Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2012, **8**, 2566–2574

www.rsc.org/molecularbiosystems

PAPER

LASSO modeling of the *Arabidopsis thaliana* seed/seedling transcriptome: a model case for detection of novel mucilage and pectin metabolism genes

Aleksandar Vasilevski,†[a] Federico M. Giorgi,†[ab] Luca Bertinetti[c] and Björn Usadel*[ade]

Received 15th March 2012, Accepted 10th May 2012
DOI: 10.1039/c2mb25096a

Whole genome transcript correlation-based approaches have been shown to be enormously useful for candidate gene detection. Consequently, simple Pearson correlation has been widely applied in several web based tools. That said, several more sophisticated methods based on *e.g.* mutual information or Bayesian network inference have been developed and have been shown to be theoretically superior but are not yet commonly applied. Here, we propose the application of a recently developed statistical regression technique, the LASSO, to detect novel candidates from high throughput transcriptomic datasets. We apply the LASSO to a tissue specific dataset in the model plant *Arabidopsis thaliana* to identify novel players in *Arabidopsis thaliana* seed coat mucilage synthesis. We built LASSO models based on a list of genes known to be involved in a sub-pathway of Arabidopsis mucilage synthesis. After identifying a putative transcription factor, we verified its involvement in mucilage synthesis by obtaining knock-out mutants for this gene. We show that a loss of function of this putative transcription factor leads to a significant decrease in mucilage pectin.

Introduction

Transcriptional coordination, also called co-expression or co-regulation, has been observed in several biological contexts between functionally related genes.[1,2] This is likely because the encoded proteins have to be present at the same time in order to functionally co-operate in the same pathway or within the same complex. While this does not imply that the underlying transcripts have to be co-expressed across an exhaustive range of different conditions, the assumption that genes that always correlate might have a similar role or might be involved in similar pathways is a valid starting hypothesis for finding new genes, a paradigm often dubbed "guilt-by-association principle".[3] This principle assumes that inferences can be made concerning the function of previously unknown molecular species (*e.g.* genes) based on the fact that they behave similarly to already characterized genes. In the field of transcriptomics, the guilt-by-association principle can be applied through co-expression, that is the assessment of transcriptionally similar behavior between two or more genes. Co-expression has been successfully exploited to find new genes in a range of model organisms, including yeast,[2] humans[4] and other mammals.[3] Using this "guilt by association" approach, transcriptome-wide gene function inference and biological pathway discovery have been shown to be possible.[5–7] To name but a few examples from the mammalian field, it has been employed successfully in the mouse transcriptome: a list of genes known by literature search to be involved in retina-related processes was used to generate a network of 673 genes with similar expression behavior, and finally a list of novel retina disease-associated genes was successfully predicted.[8] In another example, the properties of cancer proteins in protein–protein interaction networks were proven to be highly discriminatory in terms of network degree, clustering coefficient and occupancy of specific network motifs, therefore paving the way for novel cancer genes discovery in areas still poorly investigated of the human proteome.[9] The approach has also been used by integrating different species data, identifying novel human cancer genes through comparative analysis of plant–animal transcriptional behavior of DNA replication and repair genes.[10]

A particular success story, however, has been the model plant Arabidopsis where the application of this approach was

[a] *Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany. E-mail: b.usadel@bio1.rwth-aachen.de*
[b] *Institute of Applied Genomics, Parco Scientifico e Tecnologico di Udine, via J.Linussio 51, 33100 Udine, Italy*
[c] *Max Planck Institute of Colloids and Interfaces, Am Mühlenberg 1, 14476 Potsdam-Golm, Germany*
[d] *Institut für Biologie, RWTH Aachen, Worringer Weg 1, 52056 Aachen, Germany*
[e] *Institute of Bio- and Geosciences, IBG-2: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany*
† These authors contributed equally.

followed through meticulously and has resulted in the establishment of many different correlation databases, such as PlaNet[11] and AttedII[12] (see[13] for an overview). These databases have led to a better understanding of *e.g.* plant cell wall synthesis: in the case of cellulose synthase genes (CESAs), it has been demonstrated that several of these genes, which code for proteins which combine to form a functional CESA complex, are tightly co-expressed in *Arabidopsis thaliana*. Due to this tight behaviour, this complex was used to angle for other genes using co-expression. Indeed, several new genes were found to be co-expressed with these CESAs and characterized as displaying cellulose synthesis deficiencies.[14,15] Further examples where these databases and/or methods have been used for candidate gene discovery encompass different processes such as starch metabolism,[16,17] seed germination[18] and shade avoidance.[19] However, many of these approaches rely primarily on simple Pearson correlation, sometimes coupled with network clustering approaches[20] or sequence analysis.[21,22] Recently, more elaborate methods than Pearson correlation have flourished in the biostatistical literature, with the aim of increasing the accuracy of expression-based gene association inference.[23–25] These methods have been shown to be better or complementary to standard correlation approaches, at least for inferring causal gene interactions.[26] Although powerful, only a few of these algorithms have yet been used exhaustively in the plant field. One such technique is the Least Absolute Shrinkage and Selection Operator, or the LASSO. The LASSO is a recently proposed linear regression technique,[27] which explains an outcome variable y as a linear combination of independent predictor variables x_i: $y = b_0 + \sum b_i x_i$. Unlike standard regression, the LASSO imposes a limit to the weights assigned to the predictor variables:

$$|b_1| + |b_2| + |b_3| + \cdots + |b_n| \leq L_1$$

Here L_1 represents a tuning parameter for the stringency of the model. Because of the nature of the constraint, making L_1 sufficiently small will cause some of the coefficients to be exactly zero, so that several variables get discarded.[27] This increases the *interpretability* of LASSO models, as relevant variables can be clearly separated from irrelevant ones. The original algorithm to obtain the solution for LASSO at all possible sum-of-absolute-weight thresholds (referred to as L_1 thresholds) is a computationally very challenging task.[28] However a more efficient algorithm to get solutions for all LASSO models has been recently developed[28] and termed Least Angle Regression for LASSO, or simply LARS. In brief, LARS starts introducing an explanatory variable into the model and continues to increase its weight in the model until a second variable reaches the same correlation with the model's residuals as the initial variable. Then, the model proceeds modifying the weights of the two variables in a direction that is *equiangular* to both. This process balances all variables in the model, while excluding indirect effects, since increasing the weight of one variable also reduces the chance to include variables from the same informational area, similarly to what happens for partial correlation.[23] Since it is intrinsically an iterative, growing modeling process (in the LARS implementation), the LASSO can work in scenarios with more variables than samples (like *e.g.* microarrays or RNASeq datasets[29]). As such it provides a robust set of predictors and the capability of removing indirect connections, like conditional correlation.[30] The LASSO has been used extensively to generate well performing models where a clear border between important and unimportant variables had to be discerned,[31] although with only a handful of biological applications so far.[25,32,33] Its limited application to gene network reconstruction is however not too surprising, as typically the LASSO is used to predict one dependent variable by a linear combination of other variables and is thus more suitable to be used in biomarker discovery and statistical learning. Some exploratory studies exist in biological contexts: for example, the LASSO has been used in identifying genes coregulated with *StHRE* transcription factors during *Solanum tuberosum* tuber development, based on the data provided by less than twenty microarray samples.[31] Another study showed the potential of the LASSO in reverse engineering simulated gene expression data.[32] Altogether, these studies indicate that the LASSO might indeed be suitable for the guilt by association approach and might help to find candidate genes complementary to those found by more "classical" approaches.

We have previously shown that the LASSO can provide candidate genes in the case of tuber development hypoxia[34] and wanted to explore if this is also possible in Arabidopsis seed coat mucilage biosynthesis. In *Arabidopsis thaliana*, the seed coat is characterized by epidermal cells showing some specialized structures. Within the epidermal cells one can find the *columella*, which is a volcano-shaped secondary cell wall structure,[35] and which is surrounded by pectinaceous mucilage, arranged in a donut-shaped ring[36] under the primary cell wall separating the epidermal cell from outside. This mucilaginous material is released upon contact with water and then completely envelopes the seed.[36,37] Thus, mucilage has been suggested to be important for seed hydration and germination, attachment to soil components and for preventing gas exchange.[38,39] Once released, mucilage is characterized by a denser, relatively insoluble, inner layer and a more soluble outer layer, composed of sparsely branched rhamnogalacturonan I (RGI), a polysaccharide formed by succession of L-rhamnose-D-galacturonic acid dimers with side chains of arabinose, galactose and arabinogalactan residues.[36,37,40] Due to its composition, it can be considered as a model to study pectin biosynthesis. Unlike in other plant tissues, mucilage can be easily extracted from Arabidopsis seeds and *Arabidopsis* plants can tolerate the absence of mucilage under laboratory conditions.[35]

Mutations in a number of genes have been associated with altered mucilage production and/or release in the *Arabidopsis* seed coat.[41] These include several transcription factors and developmental regulators, such as *AP2*, and the factors *TTG1*, *TT8*, *EGL3*, *TT2*, *MYB5*, which comprise a WD–bHLH–MYB complex.[42] This complex and *AP2* modulate the expression of *GL2* and *TTG2* representing two transcriptional subpathways (Fig. 1).[43] In addition *MYB61* seems to drive independent genes.[44]

Furthermore, through screening of mucilage-defective mutants, five "MUcilage-Modified" (MUM) loci have been identified, which seem to act specifically in certain steps of

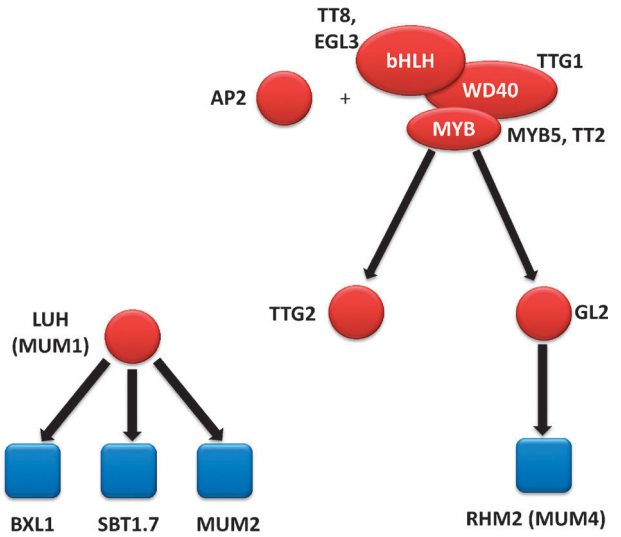

Fig. 1 Putative regulatory network for the seed coat mucilage pathway in *Arabidopsis thaliana*, inferred from the current literature.[46,50] Red circles represent transcription factors and blue squares represent genes with products characterized by enzymatic activity.

mucilage production and release.[45] A sub-pathway whose expression is at least partly positively regulated by *MUM1* (also known as *LUH* or Leunig Homolog[46,47]) has a key role in mucilage release and/or modification. This sub-pathway comprises *MUM2* (also known as *BGAL6* or beta-galactosidase 6[36,48]) as well as *BXL1* and *SBT1.7* and is required for pectin modification.[46,49–51] *MUM4* is clearly responsible for mucilage biosynthesis as cloning of the underlying gene revealed this to be *RHM2* (rhamnose biosynthesis 2) which codes for an UDP-L-rhamnose synthase.[43] In addition to these genes, eight enhancer loci (called MEN: Mum-ENhancers) have been identified in the context of an already present *RHM2* inactivation, showing reduced mucilage production and release.[50] Finally, *mum3* and *mum5* mutants show mucilage with altered composition.[45] It is noteworthy to observe that among this collection of cloned loci, only a few have been associated with the early biosynthesis steps of seed coat mucilage, but much more has been discovered about the upstream signaling cascades (Fig. 1). Although these genes have been shown to be involved in mucilage synthesis and/or modification,[52] several players are likely to be still missing from the network summarized above (Fig. 1). The existence of genes known to be involved in this pathway and transcriptionally measurable by the *Arabidopsis thaliana* Affymetrix microarray[53] makes this scenario ideal for expression-based gene network reverse engineering. Therefore, we decided to apply the LASSO method on a subset of Affymetrix microarrays comprising seed and seedling samples through several developmental stages of healthy wild-type Columbia-0 Arabidopsis plants.[54]

Results and discussion

Prediction of candidate genes

In order to extend previously used correlation based approaches, we wanted to explore the use of the LASSO on a specific dataset to identify novel candidate genes involved in Arabidopsis mucilage biosynthesis. We decided to use a small dataset, focusing on samples where seed coat mucilage is synthesized or its synthesis pre-programmed, and where genes known to be involved in mucilage biosynthesis and regulation were known to be expressed. Thus we used the AtGENEXPRESS tissue dataset which measures the expression of more than 22 000 genes using ATH1 affymetrix slides under many developmental conditions and in different tissues,[54] each being replicated three times. As we were interested in seed coat mucilage, we extracted slides from pollen, flower, seed and silique development only. Within this small dataset, we focused on the GL2 sub-network involved in mucilage biosynthesis (Fig. 1). Thus we tried to predict *RHM2* expression as well as that of its upstream transcription factor *GL2* by all other genes measured by the Arabidopsis ATH1 chip (>20 k) using the LASSO.

The LASSO needs to be parameterized, meaning that a single "best" model must be chosen for each bait variable (in our case, a gene), amongst the many weight constrained models explored by the LASSO.[30] In order to perform this parameterization, we used cross-validation as calculated in the LARS implementation[28] to identify the model(s) yielding the lowest error rate. In the case of *RHM2*, the model comprised 16 genes, whereas the *GL2* model provided 14 genes (Tables 1 and 2).

A manual inspection of the genes identified in the LASSO regression showed promising results. In the case of *RHM2* (Table 1), the genes identified contained glucuronoxylan glucuronosyltransferase (*GUT2*) also known as irregular xylem 10 (*IRX10*) due to its mild irregular xylem phenotype. Due to the mutant phenotype it was suggested to be involved in glucuronoxylan biosynthesis together with IRX10-like.[55] Furthermore this list included *NRS/ER*, a gene which shows strong similarity to *RHM2* but lacks one of its domains and is thus likely to be involved in the synthesis of UDP-L-rhamnose as well.[56] In addition, the model contained *AtNST-KT1* (At4g39390), a gene which represents a monospecific nucleotide sugar transporter.[57] Specifically it was shown that this transporter exchanges UDP-Gal for UMP and is localized to Golgi membranes. Furthermore, the network included *ATMAN7* (At5g66460), a putative mannanase, which is expressed strongly during early Arabidopsis seed germination and whose knock-out mutants show a lower germination frequency.[58] Finally, At2g04690 had been found in a proteomics study for cell wall proteins identifying less than 200 proteins in total.[59] In order to check whether these genes would also have been identified using simple Pearson correlation, we analyzed both the Pearson correlation coefficient and their rank. The gene *NRS/ER*, characterized by a partial homology to the bait gene *RHM2*, showed an extremely high correlation of >0.93 and concomitantly was the gene with the second highest correlation to the bait (Table 1). Therefore this gene would also have been found as a gene potentially involved in mucilage biosynthesis, not only because of its sequence homology, but also because of its overwhelmingly similar expression behavior to RHM2. Moreover At2g04690 was ranked 4th by LASSO and 12th by PCC and would likely have been included as a candidate when using either technique. However this situation drastically changes when looking at the three genes likely involved in cell wall precursor biosynthesis and modification.

Table 1 List of genes included in the *RHM2* lowest prediction error LASSO model (assessed by 10-fold cross-validation). The expression behavior of *RHM2* is explained by a linear combination of these genes, weighted by the coefficient indicated in the third column. Pearson Correlation Coefficients (PCCs) and absolute ranks for each gene *vs. RHM2* are also indicated

Gene symbol	Protein function[79]	Weight in the *RHM2* LASSO model	PCC *vs.* *RHM2*	Absolute LASSO rank	Absolute PCC rank
At1g63000	Nucleotide-rhamnose synthase/epimerase-reductase (NRS/ER)	0.6536294	0.9300414	1	2
At1g61440	S-locus lectin protein kinase	0.5916742	0.615933	2	832
At4g38200	SEC7-like guanine nucleotide exchange family protein	0.1743362	0.658125	3	533
At2g04690	Pyridoxamine 5′-phosphate oxidase family protein	−0.12075	−0.8418919	4	12
At1g54110	Membrane fusion protein Use1	−0.128731	−0.7416396	5	141
At4g10030	alpha/beta-Hydrolases superfamily protein	0.08882269	0.7913813	6	43
At5g66460	Glycosyl hydrolase superfamily protein (ATMAN7)	0.09562279	0.719603	7	204
At4g39390	Golgi-localized nucleotide sugar transporter (AtNST-KT1)	0.07519697	0.7088117	8	249
At5g53540	P-loop containing nucleoside triphosphate hydrolases superfamily protein	−0.06086272	−0.6424797	9	639
At1g67360	Rubber elongation factor protein (REF)	0.046068	0.6733259	10	449
At1g27440	Glucuronoxylan glucuronosyltransferase (GUT2)	0.04102195	0.775059	11	67
At1g56300	Chaperone DnaJ-domain protein	0.03710941	0.6461923	12	613
At3g01210	RNA-binding (RRM/RBD/RNP motifs) family protein	−0.03595778	−0.7981043	13	35
At2g46660	Member of CYP78A	−0.009253494	−0.6350844	14	686
At5g38530	Type 2 tryptophan synthase	−0.001077749	−0.8126333	15	27
At1g73440	Calmodulin-related	−0.04172261	−0.5315293	16	1721

Table 2 List of genes included in the *GL2* lowest prediction error LASSO model (assessed by 10-fold cross-validation). The expression behavior of *GL2* is explained by a linear combination of these genes, weighted by the coefficient indicated in the third column. Pearson Correlation Coefficients (PCCs) and absolute ranks for each gene *vs. GL2* are also indicated

Gene symbol	Protein function[79]	Weight in the *GL2* LASSO model	PCC *vs.* *GL2*	Absolute LASSO rank	Absolute PCC rank
At1g76880	Homeodomain-containing putative transcription factor	0.3351552	0.9347944	1	43
At2g23260	UDP-glucosyl transferase 84B1	0.0451604	0.9764201	2	2
At4g09820	Regulator flavonoid pathways (TT8)	0.1777512	0.9518524	3	18
At1g77990	Sulfate transporter AST56	−0.1511854	−0.8541893	4	252
At5g15180	Peroxidase superfamily protein	0.07057812	0.9443255	5	26
At3g13540	MYB transcription factor, negative regulator of trichome branching (MYB5)	0.1685542	0.9480528	6	22
At1g20500	AMP-dependent synthetase and ligase	0.156592	0.9563064	7	13
At5g49270	Involved in successfully establishing tip growth in root hairs (MRH4)	−0.04394491	−0.776442	8	999
At1g12880	Nudix hydrolase homolog 12 (NUDT12)	−0.08522669	−0.844569	9	312
At1g56170	Nuclear factor *y*, subunit c2 (HAP5B)	0.0860337	0.8560389	10	240
At5g03190	Putative methyltransferase	0.1012894	0.7836278	11	909
At1g63300	Myosin heavy chain-related protein	0.05661009	0.9576575	12	10
At1g04040	HAD IIIB acid phosphatase	−0.006245497	−0.6764834	13	2722
At4g27860	Vacuolar iron transporter (VIT)	−0.01473578	−0.7953438	14	766

Whilst in each case the Pearson correlation coefficient is above 0.7 which is used as a common threshold,[13] these genes are in no case amongst the top 50 co-expressors and indeed the nucleotide sugar transporter is found only at rank 249. Thus, these genes would not have been picked up by mutual rank based correlation approaches such as AraNet[60] focusing on the top 30 correlators, nor would they have been short-listed as likely candidates from a simple correlation based approach.

We next turned our attention to the genes identified using *GL2* as a bait. As opposed to *RHM2*, this LASSO model did not contain any gene likely involved in nucleotide sugar and/or cell wall polymer biosynthesis or modification (Table 2). Despite the absence of biosynthetic genes, some genes in the *GL2* model are already known to be involved in seed coat development and differentiation, such as *MYB5* (At3g13540)[61] and *TT8* (At4g09820), both of which are thought to be part of a ternary complex with *TTG1*[42] regulating flavonoid synthesis in the seed coat and the silique.[62] Interestingly, a loss of function of these genes not only leads to changes in the seed coat but also results in reduced mucilage release. Once again, we compared the simple Pearson correlation results and found both *MYB5* and *TT8* to be also strongly co-regulated with *GL2* (Pearson correlation >0.93 in both cases). That said, despite their extremely high correlation, both genes come out on ranks 22 and 18 in the Pearson correlation based approach respectively. In the LASSO model these were on ranks 6 and 3 respectively. Furthermore LASSO identified At5g49270 which is a cobra like protein (COBL9) and whose mutant shows a short root hair phenotype.[63] This is interesting as COBRAs have been speculated to signal between the membranes and the cell walls.[64] Due to its high PCC rank (999) it is however unlikely that it would have been considered as a likely candidate when using Pearson correlation alone. That said, several candidates in the list likely represent false-positives, *i.e.* genes not involved in mucilage or seed coat development or regulation.

Thus in order to investigate the similarity in outcome between the LASSO and Pearson correlation, we plotted the LASSO weight and the PCC (Fig. 2). In both *RHM2* and *GL2* models,

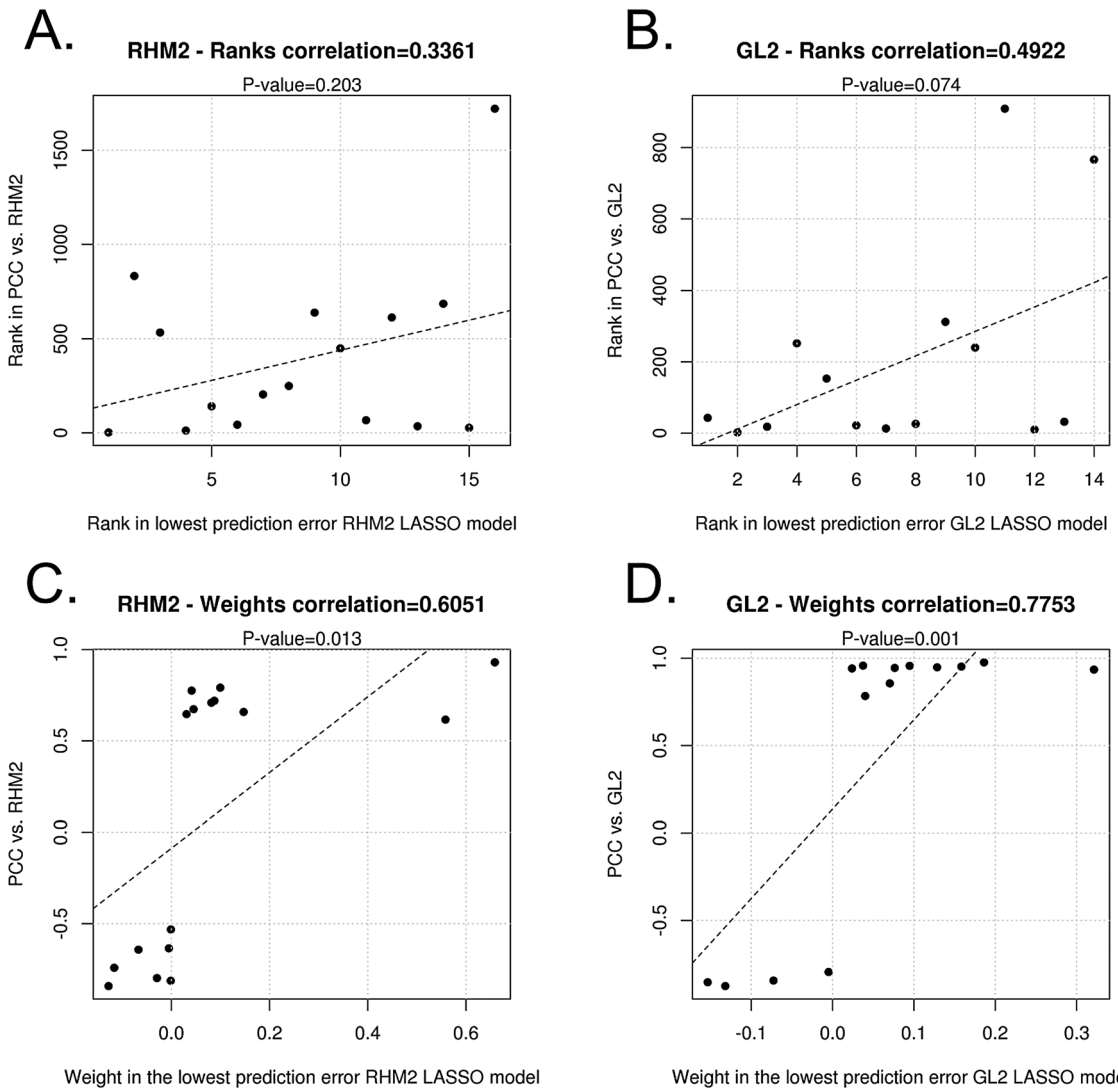

Fig. 2 Comparison between LASSO weights and Pearson correlation coefficients in the *GL2* and *RHM2* networks. The ranks of absolute LASSO weights and PCCs are positively correlated amongst the genes included in both the *RHM2* (A) and *GL2* (B) final LASSO models, although this correlation is not significant. In the lower panels ((C) for the *RHM2* model and (D) for the *GL2* model) we show the correlation between the LASSO weights and PCCs directly (without ranking). In all panels, the outcome of linear regression is drawn as a dashed line; correlation coefficients are indicated, together with their *P*-values.

it is possible to note a general agreement between Pearson correlation and LASSO (Fig. 2). The correlation between Pearson and LASSO absolute ranks for the genes found using *RHM2* (Fig. 2A) and *GL2* (Fig. 2B) is positive, although not significant. The similarity between the Pearson correlation coefficients and the LASSO weights is even stronger and significant for both *RHM2* (Fig. 2C, *P*-value 0.013) and *GL2* (Fig. 2D, *P*-value 0.001). Moreover, we can see how genes with opposite behavior to the two gene baits are given negative PCCs and negative LASSO weights, in the bottom left of the scatter plots, while positive Pearson correlators are deemed to be positive contributors also in the LASSO model (Fig. 2C and D). Altogether, it is possible to say that while they mostly differ in the order of candidate genes provided, both Pearson correlation and LASSO provide an overall comparable assessment of the nature of individual genes' contribution to the behavior of the baits *RHM2* and *GL2*.

Nevertheless as all guilt by association approaches might yield false positives, we checked the functional annotation of the genes included in both candidate lists for either enzymes that might play a role in mucilage synthesis or modification or transcription factors.

In the case of *GL2*, this yielded a list of transcription factors plus two genes encoding enzymes putatively involved in nucleotide sugar related processes: the UDP-glycosyl transferase At2g23260 and the AMP-dependent synthetase and ligase At1g20500 (Table 2). Among the transcription factors, we could identify At1g76880: a putative MYB-like transcription factor member of the trihelix DNA binding family. At1g76880 has a strong positive weight in the model (0.335, highest in the model, see Table 2). Commonly used co-expression approaches, such as Pearson correlation would have had problems in identifying it, since it ranks at position 43 (Table 2) and is not indicated as a top co-expressor of *GL2* in any of the public data mining tools

known to the authors. At1g76880 has been recently named *DF1* in a general genomic study describing all *Arabidopsis* trihelix proteins.[65] It is already known that *DF1* is able to bind DNA with consensus sequences GGTAATT or TACAGT in pea;[66] furthermore, this gene has been shown to be repressed by light in both pea and soybean.[66,67] However, no detail on *DF1* specific function has been provided so far and no obvious phenotype was found in a loss-of-function mutant,[68] therefore we decided to characterize this gene in more detail.

DF1 mucilage analysis

In order to establish whether *DF1* might play a role in mucilage synthesis, we sought T-DNA insertion lines for this gene. After querying the SALK database,[69] we could identify two independent insertion lines. Both insertions were mapped into the second exon by SALK and were relatively close to each other (Fig. 3). The lines were screened by PCR for the insertion and homozygosity. Homozygous loss of function plants were grown side by side with the recurrent WT Columbia-0 and seeds were harvested from at least seven independent plants per line. The mature seeds were stained with ruthenium red, a dye used to visualize the release of the Arabidopsis seed coat mucilage. Interestingly, while mucilage release was clearly visible in the case of the WT (Fig. 4A), it was not possible to detect any staining in either of the mutant alleles (Fig. 4B and C). This could be explained by (i) a failure to release mucilage upon contact with water due to changes in the outer cell wall, (ii) only low amounts of mucilage being released due to synthesis problems, or (iii) failure to stain released mucilage due to changes in its physicochemical properties. To distinguish these possibilities we conducted additional staining experiments.

Firstly seeds were shaken prior to staining, which removed the outer mucilage layer in the WT (Fig. 4D), but this treatment only revealed weak and irregular staining in the mutant seeds (Fig. 4E and F). It has been previously shown that the release of mucilage can be encouraged by physical damage to the seed coat if mucilage is produced but not released. In the mutants, the seed coat was therefore scraped while in the staining solution. Once again, hardly any mucilage was released in either mutant alleles (Fig. 4G and H). Another method to induce mucilage release is the treatment with chelators such as EDTA (Fig. 4I).[50] Interestingly, after the treatment

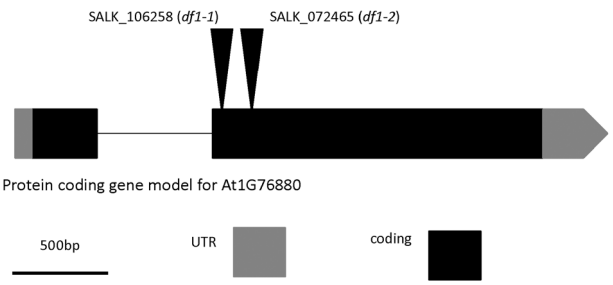

Fig. 3 Structure of the At1g76880 gene and the positions of the two T-DNA insertions. The position of the T-DNA insertions for *df1-1* and *df1-2* for At1g76880 mutants is indicated by triangles. Gray boxes represent the untranslated regions, whereas black boxes show the coding region.

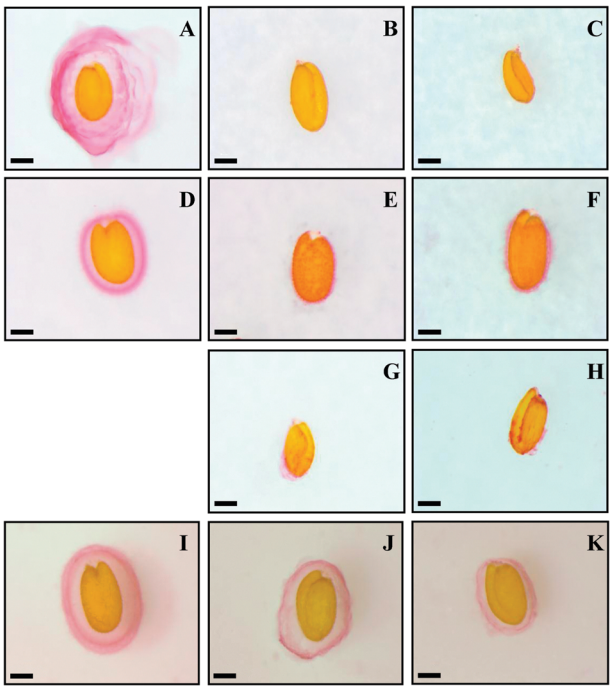

Fig. 4 Cytological analysis of the two knock-out lines *df1-1* and *df1-2*. *Arabidopsis thaliana* seeds were stained using ruthenium red. Staining was performed on the seeds directly for the WT (A) as well as for the *df1-1* (B) and the *df1-2* (C) mutant, as well as after shaking WT (D), *df1-1* (E) and *df1-2* (F) and after shaking with EDTA WT (I), *df1-1* (J) and *df1-2* (K). The two mutant lines were also stained after scraping the seed coat *df1-1* (G) and *df1-2* (H). Scale bars = 10 μm.

with EDTA both knock-out lines were showing release of the mucilage (Fig. 4J and K).[52] This result suggests that mucilage release might be impaired in mutant seeds. Nevertheless these experiments alone could not exclude that some physicochemically modified mucilage is released when the seeds are immersed in water. We therefore quantified the differences in mucilage release by extracting soluble mucilage from the mutants and the WT by shaking them in water. The supernatant was subjected to monosaccharide composition analysis. Both lines showed a drastic reduction in galacturonic acid and rhamnose by more than 80% in each case. As stated earlier, the outer mucilage is largely composed of a relatively unbranched RGI, thus consisting mostly of rhamnosyl and galacturonosyl residues (Table 3). Therefore the drastic reduction in these sugars

Table 3 Monosaccharide levels (nmol mg^{-1} seed weight) in water extractable mucilage of *Arabidopsis thaliana* wild type seeds and the two independent insertion lines for At1g76880 (*df1-1* and *df1-2*). Values are represented as mean ± standard deviation for at least 7 biological replicates. Significant values $P < 0.001$ are marked by an asterisk

	WT Col-0	*df1-1*	*df1-2*
Fucose	0.13 ± 0.05	0.07 ± 0.03*	0.05 ± 0.02*
Rhamnose	27.88 ± 9.85	5.13 ± 3.33*	3.67 ± 4.30*
Arabinose	1.50 ± 0.50	0.71 ± 0.24*	0.52 ± 0.21*
Galactose	1.36 ± 0.56	0.73 ± 0.21*	0.61 ± 0.21*
Glucose	2.18 ± 1.40	1.98 ± 0.67	2.57 ± 1.20
Xylose	2.73 ± 0.96	0.64 ± 0.37*	0.43 ± 0.34*
Mannose	0.64 ± 0.22	0.26 ± 0.10*	0.19 ± 0.08*
Galacturonic acid	26.24 ± 9.05	4.2 ± 2.58*	3.64 ± 3.39*

confirms the staining results and indicates that the observed phenotype is likely induced by a soluble mucilage reduction as opposed to reduced dye accessibility. Interestingly, this phenotype extended also to the "minor" seed coat sugars. Xylose was reduced by more than 70% and fucose, arabinose, galactose as well as mannose were reduced by more than 50% in both lines (Table 3). However glucose was unchanged in both lines when compared to the WT.

DF1 seed coat analysis

In order to further explore the reason for these changes, we investigated the seeds using environmental scanning electron microscopy (ESEM). Here we could show that mature dry mutant seeds showed a significantly disturbed seed coat (Fig. 5B and C) when compared to the WT (Fig. 5A). Both alleles showed very irregular structures and an apparently changed columella shape. This was visible in the case of a whole seed coat scan, but became more apparent when zooming into a detailed section of the seed coat (Fig. 5E and F).

We then explored how hydration of the seeds would influence the seed coat epidermis morphology. For this the seeds were observed by ESEM after hydration with water. Interestingly, we again saw consistent changes between the mutant lines and the WT. Though the epidermis seemed more regular in the mutant seeds than before hydration, both mutant lines showed a grossly different columella structure than the WT seeds, exhibiting a ring like structure instead of a simple flat columella (Fig. 5G, H and I).

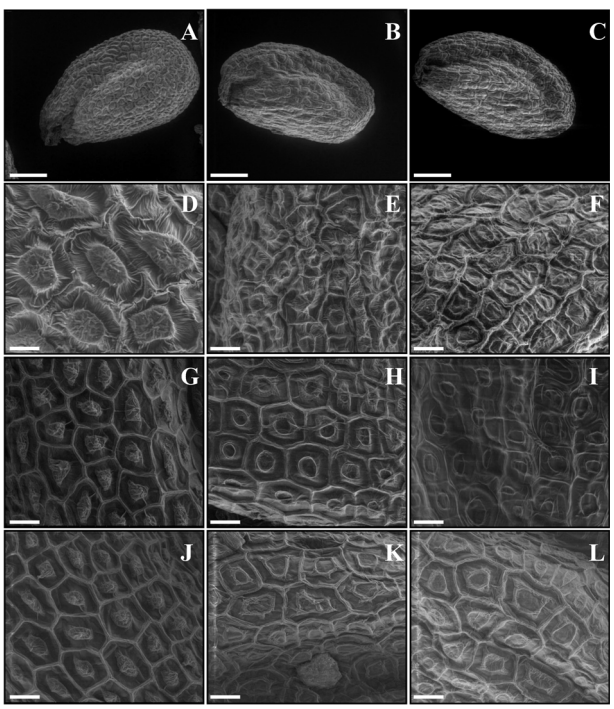

Fig. 5 Environmental Scanning Electron Microscopy (ESEM) of seed coat epidermis of dry and hydrated seed surfaces. Seeds were visualized directly, WT (A, D), *df1-1* (B, E) and *df1-2* (C, F) or after being hydrated WT (G), *df1-1* (H) and *df1-2* (I). In addition seeds were visualized after treatment with EDTA WT (J), *df1-1* (K) and *df1-2* (L). Scale bars (A, B, C) = 100 μm; scale bars (D, E, F, G, H, I, J, K, L) = 20 μm.

As stated above it is most likely that the change in mucilage release seen in our mutants is due to changes that prevent mucilage hydration and release. This is in line with what has been observed for several mutants such as *mum3*, *mum5* and *sbt1.7*,[45,51] where treatment with metal chelators promoted the mucilage release as in our mutants (Fig. 5J, K and L).

Experimental

Data processing

The transcriptome dataset used to build the LASSO models was the Affymetrix AtGenExpress[54] (GEO accessions: GSE5634 and GSE5632) seed, silique, flower and pollen developmental series, comprising 90 Affymetrix entries, and normalized *via* tRMA[70] with the CustomCDF v14.1.0 probeset annotation.[71] The entries were grouped by biological replicates and averaged after normalization, in order to give identical weights to each developmental condition, thus ending up with 30 samples. The relatively small number of samples allowed for a full LASSO modeling over all the 21492 (CustomCDF) genes without further gene filtering. The models were built using the following list of genes as dependent variables, measured according to the probes annotated in the CustomCDF project for *RHM2* and *GL2*. The path of the LASSO calculation was obtained *via* the Least Angle Regression for LASSO (LARS) algorithm,[28] in the implementation available from the R package *lars*. The best model obtained by LARS for each dependent variable was selected as the one yielding the smallest mean squared prediction error during a 10-fold cross validation analysis. Correlations between absolute ranks (Fig. 2A and B) and between LASSO weights and PCC (Fig. 2C and D) have been calculated *via* PCC; the *P*-value of these coefficients has been obtained *via* Fisher's Z transform based on the assumption that Pearson's correlation coefficients follow a t distribution with length(x) − 2 degrees of freedom.[72]

Plant material

Arabidopsis plants were grown in soil under standard conditions (120 μmol m^{-2} s^{-1}, 60% humidity, 20 °C, 16 h light/8 h dark) side by side with the respective Columbia-0 WT plants. Seeds were harvested and kept at 10 °C for at least two weeks prior to their analysis.

Identification of T-DNA lines

Two independent T-DNA insertion lines were obtained for At1g76880 (DF1) from the Nottingham Arabidopsis Stock Centre SALK collection.[69] Both lines (*df1-1*, SALK_106258 and *df1-2*, SALK_072465) showed a TDNA insertion in an exon and were confirmed by PCR using the primers (5′-ATTTTGCCGATTTCGGAAC-3′) and (5′-AACCAATCTCTCGTGTTCTCGC-3′) to confirm the insertion and (5′-GCGGAGCATGGTTACATAAG-3′) and (5′-AACCAATCTCTCGTGTTCTCGC-3′) to test for the presence of the WT allele.

Mucilage monosaccharide composition analysis

From a minimum of 7 different individual plants for each line, mucilage was extracted from 6 to 8 mg two week old seeds by shaking them in 1 mL bi-distilled water for 2 h at 37 °C.

After the addition of ribose as an internal standard, the released mucilage material was dried under a stream of air and hydrolyzed to monosaccharides by incubation with 2 M trifluoroacetic acid for 1 hour at 121 °C.

Monosaccharide composition analysis of hydrolyzed material was performed[73] using a High-Performance Anion Exchange Chromatography with Pulse Amperometric Detection[74] ICS 3000 (Dionex, California) equipped with a CarboPac PA20 column. Under a constant flow of 0.45 mL min^{-1} a NaOH gradient was applied as follows: start at 4 mM NaOH, decrease to 2 mM in 2 minutes, isocratic at 2 mM NaOH for 19 minutes, rise to 616 mM NaOH in 2 minutes, isocratic at 616 mM NaOH for 16 minutes, decrease to 4 mM NaOH in 3 minutes and re-equilibration of the column for 11 minutes at 4 mM NaOH. A constant post-column addition of 0.15 mL min^{-1} 100 mM NaOH was used to increase sensitivity.

Statistical analysis on the measured sugar levels was performed using a student's *t* test as implemented in the statistical environment R/Bioconductor.[75]

Seed staining

Arabidopsis thaliana ecotype Columbia-0 seeds and seeds from two knock-out lines (*df1-1* and *df1-2*) were stained with a 0.01% (w/v) aqueous solution of ruthenium red for 5–10 minutes both after 2 h shaking with water at 37 °C and without shaking. For the experiment of mucilage release by EDTA the seeds were shaken in 50 mM EDTA for 2 h at 37 °C, rinsed with water and stained with 0.01% (w/v) aqueous solution of ruthenium red. Seeds were visualized using a Leica MZ 12.5 Stereomicroscope (software: Leica Application Suite).

Environmental scanning electron microscopy

Environmental scanning electron micrographs were obtained with an FEI FE-ESEM Quanta 600 scanning electron microscope. Images were acquired at 5 kV of accelerating voltage and at 0.75 Pa of water vapor pressure. Mature dry seeds, hydrated seeds (aqueous shaking for 2 h at 37 °C) and EDTA treated seeds (shaking for 2 h at 37 °C with 50 mM EDTA) were analyzed.

Conclusions

The finding of "missing links" in gene network reverse engineering has always been a challenge both for experimental biology and bioinformatics since the advent of transcriptomics.[76] The capability to create a short list of novel gene candidates can give a considerable advantage in our understanding of biological systems, with tremendous benefits in *e.g.* cancer[24] and crop[77] research. In the present study, we propose a novel expression-based approach built on LASSO modeling, with the aim of discovering novel genes involved in mucilage biosynthesis in *Arabidopsis thaliana*. Our method follows the well-established "guilt-by-association" principle:[3] specifically, it scouts a developmental series transcriptomic dataset, using as a bait the expressional behavior of two genes known to be already involved in the process: the transcription factor *GL2* and the UDP-L-rhamnose synthase *RHM2*. Despite LASSO being developed for linear modeling of genes, we show that it provides realistic candidates (see Tables 1 and 2), some of which were already known to be involved in pectin metabolism. Furthermore, we show that it is a rather complimentary technique when compared with common Pearson correlation, the most widely-used method in expression-based gene network reconstruction studies.[13] The LASSO, unlike simple regression techniques, can operate in datasets with far less samples than genes (in our case, 30 developmental samples and 21 492 measured transcripts), which is a common case for microarray data, and will be an even more critical issue for RNASeq data,[78] therefore LASSO might see a further adaptation in the future. In our study, we experimentally proved a loss of function mutation for a putative transcription factor yielded by the *GL2* model to be largely devoid of mucilage as would be expected for a transcription factor involved in seed (mucilage) development. Furthermore, similar to other transcription factors involved in seed development, the mutants show an altered seed coat surface. Thus showing that the LASSO can not only be used to find already known candidate genes, our approach thus offers itself as a complementary method for guilt-by-association gene finding for transcriptomic studies.

Acknowledgements

Hereby, we thank all colleagues from AG Usadel for help with the harvesting of the plant material, Thomas Herter for sequence analysis, Lupo Giorgi for his support, and Marie Bolger and Anthony Bolger for critical reading of the manuscript.

References

1 J. M. Stuart, E. Segal, D. Koller and S. K. Kim, *Science*, 2003, **302**, 249.
2 H. Yu, N. M. Luscombe, J. Qian and M. Gerstein, *Trends Genet.*, 2003, **19**, 422–427.
3 C. J. Wolfe, I. S. Kohane and A. J. Butte, *BMC Bioinf.*, 2005, **6**, 227.
4 P. R. Lee, J. E. Cohen, E. A. Tendi, R. Farrer, G. H. De Vries, K. G. Becker and R. D. Fields, *Neuron Glia Biol.*, 2004, **1**, 135–147.
5 H. Wei, S. Persson, T. Mehta, V. Srinivasasainagendra, L. Chen, G. P. Page, C. Somerville and A. Loraine, *Plant Physiol.*, 2006, **142**, 762.
6 K. Yonekura-Sakakibara, T. Tohge, F. Matsuda, R. Nakabayashi, H. Takayama, R. Niida, A. Watanabe-Takahashi, E. Inoue and K. Saito, *Plant Cell*, 2008, **20**, 2160.
7 B. Usadel, F. Poree, A. Nagel, M. Lohse, A. Czedik-Eysenberg and M. Stitt, *Plant, Cell Environ.*, 2009, **32**, 1211–1229.
8 J. Hu, J. Wan, L. Hackler Jr., D. J. Zack and J. Qian, *Bioinformatics*, 2010, **26**, 2289–2297.
9 D. Rambaldi, F. M. Giorgi, F. Capuani, A. Ciliberto and F. D. Ciccarelli, *Trends Genet.*, 2008, **24**, 427–430.
10 M. Quimbaya, K. Vandepoele, E. Raspé, M. Matthijs, S. Dhondt, G. T. S. Beemster, G. Berx and L. De Veylder, *Cell. Mol. Life Sci.*, 2012, 1–15.
11 M. Mutwil, S. Klie, T. Tohge, F. M. Giorgi, O. Wilkins, M. M. Campbell, A. R. Fernie, B. Usadel, Z. Nikoloski and S. Persson, *Plant Cell*, 2011, **23**, 895.
12 T. Obayashi, K. Kinoshita, K. Nakai, M. Shibaoka, S. Hayashi, M. Saeki, D. Shibata, K. Saito and H. Ohta, *Nucleic Acids Res.*, 2007, **35**, D863–D869.
13 B. Usadel, T. Obayashi, M. Mutwil, F. M. Giorgi, G. W. Bassel, M. Tanimoto, A. Chow, D. Steinhauser, S. Persson and N. J. Provart, *Plant, Cell Environ.*, 2009, **32**, 1633–1651.
14 S. Persson, H. Wei, J. Milne, G. P. Page and C. R. Somerville, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 8633–8638.
15 D. M. Brown, L. A. H. Zeef, J. Ellis, R. Goodacre and S. R. Turner, *Plant Cell*, 2005, **17**, 2281–2295.

16 L. Li, C. M. Foster, Q. Gan, D. Nettleton, M. G. James, A. M. Myers and E. S. Wurtele, *Plant J.*, 2009, **58**, 485–498.
17 F. F. Fu and H. W. Xue, *Plant Physiol.*, 2010, **154**, 927.
18 G. W. Bassel, E. Glaab, J. Marquez, M. J. Holdsworth and J. Bacardit, *Plant Cell*, 2011, **23**, 3101–3116.
19 J. M. Jiménez-Gómez, A. D. Wallace and J. N. Maloof, *PLoS Genet.*, 2010, **6**, e1001100.
20 H. Less, R. Angelovici, V. Tzin and G. Galili, *Plant Cell*, 2011, **23**, 1264.
21 K. Vandepoele, M. Quimbaya, T. Casneuf, L. De Veylder and Y. Van de Peer, *Plant Physiol.*, 2009, **150**, 535.
22 M. Mutwil, J. Øbro, W. G. T. Willats and S. Persson, *Nucleic Acids Res.*, 2008, **36**, W320.
23 A. de la Fuente, N. Bing, I. Hoeschele and P. Mendes, *Bioinformatics*, 2004, **20**, 3565–3574.
24 K. Basso, A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera and A. Califano, *Nat. Genet.*, 2005, **37**, 382–390.
25 Y. Lu, Y. Zhou, W. Qu, M. Deng and C. Zhang, *Bioinformatics*, 2011, **27**, 2406–2413.
26 M. Zampieri, N. Soranzo and C. Altafini, *Bioinformatics*, 2008, **24**, 1510–1515.
27 R. Tibshirani, *J. R. Stat. Soc. B*, 1996, **58**, 267–288.
28 B. Efron, T. Hastie, I. Johnstone and R. Tibshirani, *Ann. Stat.*, 2004, **32**, 407–451.
29 Z. Wang, M. Gerstein and M. Snyder, *Nat. Rev. Genet.*, 2009, **10**, 57–63.
30 J. Friedman, T. Hastie and R. Tibshirani, *The elements of statistical learning*, Springer, 3rd edn, 2008.
31 T. Hastie, R. Tibshirani and J. Friedman, *The elements of statistical learning*, 2nd edn, 2001.
32 T. Shimamura, S. Imoto, R. Yamaguchi and S. Miyano, *Genome Inf.*, 2007, **19**, 142–153.
33 M. Gustafsson, M. Hörnquist, J. Lundström, J. Björkegren and J. Tegnér, *Ann. N. Y. Acad. Sci.*, 2009, **1158**, 265–275.
34 F. Licausi, F. M. Giorgi, E. Schmälzlin, B. Usadel, P. Perata, J. T. van Dongen and P. Geigenberger, *Plant Cell Physiol.*, 2011, **52**, 1957–1972.
35 T. L. Western, *Botany*, 2006, **84**, 622–630.
36 A. Macquet, M. C. Ralet, J. Kronenberger, A. Marion-Poll and H. M. North, *Plant Cell Physiol.*, 2007, **48**, 984.
37 T. L. Western, D. J. Skinner and G. W. Haughn, *Plant Physiol.*, 2000, **122**, 345–356.
38 L. V. Caeseele, J. T. Mills, M. Sumner and R. Gillespie, *Can. J. Bot.*, 1981, **59**, 292–300.
39 F. D. Boesewinkel and F. Bouman, *Seed Dev. Germination*, 1995, **95**, 1–24.
40 W. G. T. Willats, L. McCartney, W. Mackie and J. P. Knox, *Plant Mol. Biol.*, 2001, **47**, 9–27.
41 G. Haughn and A. Chaudhury, *Trends Plant Sci.*, 2005, **10**, 472–477.
42 A. Baudry, M. A. Heim, B. Dubreucq, M. Caboche, B. Weisshaar and L. Lepiniec, *Plant J.*, 2004, **39**, 366–380.
43 T. L. Western, D. S. Young, G. H. Dean, W. L. Tan, A. L. Samuels and G. W. Haughn, *Plant Physiol.*, 2004, **134**, 296–306.
44 A. Gonzalez, J. Mendenhall, Y. Huo and A. Lloyd, *Dev. Biol.*, 2009, **325**, 412–421.
45 T. L. Western, J. Burn, W. L. Tan, D. J. Skinner, L. Martin-McCaffrey, B. A. Moffatt and G. W. Haughn, *Plant Physiol.*, 2001, **127**, 998.
46 J. Huang, D. Bowles, E. Esfandiari, G. Dean, N. C. Carpita and G. W. Haughn, *Plant Physiol.*, 2011, **156**, 491–502.
47 M. Walker, M. Tehseen, M. S. Doblin, F. A. Pettolino, S. M. Wilson, A. Bacic and J. F. Golz, *Plant Physiol.*, 2011, **156**, 46–60.
48 G. H. Dean, H. Zheng, J. Tewari, J. Huang, D. S. Young, Y. T. Hwang, T. L. Western, N. C. Carpita, M. C. McCann and S. D. Mansfield, *Plant Cell*, 2007, **19**, 4007.
49 B. Usadel, A. M. Kuschinsky, M. G. Rosso, N. Eckermann and M. Pauly, *Plant Physiol.*, 2004, **134**, 286.
50 A. A. Arsovski, M. M. Villota, O. Rowland, R. Subramaniam and T. L. Western, *J. Exp. Bot.*, 2009, **60**, 2601–2612.
51 C. Rautengarten, B. Usadel, L. Neumetzler, J. Hartmann, D. Büssis and T. Altmann, *Plant J.*, 2008, **54**, 466–480.
52 A. A. Arsovski, G. W. Haughn and T. L. Western, *Plant Signaling Behav.*, 2010, **5**, 796–801.
53 Affymetrix, http://www.affymetrix.com/.
54 M. Schmid, T. S. Davison, S. R. Henz, U. J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel and J. U. Lohmann, *Nat. Genet.*, 2005, **37**, 501–506.
55 D. M. Brown, Z. Zhang, E. Stephens, P. Dupree and S. R. Turner, *Plant J.*, 2009, **57**, 732–746.
56 G. Watt, C. Leoff, A. D. Harper and M. Bar-Peled, *Plant Physiol.*, 2004, **134**, 1337–1346.
57 I. Rollwitz, M. Santaella, D. Hille, U. I. Flügge and K. Fischer, *FEBS Lett.*, 2006, **580**, 4246–4251.
58 R. Iglesias-Fernández, M. C. Rodríguez-Gacio, C. Barrero-Sicilia, P. Carbonero and A. Matilla, *Planta*, 2011, 1–12.
59 M. Irshad, H. Canut, G. Borderies, R. Pont-Lezica and E. Jamet, *BMC Plant Biol.*, 2008, **8**, 94.
60 M. Mutwil, B. Usadel, M. Schutte, A. Loraine, O. Ebenhoh and S. Persson, *Plant Physiol.*, 2010, **152**, 29–43.
61 S. F. Li, O. N. Milliken, H. Pham, R. Seyit, R. Napoli, J. Preston, A. M. Koltunow and R. W. Parish, *Plant Cell*, 2009, **21**, 72–89.
62 N. Nesi, I. Debeaujon, C. Jond, G. Pelletier, M. Caboche and L. Lepiniec, *Plant Cell*, 2000, **12**, 1863–1878.
63 M. A. Jones, M. J. Raymond and N. Smirnoff, *Plant J.*, 2006, **45**, 83–100.
64 G. Schindelman, A. Morikami, J. Jung, T. I. Baskin, N. C. Carpita, P. Derbyshire, M. C. McCann and P. N. Benfey, *Genes Dev.*, 2001, **15**, 1115–1127.
65 R. N. Kaplan-Levy, P. B. Brewer, T. Quon and D. R. Smyth, *Trends Plant Sci.*, 2012.
66 Y. Nagano, *Plant Physiol.*, 2000, **124**, 491.
67 K. O'Grady, V. H. Goekjian, C. J. Nairn, R. T. Nagao and J. L. Key, *Plant Mol. Biol.*, 2001, **47**, 367–378.
68 C. Breuer, A. Kawamura, T. Ichikawa, R. Tominaga-Wada, T. Wada, Y. Kondou, S. Muto, M. Matsui and K. Sugimoto, *Plant Cell*, 2009, **21**, 2307–2322.
69 J. M. Alonso, A. N. Stepanova, T. J. Leisse, C. J. Kim, H. Chen, P. Shinn, D. K. Stevenson, J. Zimmerman, P. Barajas and R. Cheuk, *Science*, 2003, **301**, 653.
70 F. M. Giorgi, A. M. Bolger, M. Lohse and B. Usadel, *BMC Bioinf.*, 2010, **11**, 553.
71 M. Dai, P. Wang, A. D. Boyd, G. Kostov, B. Athey, E. G. Jones, W. E. Bunney, R. M. Myers, T. P. Speed, H. Akil, S. J. Watson and F. Meng, *Nucleic Acids Res.*, 2005, **33**, e175.
72 R. A. Fisher, *Biometrika*, 1915, **10**, 507–521.
73 J. Øbro, J. Harholt, H. V. Scheller and C. Orfila, *Phytochemistry*, 2004, **65**, 1429–1438.
74 M. R. Hardy, R. R. Townsend and Y. C. Lee, *Anal. Biochem.*, 1988, **170**, 54–62.
75 R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang and J. Zhang, *Genome Biol.*, 2004, **5**, R80.
76 M. Yeung, J. Tegnér and J. J. Collins, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 6163.
77 A. Fukushima, S. Kanaya and M. Arita, *Plant Biotechnol.*, 2009, **26**, 485–493.
78 N. Street, Large-scale RNA-seq Transcriptomics Studies Exploring Wood Development and Natural Variation in Aspen (*P. tremula*): Projects and Resource Development, 2012.
79 P. Lamesch, K. Dreher, D. Swarbreck, R. Sasidharan, L. Reiser and E. Huala, *Databases*, 2010, **1**, 1.

