP-seq or ChIP-exo data can be cross-checked with transcriptomic data from different tissues or stages or from mutant versus wild-type plants etc. Furthermore, induction experiments may be performed in which the transcription factor of interest, often in a transgenic context, is up-regulated such that its immediate effects on transcription can be monitored, typically using RNA-seq or whole transcriptome microarrays (for example, Wagner *et al.*, 2004). The most commonly used induction system in such experiments is based on the addition, to a transcription factor of interest, of the hormone-binding domain of the rat glucocorticoid receptor protein. This GR domain is believed to interact with heat-shock proteins in the cytoplasm, which prevents the nuclear translocation of the recombinant protein. However, nuclear translocation of the recombinant transcription factor can be rapidly induced by treatment of plants with the hormone analogue dexamethasone (DEX) which binds strongly to the GR domain causing the displacement of the associated heat-shock proteins. Interestingly, a variant of the DEX-induction system has been developed recently to identify transient transcriptional interactions that escape detection by ChIP-based methods (Para *et al.*, 2014).

The DNA-binding preferences of a transcription factor of interest may be deduced in model species from ChIP-seq data-sets using computational methods (Thomas-Chollier *et al.*, 2011; Medina-Rivera *et al.*, 2015). However, for comparisons between model and non-model species, it may be necessary to use *in vitro* methods to deduce the DNA-binding preferences of transcription factors. The two principal methods available for this are Systematic Evolution of Ligands by Exponential Enrichment (SELEX; e.g. Jolma *et al.*, 2013) and Protein Binding Microarrays (PBMs;e.g. Godoy *et al.*, 2011). SELEX is based on *in vitro* interactions between transcription factors and a large pool of free, double-stranded oligonucleotides in solution. Sequences that bind to the factor of interest are sequentially enriched, eluted, and PCR-amplified before being identified by next generation sequencing (NGS) and subjected to statistical analysis to provide a mathematical description of transcription factor binding preferences (see below). By contrast, PBM analyses are based on a large set of double-stranded oligonucleotides attached to a microarray. For example, Godoy *et al.* (2011) use a set of approximately 240 000 35-mer oligonucleotides in which all possible 11-mer sequences are present, once each. The affinity of a tagged recombinant transcription factor for each sequence of a defined length can then be determined by imaging the array using a fluorescent antibody directed against the protein tag and performing statistical analyses on the data obtained (Berger and Bulyk, 2009). Both SELEX and PBM analyses can be performed using a naïve pool of oligonucleotides or on a pool in which certain positions have been fixed in relation to a previously determined core binding sequence. Both methods can be used to measure the binding of individual transcription factors or of transcription factor dimers or complexes to a motif representing a single TFBS.

Transcription factor binding preferences derived from ChIP-seq data-sets or by using SELEX or PBM procedures can be expressed as a Position-Specific Scoring Matrix (PSSM), also known as a Position Weight Matrix (PWM), as illustrated in Fig. 2. Such matrices for 63 *A. thaliana* transcription factors, many of which are involved in developmental processes, were recently published by Franco-Zorrilla *et al.* (2014), further adding to the extensive databases of PSSMs such as JASPAR (Mathelier *et al.*, 2016). PSSMs can be used to scan whole or partial genome sequences to indicate potential instances of regulation by transcription factors of interest using a number of bioinformatics methods (Turatsinze *et al.*, 2008; Korhonen et al; 2009; Grant *et al.*, 2011**).** These programmes typically retain sites that produce a score higher than a predefined cut-off value or eliminate sites with a *P*-value higher than a given limit, the *P*-value being a statistic that indicates the probability of obtaining a given score by chance. If a list of direct targets for a given transcription factor is available for a model species of interest, the PSSMs of orthologous transcription factors from non-model species can then be used to scan the *cis*-regulatory regions of genes from those genomes that are orthologous to the (known) targets in the model species. In many cases, several high-scoring

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

A	0.1834	0.0044	0.0090	0.9866	0.0010	0.0284	0.0024	0.0023	0.0059	0.0912
C	0.3664	0.9566	0.9853	0.0060	0.9681	0.0025	0.0050	0.0034	0.0331	0.4031
G	0.3657	0.0331	0.0034	0.0050	0.0025	0.9681	0.0060	0.9853	0.9566	0.3482
T	0.0844	0.0059	0.0023	0.0024	0.0284	0.0010	0.9866	0.0090	0.0044	0.1576

Fig. 2. Position-Specific Scoring Matrices. (A) Position-Specific Scoring Matrix (PSSM) in JASPAR format for the *Arabidopsis thaliana* transcription factor PHYTOCHROME INTERACTING FACTOR3 (PIF3), taken from Franco-Zorilla et al. (2014), and the corresponding graphical representation (logo), generated using enoLOGOS (Workman *et al.*, 2005). Probabilities of finding each possible nucleotide are shown over ten contiguous bases. Information-rich sites, showing clear nucleotide preferences, generate tall letters or letter-combinations in the logo.

sites for a given transcription factor may be found within the putative regulatory regions of its target genes. In such cases, it is possible to integrate the number of sites present and their individual scores or *P*-values to provide an overall likelihood of occupancy of each potential target promoter (Turatsinze *et al.*, 2008). A generalized workflow for the comparison of direct transcriptional interactions involving orthologous transcription factors from a model and a non-model species is indicated in Fig. 3.

PSSMs are very simple and practical to use but do not take into account interdependency between nucleotide positions within binding sites. Hence, for some transcription factors, two or more PSSMs may be required to represent the full spectrum of possible interactions accurately (and many examples of this are provided by Franco-Zorilla *et al.*, 2014). An alternative approach is to calculate a scoring matrix that takes into account the interdependency of sites (e.g. Mathelier and Wasserman, 2013; Minguet *et al.*, 2015), although these methods are typically limited to short-range dependencies between adjacent sites or within contiguous trinucleotide sequences. A further alternative, which should systematically account for the interdependence of nucleotide positions, is to search for candidate binding sites in target DNA using K-mer tables derived from SELEX or PBM data (Berger and Bulyk, 2009), rather than using PSSMs.

The use of PSSMs to scan the *cis*-regulatory regions of potential target genes in non-model species (by comparison with known targets in model species) should indicate instances of possible conservation or non-conservation in regulatory interactions between the model and non-model species being compared. However, this approach cannot be used to identify novel regulatory interactions in non-model species (which are not present in the model species being used for comparison). Such novel interactions in non-models might be uncovered by scanning whole genome sequences with PSSMs, although these motifs are typically quite short and may show low variability between different members of a transcription factor family. Therefore, the conclusive identification of novel transcriptional interactions in a non-model species may require corroborative evidence from further sources such as transcriptomics and other data-sets (see the section on Data Integration).

In certain cases, the order, orientation, and spacing of two or more binding sites within a *cis*-regulatory region may be critical to its physical interactions with transcription factor complexes and the accurate measurement of transcriptional interactions in these cases will almost certainly require the use of longer DNA molecules than the oligonucleotides typically used in SELEX or PBM analyses. For example, Electrophoretic Mobility Shift Assays (EMSAs) and DNase I protection assays have been used to study the binding of tetramers of MADS-box transcription factors to pairs of CArG-box motifs with an approximate spacing of 65 bp (Melzer *et al.*, 2009). However, for the study of much longer cis-regulatory stretches of DNA, Surface Plasmon Resonance (SPR) analysis, which has been used to quantify the *in vitro* binding of transcription factors containing multiple binding motifs to promoter fragments of up to 3 kb in length, may prove useful (Moyroud *et al.*, 2009). SPR analysis permits the quantitation of positively or negatively cooperative binding between transcription factor complexes and multiple DNA binding sites (Majka and Speck, 2007). Accordingly, this method has recently been used to investigate the binding of dimers of AUXIN RESPONSE FACTORs (ARFs) to pairs of Auxin Response Elements (AuxREs) in target DNA whose exact spacing may be responsible for defining the specificity of interactions between different ARF family members and their distinct sets of downstream targets (Boer *et al.*, 2014).

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

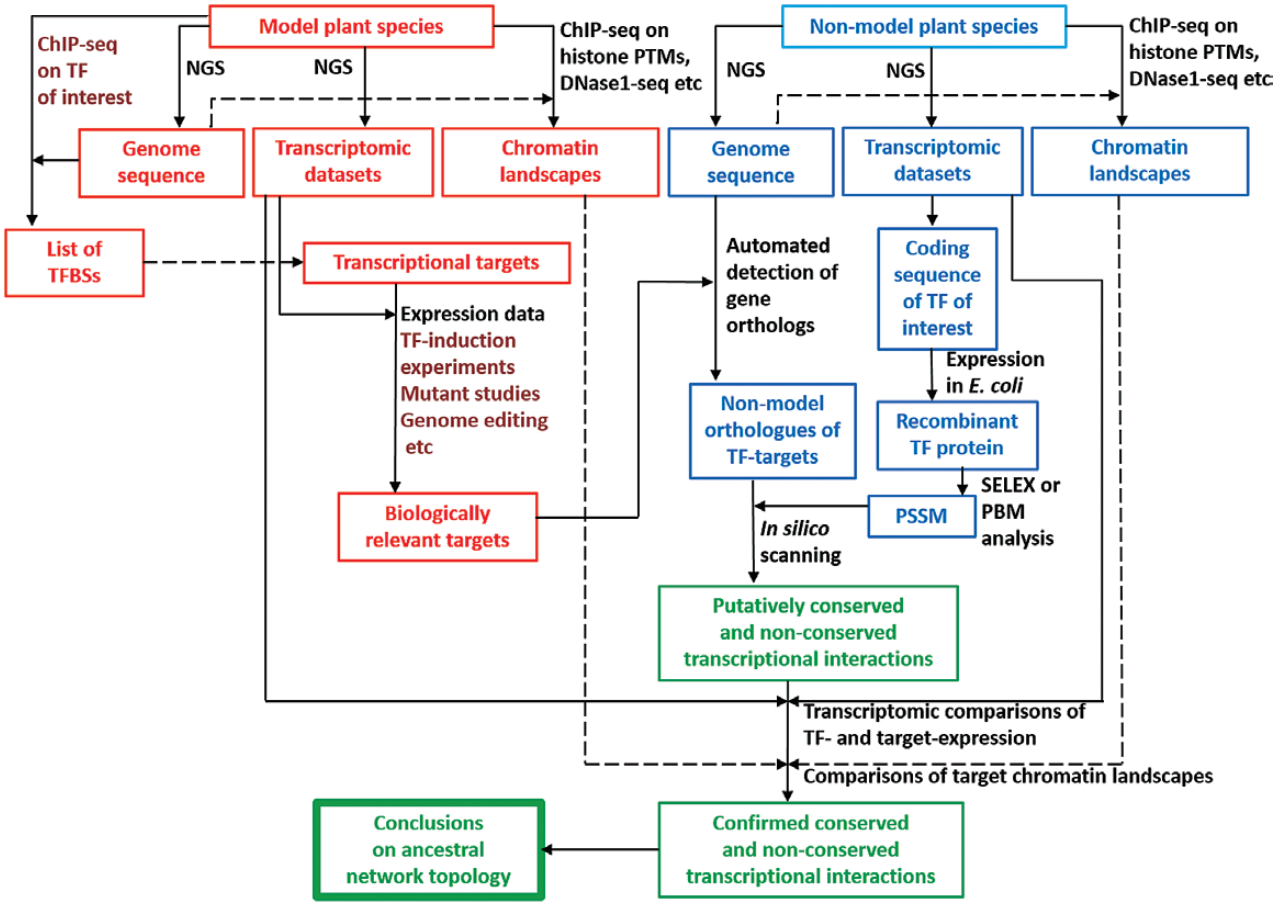

Fig. 3. A workflow for the evo-devo comparison of transcriptional interactions involving orthologous transcription factors from a model and a non-model species. Objects and data-sets derived exclusively from the model and non-model species are shown in red and blue, respectively. Data-sets derived from a comparison of the model and non-model species are shown in green. Procedures that can be applied to either species are shown in black, while those applicable only to the model are shown in brown. For clarity, a few paths in the workflow are indicated by broken lines. Abbreviations: NGS, Next generation sequencing; PBM, Protein Binding Microarray; PSSM, Position-Specific Scoring Matrix;, PTM, post-translational modification; SELEX, Systematic Evolution of Ligands by Exponential Enrichment; TF, transcription factor; TFBS, Transcription Factor Binding Site.

Further methods for the analysis of interactions involving complex *cis*-regulatory regions are based on the transient expression of transcription factors in a heterologous plant system, such as that devised by Hellens *et al.* (2005).

Post-transcriptional regulation

Small RNAs form an important class of regulatory component in many GRNs. The biogenesis and mode of action of small RNAs in plants is highly complex and has been studied principally in *A. thaliana*. However, as regulatory interactions involving small RNAs occur through base pairing with target sequences, the conservation or non-conservation of these interactions is entirely accessible to studies in non-model plants in an evo-devo context.

Small RNAs include microRNAs (miRNAs) and several classes of short interfering RNAs (siRNAs). As reviewed by Borges and Martienssen (2015), the primary transcripts of most miRNAs in plants form a hairpin structure critically involving duplex formation between a mature miRNA sequence of 20–22 nt and an imperfectly matched reverse complementary miR* sequence. These transcripts are processed by the dicer-like nuclease DCL1 to the mature miRNA/miR* duplex which is then loaded onto an ARGONAUTE (AGO) protein to become incorporated in an RNA-Induced Silencing Complex (RISC). AGO proteins, which are encoded in *A. thaliana* by a multigene family, show different specificities towards miRNAs, depending on the nucleotide at the 5′-extremity of the mature miRNA and on the positions of mis-matches within its RNA duplex (Poulsen *et al.*, 2013). Following removal of the miR* strand, the miRNA-containing RISC is able to target, by base-pairing, specific mRNAs in the cytoplasm, either causing these to be degraded or, in a minority of cases, by preventing translation. In plants, mature miRNA sequences typically show high similarity to both their complementary miR* sites and to their target sequences in mRNAs. This feature facilitates the accurate identification, using bioinformatics procedures, of both MIRNAs (i.e. the genes encoding miRNAs) in whole genome sequences and target mRNAs from genomic or transcriptomic databases. Indeed, the conserved features of MIRNAs can even be used to discover novel orthologues of these genes using PCR-based procedures (Jasinski *et al.*, 2010).

A useful workflow for the bioinformatics detection of MIRNA genes and their putative miRNA targets in plants has

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

been devised by Yin *et al.* (2008). In this method (Fig. 4), specialized BLAST searches are performed, optimized for the detection of short, well-matched sequences. miRNAs are identified from the results of these searches by their capacity to form stable hairpin structures, crucially involving base-pairing between the sequences that will form the mature double-stranded miRNA duplex. Potential target sequences in non-models can then be identified in transcriptomic or genomic databases, again based on BLAST-searching for perfect or close matches to mature miRNA sequences. The targets of small RNA regulators that induce the cleavage of mRNAs can also be identified from degradome sequencing (e.g. Formey *et al.*, 2015).

Numerous developmental processes in plants are regulated by a class of siRNA termed *trans*-acting siRNA (tasiRNAs) which, like miRNAs, are able to target specific mRNAs in the cytoplasm by base-pairing with these as part of a RISC. These small RNAs are derived from *TAS* genes whose primary transcripts, produced by RNA Pol II, are rendered double-stranded by the action of RNA-DEPENDENT RNA POLYMERASE 6 (RDR6), as reviewed by Borges and Martienssen (2015). TasiRNAs are then produced by the progressive cleavage of the double-stranded *TAS* transcripts into 21–22 nt fragments by the action of DCL4, the initial position of cleavage being determined by interactions with a guiding miRNA. Primary *TAS* transcripts may possess either one miRNA target site or two such sites flanking the mature tasiRNA sequences to be generated (Axtell *et al.*, 2006). Bioinformatics workflows can be devised to identify *TAS* gene orthologues in any sequenced plant genome based on the presence and positions of tasiRNA sequences and the miRNA-targeted sites. The downstream protein-coding genes targets of tasiRNAs can then be identified by bioinformatics analyses, as for the targets of miRNAs (e.g. Yin *et al.*, 2008).

Conclusions on the presence or absence of post-transcriptional regulation in a given situation will usually require studies of the expression of the small RNA of interest and its putative targets. It should be borne in mind that some interactions between small RNAs and their targets lead to the near-total elimination of target mRNAs whereas others act quantitatively such that a measurable level of both the small RNA regulator and its targets may be present in the same cells (e.g. Nikovics *et al.*, 2006). Expression levels of mature small RNAs can be investigated by small RNA-seq, as mentioned above, and many small RNAs are also sufficiently highly expressed to enable their detection by *in situ* hybridization using locked nucleic acid (LNA) probes (Valoczi *et al.*, 2004). Alternatively, the expression of small RNAs can be studied through the detection of their primary transcripts, either in RNA-seq data-sets or by Reverse Transcriptase-PCR (RT-PCR). Conclusions on the presence or absence of a regulatory interaction between a small RNA of interest and its putative targets may also need to take into account the spectrum of AGO proteins expressed in the tissue in question, as discussed by Borges and Martienssen (2015).

Epigenetic regulation

A large component of transcriptional regulation in eukaryotes occurs via epigenetic modifications to chromatin. As an example, Kaufmann *et al.* (2010) cite a simplified regulatory network of around 20 genes that control the transition to flowering in *A. thaliana*, only three of which are not known to be specifically regulated through epigenetic mechanisms. As reviewed by Engelhorn *et al.* (2014), features of chromatin landscapes that contribute to the regulation of gene expression in plants include nucleosome positioning, post-translational

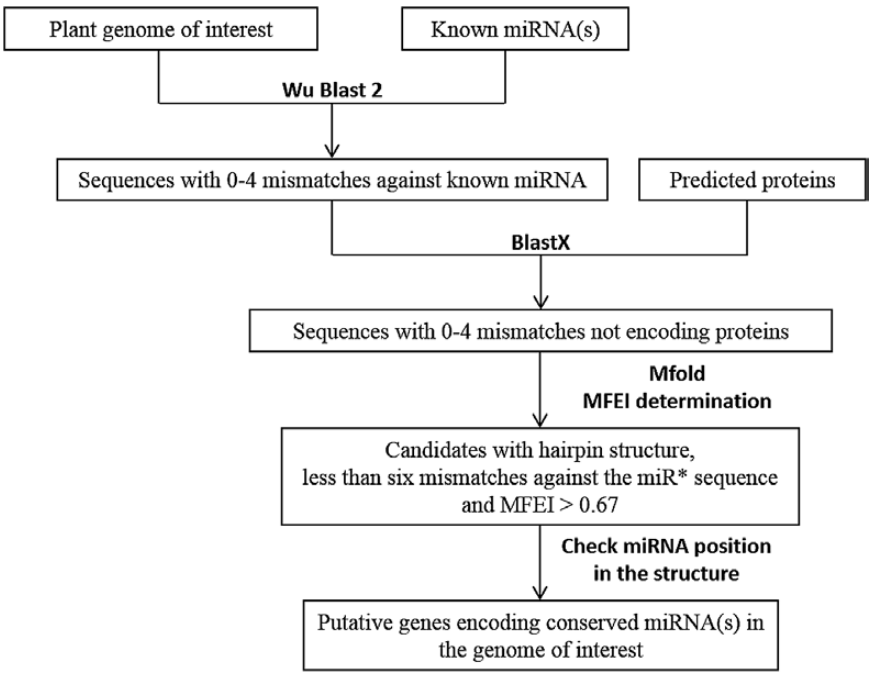

Fig. 4. A bioinformatics workflow for the identification of MIRNA orthologues from sequenced plant genomes (after Yin *et al.*, 2008).

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

modifications of histones, the presence of histone variants, and the direct cytosine methylation of DNA.

As for post-transcriptional regulation, the details of epigenetic regulation in plants are under intense investigation in model systems, chief among which is *A. thaliana*. However, an expanding number of techniques that permit the whole-genome survey of chromatin landscapes in model plants can also be used in any non-model species, conditionally on the availability of its genome sequence. Thus the key regulatory characteristics of chromatin in different tissues and over a range of developmental stages can be monitored in non-models plants of use in evo-devo research.

The openness of chromatin and hence the accessibility of *cis*-regulatory regions to the transcriptional machinery is largely controlled by the positioning of nucleosomes. Nucleosomes can be actively slid along chromatin by SWI/SNF-related enzymes and this mechanism is known to control gene expression in diverse developmental processes including stem cell maintenance (Kwon *et al.*, 2005), embryogenesis (Sang *et al.*, 2012), and root development (Aichinger *et al.*, 2011). The openness of chromatin can be globally mapped in plants with sequenced genomes using techniques such as DNase-seq (Du *et al.*, 2013) and Fomaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq; Omidbakhshfard *et al.*, 2014). In the former technique, isolated chromatin is digested with DNase I and subjected to NGS analysis thereby identifying cut-sites at the start of sequence reads. The frequency of cutting at each site gives a measure of its accessibility, correlating negatively with nucleosome occupancy. In FAIRE-seq, nucleosome-depleted regions of chromatin that are bound by transcription factors are preferentially isolated and sequenced following formaldehyde cross-linking.

Post-translational modifications of the histone proteins of which nucleosomes are composed include acetylation, methylation, phosphorylation, and ubiquitination (reviewed by Berr *et al.*, 2011). The deposition of these epigenetic marks can lead to the activation or repression of transcription at the locus concerned. Histone acetylation is generally linked to transcriptional activation (Zhang *et al.*, 2015) and targets for this form of modification in *A. thaliana* include five lysine residues in each of the histones H3 and H4. Histone acetylation/deacetylation is typically a dynamic process regulated by the opposing activities of Histone Acetyl Transferases (HACs) and Histone Deactyl Transferases (HDACs) both of which are encoded by multigene families in *A. thaliana*. Histone methylation, by contrast, can be linked either to the repression or activation of transcription. For example, three different Polycomb Repressive Complex 2 (PRC2) complexes act in *A. thaliana* to depose stable repressive trimethyl marks (me3) on Lysine 27 of H3 (H3K27me3), whereas the Trithorax Group (TxG) complex acts to depose activating H3K4me3 marks (reviewed by Berr *et al.*, 2011). As for methylation, the ubiquitination of histones can be associated with either the activation or repression of transcription: monoubiquitination of H2A which is mediated in *A. thaliana* by the PRC1-like complex is known to stably repress transcription, whereas monoubiquination of H2B is associated with transcriptional activation, as reviewed by Engelhorn *et al.* (2014).

Although some differences exist in the mechanisms of histone modification and in the precise effects of certain marks on DNA accessibility between distantly related eukaryotes (Shu *et al.*, 2012) there is, nonetheless, a general degree of conservation in the marks deposed and their effects on transcription (Suganuma and Workman, 2011; van Steensel, 2011). This broad conservation of histone modifications makes it possible to predict the likely repressed or activated state of chromatin at a given locus in a non-model species through a detailed analysis of associated post-translational histone modifications.

Post-translational modifications of histones can be globally mapped using ChIP-seq procedures based on antibodies against the marks of interest (e.g. Ha *et al.*, 2011). Such methods have been applied to numerous epigenetic marks in *A. thaliana* to define four main chromatin states (Roudier *et al.*, 2011). As histones are among the most highly conserved proteins in nature and as their epigenetic marks are also highly conserved, even between plants and metazoans, antibodies and methods developed for research in model species can be used without difficulty in non-models. Signals obtained in ChIP-seq procedures using anti-histone antibodies are typically tens to hundreds of fold higher than those obtained using gene-specific transcription factors, greatly simplifying the mapping of histone modifications in non-model plants. A further method, based on mass-spectroscopy, has also recently been applied to study histone post-translational modifications in the genomes of diatoms (Veluchamy *et al.*, 2015) and sugarcane (Moraes *et al.*, 2015), among other organisms.

The presence of histone variants which can replace canonical histones within nucleosomes is also known to control transcriptional activation and repression in plants and other eukaryotes. Specific histone variants may be associated with particular loci, tissues, and developmental stages and may also combine with specific epigenetic marks to regulate gene expression. As for the post-translational modification of histones, histone variants can be studied by ChIP-seq using antibodies specific to the histone variants of interest (Ku *et al.*, 2012).

A final type of epigenetic modification that can be compared between model and non-model plants in the context of evo-devo analyses involves the direct cytosine-methylation of DNA. This type of covalent modification is frequently associated with non-expressed regions of the genome, although it also occurs in promoters and may thus contribute to tissue-specific expression through transcriptional silencing (Zhang *et al.*, 2006). A further role of cytosine methylation, within transcribed regions, may be to suppress the activity of cryptic promoter elements that would otherwise lead to the production of truncated or nonsense transcripts (Zilberman *et al.*, 2007). Global patterns of cytosine methylation can be determined in any organism with a sequenced genome using Reduced Representation Bisulphite Sequencing (RRBS; Gu *et al.*, 2011; Seymour *et al.*, 2014). In this method, extracted genomic DNA is treated with bisulphite solution which converts unmethylated cytosine to uracil residues. DNA-seq is then performed and the resulting dataset compared with the

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

reference genome sequence to determine the frequency of methylation at each cytosine residue in the genome.

Data integration for the assembly of Gene Regulatory Networks and inferences on ancestral network topologies

In the interests of simplicity in this article we treat transcriptional, post-transcriptional, and epigenetic regulatory interactions as separate phenomena that can be studied in an evo-devo context using discrete methods of analysis. However, these regulatory processes are clearly intimately linked. For example, distinct classes of transcription factors control gene expression by operating upstream or downstream of enzymes that modify chromatin landscapes (Engelhorn *et al.*, 2014) while specific classes of small RNA regulators are also known to bring about transcriptional silencing through epigenetic modifications (Castel and Martienssen, 2013). The interrelatedness of the forms of developmental regulation discussed here has important practical consequences for the study of GRNs. The degree of confidence that can be accorded to, for example, a putative transcription factor binding site in a given target promoter may be improved if that site correlates with the activating epigenetic marks of an open chromatin landscape (Fig. 3). Thus, the integration of data from transcriptomic, genomic, and epigenomic surveys is particularly important for the assembly of GRNs in non-model plants in which other forms of functional experimentation may be impractical or impossible. The comparison of these different types of data-sets is indicated in the last few steps in the procedure outlined in Fig. 3. The task of data integration for plant evo-devo studies could be facilitated by the use of specialized bioinformatics platforms designed to correlate and integrate different types of large-scale data-sets, similar to those already available in some animal systems (Roy *et al.*, 2010; Dunham *et al.*, 2012; Yue *et al.*, 2014).

A further requirement for the analysis and comparison of GRNs is to be able to represent these graphically and numerous programmes are available for this. For example, BioTapestry (Longabaugh *et al.*, 2005) has been widely used for the detailed representation of transcriptional circuitry while Cytoscape (Shannon *et al.*, 2003) is particularly useful for evo-devo analyses as it can be used to compare network topologies in different species.

Protein Resurrection for the direct study of ancestral Gene Regulatory Network components

The previous sections of this review outlined experimental approaches that can be used to compare GRNs in living taxa and thereby infer ancestral network topologies in the common ancestors of those species. However, a further evo-devo approach, termed Protein Resurrection (or Ancestral Sequence Reconstruction), reverses the order of experimentation and inference. Using this method, the sequences of ancestral genes or proteins are first inferred using specialized techniques of molecular phylogeny based on the computation of posterior probabilities for all possible states of each site in ancestral sequences (reviewed by Thornton, 2004). Coding sequences corresponding to inferred ancestral genes can then be physically constructed from which the encoded proteins can be expressed and studied *in vitro* and/or *in vivo* to assay their activities, properties, and molecular interactions directly. This 'Resurrection' approach is, in principle, applicable to the study of any regulatory protein including receptors, transducers, transcription factors, protein kinases and phosphates, and chromatin remodelling enzymes etc (Harms and Thornton, 2014; McKeown *et al.*, 2014). Recent advances in Protein Resurrection demonstrate that accuracy in sequence reconstruction can be increased by constraining gene/protein phylogenies in molecular families of interest to the more accurately reconstructed topologies of organismal phylogenies (Groussin *et al.*, 2015).

Following ancestral reconstruction work, coding sequences corresponding to resurrected ancestral proteins can be custom-ordered from commercial suppliers and expressed in heterologous systems such as *E. coli* for the preparation of free native or recombinant protein for use in *in vitro* assays. For example, ancestral transcription factors or their isolated DNA-binding regions can be subjected to SELEX or PBM analyses as can their present-day descendants (as described in the section on Transcriptional Regulation). Such experiments may reveal which descendant transcription factors in living taxa have conserved the DNA-binding activity of their (inferred) common ancestor and which have undergone a change in their binding-site preferences. Resurrected genes can also be expressed, using present-day *cis*-regulatory sequences, in transgenic plants to test, for example, their ability to complement mutations in related genes or to generate mis-expression phenotypes. Such genes can also be expressed in yeast-two-hybrid experiments to test their interactions with other (present-day or ancestral) regulatory proteins. All of the above types of *in vitro* and *in vivo* experiments can be used to confirm predictions relating to ancestral GRNs that have been inferred by the comparison of GRNs between living taxa. It is possible in some ancestral reconstructions that sequence ambiguities may arise in particular residues. In such cases, researchers may have to construct genes that encode both (or all) of the inferred alternative ancestral proteins in order to test whether their biochemical properties differ from each other.

A further potential use of Protein Resurrection which may be particularly applicable to plant evo-devo makes use of the whole genome duplication (WGD) events that are more frequently found in plant than in metazoan evolution. Many of these WGDs appear to correlate with the emergence of major new plant clades such as the zeta duplication (Jiao *et al.*, 2011) that occurred before the most recent common ancestor (MRCA) of living seed plants, the epsilon duplication (Jiao *et al.*, 2011) that occurred before the MCRA of living angiosperms, and the gamma triplication (Bowers *et al.*, 2003; Tang *et al.*, 2008) that occurred before the MCRA of living core eudicots (Fig. 5). Many more recent, clade-specific

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

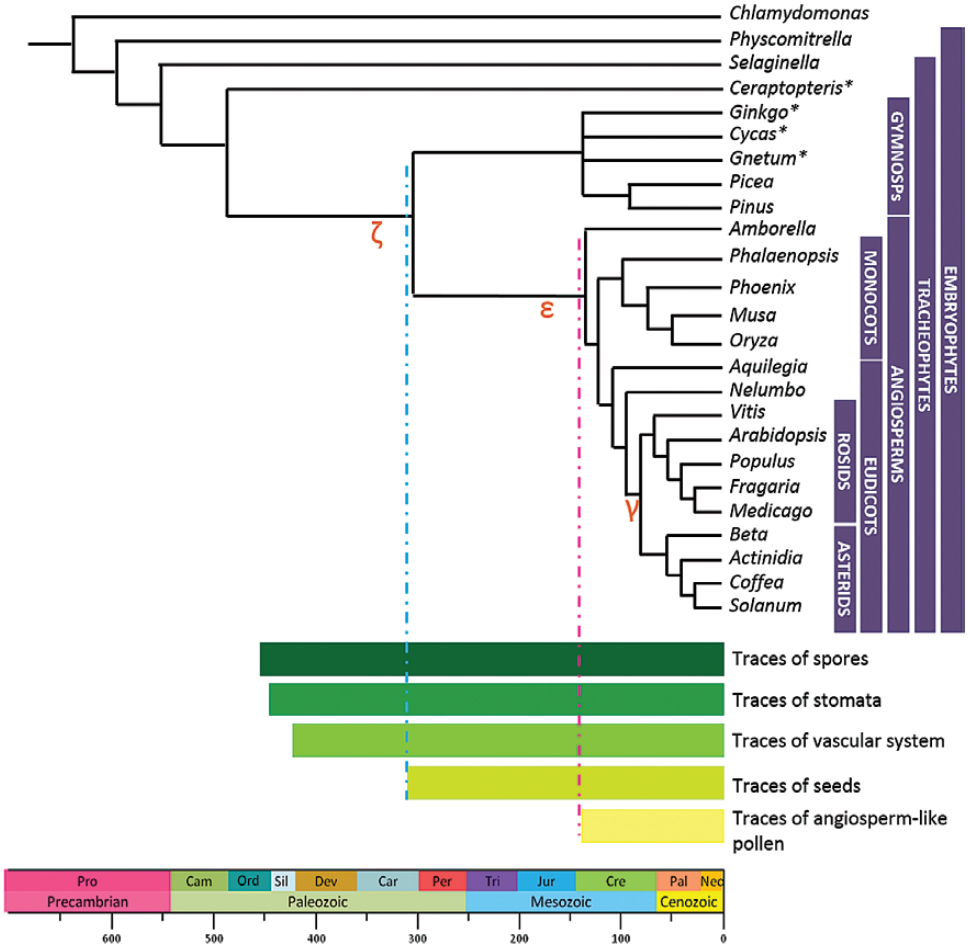

Fig. 5. A selection of major evolutionary novelties in land plant evolution and a schematic phylogeny of key living taxa for the evo-devo investigation of Gene Regulatory Networks. Timings for the earliest fossil evidence of land plant features are taken from Edwards and Kenrick (2015), Serbet and Rothwell (1992), and Barrett and Willis (2001). Two nodes on the phylogeny (vertical broken lines) have been constrained to the earliest evidence of seeds and angiosperm pollen, respectively. All taxa shown, except those marked with an asterisk, include at least one species with a sequenced genome (Table 1). Approximate positions of the zeta and epsilon whole-genome duplications, and of the gamma whole-genome triplication, are shown on the stem lineages of the seed plants, angiosperms, and core eudicots, respectively.

WGD events can also be inferred from present-day taxa (Cui *et al.*, 2006). Classical evo-devo approaches in which living taxa and their components are studied directly are limited to inferences on the topology of ancestral GRNs situated at nodes of the organismal phylogenetic tree that correspond to speciation events. Protein Resurrection, by contrast, can be used to infer ancestral sequences at nodes in gene/protein trees that were caused by gene (or genome) duplication events. Thus, Protein Resurrection can be used to infer the sequences of ancestral components of GRNs that occurred along the stem lineages of major new clades of organisms. For example, the initial radiation of the living angiosperm which perhaps occurred some 150 million years ago (MYA) was preceded by a long stem lineages (of perhaps 150 MY) along which occurred the epsilon WGD event (Jiao *et al.* 2011). This event or more localized gene duplications that took place over the same period appears to have generated many pairs of paralogues that play important roles in flower development in model angiosperms (Pabon-Mora *et al.*, 2014). Using Protein Resurrection, it should be possible to generate the ancestors of these genes and proteins from a stage relatively soon before the origin of the flower and, in further studies, experimentally determine any features of these molecules that may have changed and thereby contributed to the origin of angiosperm-specific morphological features.

The integration of inferred ancestral Gene Regulatory Networks with non-molecular data to provide more comprehensive insights into plant evolution

The main objective of this article has been to outline methods for the assembly of GRNs in model and non-model plant species with a view to the inference and testing (using Protein Resurrection, for example) of ancestral GRNs. Apparent changes to network topologies during evolution can then be correlated with evolutionary transitions at the morphological level and the emergence of new biological forms. Figure 5 shows some of the major transitions and evolutionary novelties acquired during land plant evolution, together with an organismal phylogeny showing the positions of key taxa

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

with sequenced genomes and of some of the WGD events that appear to have fuelled the processes of neo-functionalization during plant evolution. However, the significance of evolutionary changes to the topology of GRNs and their further effects on plant development and morphology may only become fully apparent in the context of the ecological, geographical, and climatic factors pertaining at the time of those changes. Accordingly, a full understanding of the processes that have shaped plant evolution can only be achieved by combining molecular data with information from many other sources to construct robust and complete evolutionary hypotheses. At present, molecular and paleobotanical lines of evolutionary enquiry are pursued mostly by separate groups of researchers and tend to generate data-sets with very different characteristics. Critically, therefore, it will be necessary to develop new methods of data integration to make robust and comprehensive inferences on plant evolution. The achievement of this objective will almost certainly require a greater degree of interdisciplinary collaboration and exchange between evolutionary biologists using molecular, whole plant, paleobotanical, and paleoecological approaches.

Acknowledgements

Work in our laboratories is funded by research grant ANR-13-BSV2-0009 'ORANGe'. AV-G is also funded by an ENS-Lyon research and teaching position.

References

Aichinger E, Villar CBR, Di Mambro R, Sabatini S, Koehler C. 2011. The CHD3 chromatin remodeler PICKLE and polycomb group proteins antagonistically regulate meristem activity in the Arabidopsis root. The Plant Cell **23,** 1047–1060.

Altschul S, Gish W, Miller W, Myers E, Lipman D. 1990. BASIC LOCAL ALIGNMENT SEARCH TOOL. Journal of Molecular Biology **215,** 403–410.

Axtell MJ, Jan C, Rajagopalan R, Bartel DP. 2006. A two-hit trigger for siRNA biogenesis in plants. Cell **127,** 565–577.

Barrett PM, Willis KJ. 2001. Did dinosaurs invent flowers? Dinosaur–angiosperm coevolution revisited. Biological Reviews **76,** 411–447.

Berger MF, Bulyk ML. 2009. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nature Protocols **4,** 393–411.

Berr A, Shafiq S, Shen W-H. 2011. Histone modifications in transcriptional activation during plant development. Biochimica et Biophysica Acta-Gene Regulatory Mechanisms **1809,** 567–576.

Bigot T, Daubin V, Lassalle F, Perriere G. 2013. TPMS: a set of utilities for querying collections of gene trees. BMC Bioinformatics **14,** 109.

Boer DR, Freire-Rios A, van den Berg WAM, *et al.* 2014. Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. Cell **156,** 577–589.

Borges F, Martienssen RA. 2015. The expanding world of small RNAs in plants. Nature Reviews Molecular Cell Biology **16,** 727–741.

Boussau B, Walton Z, Delgado JA, *et al.* 2014. Strepsiptera, phylogenomics and the long branch attraction problem. Plos One **9,** e107709.

Bowers JE, Chapman BA, Rong JK, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422,** 433–438.

Bremer B, Bremer K, Chase MW, *et al.* 2009. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Botanical Journal of the Linnean Society **161,** 105–121.

Carroll SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134,** 25–36.

Castel SE, Martienssen RA. 2013. RNA interference in the nucleus: roles for small RNAs in transcription, epigenetics and beyond. Nature Reviews Genetics **14,** 100–112.

Crosetto N, Bienko M, van Oudenaarden A. 2015. Spatially resolved transcriptomics and beyond. Nature Reviews Genetics **16,** 57–66.

Cui L, Wall PK, Leebens-Mack JH, *et al.* 2006. Widespread genome duplications throughout the history of flowering plants. Genome Research **16,** 738–749.

Doudna JA, Charpentier E. 2014. The new frontier of genome engineering with CRISPR-Cas9. Science **346,** 1258096.

Du Z, Li H, Wei Q, *et al.* 2013. Genome-wide analysis of histone modifications: H3K4me2, H3K4me3, H3K9ac, and H3K27ac in *Oryza sativa* L. Japonica. Molecular Plant **6,** 1463–1472.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perriere G. 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics **21,** 2596–2603.

Dunham I, Kundaje A, Aldred SF, *et al.* 2012. An integrated encyclopedia of DNA elements in the human genome. Nature **489,** 57–74.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5,** 1–19.

Edwards D, Kenrick P. 2015. The early evolution of land plants, from fossils to genomics: a commentary on Lang (1937) 'On the plant-remains from the Downtonian of England and Wales'. Philosophical transactions of the Royal Society of London, Series B, Biological Sciences **370**.

Engelhorn J, Blanvillain R, Carles CC. 2014. Gene activation and cell fate control in plants: a chromatin perspective. Cellular and Molecular Life Sciences **71,** 3119–3137.

Finet C, Fourquin C, Vinauger M, Berne-Dedieu A, Chambrier P, Paindavoine S, Scutt CP. 2010*a*. Parallel structural evolution of auxin response factors in the angiosperms. The Plant Journal **63,** 952–959.

Finet C, Timme RE, Delwiche CF, Marletaz F. 2010*b*. Multigene phylogeny of the green lineage reveals the origin and diversification of land plants. Current Biology **20,** 2217–2222.

Formey D, Pedro Iniguez L, Pelaez P, Li Y-F, Sunkar R, Sanchez F, Luis Reyes J, Hernandez G. 2015. Genome-wide identification of the *Phaseolus vulgaris* sRNAome using small RNA and degradome sequencing. BMC Genomics **16,** 423.

Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R. 2014. DNA-binding specificities of plant transcription factors and their potential to define target genes. Proceedings of the National Academy of Sciences, USA **111,** 2367–2372.

Godoy M, Franco-Zorrilla JM, Perez-Perez J, Oliveros JC, Lorenzo O, Solano R. 2011. Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors. The Plant Journal **66,** 700–711.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics **27,** 1017–1018.

Groussin M, Hobbs JK, Szoellosi GJ, Gribaldo S, Arcus VL, Gouy M. 2015. Toward more accurate ancestral protein genotype–phenotype reconstructions with the use of species tree-aware gene trees. Molecular Biology and Evolution **32,** 13–22.

Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. 2011. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nature Protocols **6,** 468–481.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology **52,** 696–704.

Ha M, Ng DW-K, Li W-H, Chen ZJ. 2011. Coordinated histone modifications are associated with gene expression variation within and between species. Genome Research **21,** 590–598.

Harms MJ, Thornton JW. 2014. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. Nature **512,** 203–207.

Hay AS, Pieper B, Cooke E, *et al.* 2014. *Cardamine hirsuta*: a versatile genetic system for comparative studies. The Plant Journal **78,** 1–15.

Hellens RP, Allan AC, Friel EN, Bolitho K, Grafton K, Templeton MD, Karunairetnam S, Gleave AP, Laing WA. 2005. Transient expression

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

vectors for functional genomics, quantification of promoter activity and RNA silencing in plants. Plant Methods **1,** 13.

Jasinski S, Vialette-Guiraud ACM, Scutt CP. 2010. The evolutionary-developmental analysis of plant microRNAs. Philosophical Transactions of the Royal Society B-Biological Sciences **365,** 469–476.

Jiao Y, Wickett NJ, Ayyampalayam S, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. Nature **473,** 97–100.

Jolma A, Yan J, Whitington T, *et al.* 2013. DNA-binding specificities of human transcription factors. Cell **152,** 327–339.

Kaufmann K, Muino JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC. 2009. Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLOS Biology **7,** 854–875.

Kaufmann K, Pajoro A, Angenent GC. 2010. Regulation of transcription in plants: mechanisms controlling developmental switches. Nature Reviews Genetics **11,** 830–842.

Kersting AR, Bornberg-Bauer E, Moore AD, Grath S. 2012. Dynamics and adaptive benefits of protein domain emergence and arrangements during plant genome evolution. Genome Biology and Evolution **4,** 316–329.

Korhonen J, Martinmaki P, Pizzi C, Rastas P, Ukkonen E. 2009. MOODS: fast search for position weight matrix matches in DNA sequences. Bioinformatics **25,** 3181–3182.

Ku M, Jaffe JD, Koche RP, Rheinbay E, Endoh M, Koseki H, Carr SA, Bernstein BE. 2012. H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions. Genome Biology **13,** R85.

Kwon CS, Chen CB, Wagner D. 2005. WUSCHEL is a primary target for transcriptional regulation by SPLAYED in dynamic control of stem cell fate in Arabidopsis. Genes & Development **19,** 992–1003.

Lemay P, Salemi M, Vandamme A-M. 2009. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing* . Cambridge: Cambridge University Press.

Longabaugh WJR, Davidson EH, Bolouri H. 2005. Computational representation of developmental genetic regulatory networks. Developmental Biology **283,** 1–16.

Maddison WP, Maddison DR. 2015. Mesquite: a modular system for evolutionary analysis . Version 3.04 http://mesquiteproject.org

Majka J, Speck C. 2007. Analysis of protein-DNA interactions using surface plasmon resonance. In: Seitz H, ed. *Analytics of protein–DNA interaction*, 13–36.

Matasci N, Hung L-H, Yan Z, *et al.* 2014. Data access for the 1,000 plants (1KP) project. Gigascience **3,** 17.

Mathelier A, Fornes O, Arenillas DJ, *et al.* 2016. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Research **44,** D110–115.

Mathelier A, Wasserman WW. 2013. The next generation of transcription factor binding site prediction. PLOS Computational Biology **9,** e1003214.

McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. 2014. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. Cell **159,** 58–68.

Medina-Rivera A, Defrance M, Sand OJ, *et al.* 2015. RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Research **43,** W50–W56.

Melzer R, Verelst W, Theissen G. 2009. The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in floral quartet-like complexes *in vitro*. Nucleic Acids Research **37,** 144–157.

Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics **12,** 116.

Minguet EG, Segard S, Charavay C, Parcy F. 2015. MORPHEUS, a webtool for transcription factor binding analysis using Position Weight Matrices with dependency. PLOS One **10,** e0135586.

Moraes I, Yuan Z-F, Liu S, Souza GM, Garcia BA, Armando Casas-Mollano J. 2015. Analysis of histones H3 and H4 reveals novel and conserved post-translational modifications in sugarcane. PLOS One **10,** e0134586

Moyroud E, Reymond MCA, Hames C, Parcy F, Scutt CP. 2009. The analysis of entire gene promoters by surface plasmon resonance. The Plant Journal **59,** 851–858.

Muino JM, de Bruijn S, Pajoro A, Geuten K, Vingron M, Angenent GC, Kaufmann K. 2016. Evolution of DNA-binding sites of a floral master regulatory transcription factor. Molecular Biology and Evolution **33,** 185–200.

Nikovics K, Blein T, Peaucelle A, Ishida T, Morin H, Aida M, Laufs P. 2006. The balance between the *MIR164A* and *CUC2* genes controls leaf margin serration in *Arabidopsis*. The Plant Cell **18,** 2929–2945.

Omidbakhshfard MA, Winck FV, Arvidsson S, Riano-Pachon DM, Mueller-Roeber B. 2014. A step-by-step protocol for formaldehyde-assisted isolation of regulatory elements from *Arabidopsis thaliana*. Journal of Integrative Plant Biology **56,** 527–538.

Pabon-Mora N, Wong GK-S, Ambrose BA. 2014. Evolution of fruit development genes in flowering plants. Frontiers in Plant Science **5,** 300.

Pantalacci S, Semon M. 2015. Transcriptomics of developing embryos and organs: a raising tool for evo-devo. Journal of Experimental Zoology, Part B-Molecular and Developmental Evolution **324,** 363–371.

Para A, Li Y, Marshall-Colon A, *et al.* 2014. Hit-and-run transcriptional control by bZIP1 mediates rapid nutrient signaling in *Arabidopsis*. Proceedings of the National Academy of Sciences, USA **111,** 10371–10376.

Poulsen C, Vaucheret H, Brodersen P. 2013. Lessons on RNA silencing mechanisms in plants from eukaryotic argonaute structures. The Plant Cell **25,** 22–37.

Price MN, Dehal PS, Arkin AP. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution **26,** 1641–1650.

Reymond MC, Brunoud G, Chauvet A, Martinez-Garcia JF, Martin-Magniette M-L, Moneger F, Scutt CP. 2012. A light-regulated genetic module was recruited to carpel development in *Arabidopsis* following a structural change to SPATULA. The Plant Cell **24,** 2812–2825.

Rhee HS, Pugh BF. 2011. Comprehensive genome-wide protein–DNA interactions detected at single-nucleotide resolution. Cell **147,** 1408–1419.

Roudier F, Ahmed I, Berard C, *et al.* 2011. Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. EMBO Journal **30,** 1928–1938.

Roy S, Ernst J, Kharchenko PV, *et al.* 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science **330,** 1787–1797.

Samanta S, Thakur JK. 2015. Importance of Mediator complex in the regulation and integration of diverse signaling pathways in plants. Frontiers in Plant Science **6,** 757.

Sang Y, Silva-Ortega CO, Wu S, Yamaguchi N, Wu M-F, Pfluger J, Gillmor CS, Gallagher KL, Wagner D. 2012. Mutations in two non-canonical Arabidopsis SWI2/SNF2 chromatin remodeling ATPases cause embryogenesis and stem cell maintenance defects. The Plant Journal **72,** 1000–1014.

Serbet R, Rothwell G. 1992. Characterising the most primitive seed ferns. 1. A reconstruction of *Elkinsia polymorpha*. International Journal of Plant Sciences **153,** 602–621.

Seymour DK, 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,** e1004785.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research **13,** 2498–2504.

Shu H, Wildhaber T, Siretskiy A, Gruissem W, Hennig L. 2012. Distinct modes of DNA accessibility in plant chromatin. Nature Communications **3,** 1281.

Suganuma T, Workman JL. 2011. Signals and combinatorial functions of histone modifications. In: Kornberg RD, Raetz CRH, Rothman JE, Thorner JW, eds. *Annual Review of Biochemistry*, **80,** 473–499.

Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. Genome Research **18,** 1944–1954.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annual Review of Genetics **38,** 615–643.

Taylor RS, Tarver JE, Hiscock SJ, Donoghue PCJ. 2014. Evolutionary history of plant microRNAs. Trends in Plant Science **19,** 175–182.

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016

Thomas-Chollier M, Hufton A, Heinig M, O'Keeffe S, El Masri N, Roider HG, Manke T, Vingron M. 2011. Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. Nature Protocols **6,** 1860–1869.

Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. Nature Reviews Genetics **5,** 366–375.

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

Valoczi A, Hornyik C, Varga N, Burgyan J, Kauppinen S, Havelda Z. 2004. Sensitive and specific detection of microRNAs by Northern blot analysis using LNA-modified oligonucleotide probes. Nucleic Acids Research **32,** e175.

Van Steensel B. 2011. Chromatin: constructing the big picture. EMBO Journal **30,** 1885–1895.

Vazquez J. 2004. Phylogenetics - MacClade 4: analysis of phylogeny and character evolution, version 4.06. American Biology Teacher **66,** 511–512.

Veluchamy A, Rastogi A, Lin X, *et al.* 2015. An integrative analysis of post-translational histone modifications in the marine diatom *Phaeodactylum tricornutum*. Genome Biology **16,** 102.

Wagner D, Wellmer F, Dilks K, William D, Smith MR, Kumar PP, Riechmann JL, Greenland AJ, Meyerowitz EM. 2004. Floral induction in tissue culture: a system for the analysis of LEAFY-dependent gene regulation. The Plant Journal **39,** 273–282.

Wickett NJ, Mirarab S, Nguyen N, *et al.* 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proceedings of the National Academy of Sciences, USA **111,** E4859–E4868.

Workman CT, Yin YT, Corcoran DL, Ideker T, Stormo GD, Benos PV. 2005. enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Research **33,** W389–W392.

Yin Z, Li C, Han M, Shen F. 2008. Identification of conserved microRNAs and their target genes in tomato (Lycopersicon esculentum). Gene **414,** 60–66.

Yue F, Cheng Y, Breschi A, *et al.* 2014. A comparative encyclopedia of DNA elements in the mouse genome. Nature **515,** 355–364.

Zhang W, Garcia N, Feng Y, Zhao H, Messing J. 2015. Genome-wide histone acetylation correlates with active transcription in maize. Genomics **106,** 214–220.

Zhang X, Yazaki J, Sundaresan A, *et al.* 2006. Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. Cell **126,** 1189–1201.

Zilberman D, Gehring M, Tran RK, Ballinger T, Henikoff S. 2007. Genome-wide analysis of *Arabidopsis thaliana* DNA methylation uncovers an interdependence between methylation and transcription. Nature Genetics **39,** 61–69.

Downloaded from http://jxb.oxfordjournals.org/ by guest on March 23, 2016