

RESEARCH ARTICLE

Open Access

TobEA: an atlas of tobacco gene expression from seed to senescence

Kieron D Edwards^{1*}, Aureliano Bombarely², Geraint W Story^{1,3}, Fraser Allen¹, Lukas A Mueller², Steve A Coates¹, Louise Jones¹

Abstract

Background: Transcriptomics has resulted in the development of large data sets and tools for the progression of functional genomics and systems biology in many model organisms. Currently there is no commercially available microarray to allow such expression studies in *Nicotiana tabacum* (tobacco).

Results: A custom designed Affymetrix tobacco expression microarray was generated from a set of over 40k unigenes and used to measure gene expression in 19 different tobacco samples to produce the Tobacco Expression Atlas (TobEA). TobEA provides a snap shot of the transcriptional activity for thousands of tobacco genes in different tissues throughout the lifecycle of the plant and enables the identification of the biological processes occurring in these different tissues. 772 of 2513 transcription factors previously identified in tobacco were mapped to the array, with 87% of them being expressed in at least one tissue in the atlas. Putative transcriptional networks were identified based on the co-expression of these transcription factors. Several interactions in a floral identity transcription factor network were consistent with previous results from other plant species. To broaden access and maximise the benefit of TobEA a set of tools were developed to provide researchers with expression information on their genes of interest via the Solanaceae Genomics Network (SGN) web site. The array has also been made available for public use via the Nottingham Arabidopsis Stock Centre microarray service.

Conclusions: The generation of a tobacco expression microarray is an important development for research in this model plant. The data provided by TobEA represents a valuable resource for plant functional genomics and systems biology research and can be used to identify gene targets for both fundamental and applied scientific applications in tobacco.

Background

Nicotiana tabacum (tobacco) is a member of the *Solanaceae* family, which also includes tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*), amongst other commercially important crop plants. It was originally studied due to its economic importance but came to the fore of plant biology when it was the first plant to be genetically modified in 1983 [1].

Tobacco is an allotetraploid ($2n = 4 \times = 48$) derived from the interspecific hybridisation of two progenitor genomes; a maternal 'S' genome, originating from an ancestor of *N. sylvestris*, and a paternal 'T' genome, most likely originating from a lineage of *N. tomentosiformis* [2-4]. It has a genome of approximately 5 Gb [5],

but despite its commercial and scientific importance relatively little is known about the plant's genome sequence. Recent efforts by the Tobacco Genome Initiative (TGI; http://www.pnng.org/cbnp/index.php?option=com_content&task=view&id=16&Itemid=40) have increased the sequence information available for the transcriptionally active regions of the tobacco genome. Some of this information is in the form of Expressed Sequence Tags (ESTs); short, single pass sequence reads derived from complementary DNA (cDNA) libraries, whilst others are methyl filtered Genome Space Sequence Reads (GSRs).

ESTs provide an insight into transcriptionally active genes in a biological sample under a given set of conditions. EST sequencing is relatively expensive and time consuming. Microarrays, however, provide a faster less costly alternative for measuring gene expression for

* Correspondence: kieron.edwards@atcbiotech.com

¹Advanced Technologies (Cambridge) Ltd, 210 Cambridge Science Park, Milton Road, Cambridge, CB4 0WA, UK

thousands of genes simultaneously that can be more easily and reproducibly applied across a broad range of conditions or treatments to identify genes showing specific expression patterns or responses. It should be noted that microarray measurements provide an estimate of the steady state level of transcripts. They do not capture all of the post-transcriptional or post-translational regulation of genes and as such do not necessarily reflect the expression of a gene in terms of functional protein produced.

Microarrays have been used in other plant species and organisms such as insects and mammals to capture the variation in gene expression in different tissues throughout development [6-10]. These studies show that organs can be differentiated based on their transcriptional signatures, with sets of genes involved in specific biological processes being expressed only in the relevant tissues. Such data sets provide researchers with information on their gene of interest's expression over a broad range of conditions.

Inference of gene function based on co-expression with other annotated genes is a key component of functional genomics. Community-wide projects, such as AtGenExpress (<http://www.weigelworld.org/resources/microarray/AtGenExpress/>; [11,12]), and the development of centralised microarray service providers and data repositories, such as the Nottingham Arabidopsis Stock Centre (NASC; <http://affymetrix.arabidopsis.info>), have resulted in a large amount of publicly available expression data for the model plant *Arabidopsis thaliana*. Development of this resource has enabled comparative analyses to be carried out leading to the identification of strongly co-expressed genes and aiding the elucidation of biochemical pathways for example [13,14]. Associated with the broad availability of such data has been the development of both stand-alone and web-based tools to allow researchers to fully interrogate the expression data sets to further their own research [8,15-19].

Transcription Factors play a key role in regulating gene expression and have been identified by comprehensive analyses in a number of plant species [20-22]. Methyl filtered genomic sequence reads have been used to identify 80-90% of the transcription factors in tobacco [23,24]. Functional genomics is based on a guilt-by-association approach, whereby co-expression of a gene of unknown function with other genes known to be involved in a common biochemical or signalling pathway may suggest that the unknown gene is also involved in this pathway. Co-expression of genes can also be used to infer putative transcriptional networks [25]. The identification of transcription factors that act as regulators of specific biochemical or signalling pathways potentially provide prime targets for manipulation to alter the agronomic or biochemical characteristics of the

plant. Such manipulation could prove very useful in projects related to secondary metabolite or biofuel production, or in the case of tobacco leaves, reducing the level of metabolites that give rise to toxicants in cigarette smoke.

This study describes the generation of a custom Affymetrix expression array for tobacco and its use in developing the Tobacco Expression Atlas (TobEA). TobEA is a map of gene expression from multiple tissues sampled throughout the life cycle of the tobacco plant and is intended to be used as a reference data set for plant researchers. The expression data is freely available via the Solanaceae Genomics Network (SGN), a web based genomic resource for plants of the Solanaceae family ([26]; <http://solgenomics.net/>). Several tools have also been developed to facilitate access to information on researcher's genes of interest.

Results

Generation of a microarray for tobacco

In order to enable gene expression studies in tobacco, generation of a microarray based on *Nicotiana tabacum* sequences was undertaken. Currently GenBank contains over 250000 tobacco ESTs, however, prior to the initiation of this project less than 30000 sequences were publicly available for tobacco. To increase the quantity of available sequence data 16 cDNA libraries were generated from multiple tissue types of several tobacco varieties, although there was a focus towards libraries originating from variety K326 and leaf samples. Table 1 summarises the variety and tissue of origin, along with the number of ESTs generated from each of the libraries.

One of two methods was applied to minimise redundancy within each library and maximise the number of unique ESTs generated. Five of the libraries (TT1, TT2, TT3, TT4 and TT5) underwent normalisation prior to sequencing ([27]; see methods). This generated 12439 ESTs, of which 10294 (82.76%) could successfully be sorted into their library of origin based on library specific sequence tags. Despite lacking tags, the remaining 2145 sequences still contained valuable sequence information so were maintained as a generic *N. tabacum* leaf EST library (Table 1). The remaining 11 libraries (KP1, KR2, KR3, KL4, KL5, KN6, KT7, KF8, KG9, BL12 and TL13) underwent virtual subtraction to reduce redundancy [28]. The virtual subtraction libraries were generated as part of the European Sequencing of Tobacco (ESTobacco) project and were included in the tobacco unigene set generated by this effort <http://www.ESTobacco.info>.

The 58969 ESTs generated by this study, along with another 27219 *N. tabacum* sequences available at the time were assembled into a set of 40642 tobacco

Table 1 Tobacco EST libraries

| Tissue | Variety | Library name | Number of ESTs |
|----------------------|-------------------------------|--------------|----------------|
| Germinated seeds | K326 | KG9 | 4361 |
| Seedlings | K326 | KP1 | 4692 |
| Roots before topping | K326 | KR2 | 4351 |
| Roots after topping | K326 | KR3 | 5288 |
| Leaf before topping | K326 | KL4 | 4540 |
| Leaf after topping | K326 | KL5 | 3755 |
| Leaf after topping | Burley 21 | BL12 | 1938 |
| Leaf after topping | TN86 | TL13 | 3205 |
| Midrib | K326 | KN6 | 5150 |
| Stem | K326 | KT7 | 4844 |
| Flowers | K326 | KF8 | 4422 |
| Trichomes | Samsun | TT1 | 3183 |
| Trichomes | Burley 21 | TT2 | 1950 |
| Trichomes | K326 | TT3 | 1161 |
| Early senescent leaf | K326 | TT4 | 1691 |
| Late senescent leaf | K326 | TT5 | 2309 |
| Tobacco leaf | K326, Samsun, Burley 21 | TT | 2145 |

Table showing the tissue type and tobacco variety sampled in generating 17 EST libraries. The number of ESTs in each library is shown.

unigenes. 57270 of the 86188 sequences assembled into 11724 contigs, with the remaining 28918 singleton sequences completing the unigene set, which ranged in length from 81 bp to 6570 bp (Additional file 1: ATC unigene lengths.ppt and Additional file 2: ATC unigene sequences.zip). Functional annotations for the unigenes were taken from the best BLASTX hit of *Arabidopsis thaliana* proteins (e -value $< 1 \times 10^{-10}$), and/or from the BLASTX hit of non-redundant proteins from Genbank using the program Blast2GO. In total 21060 (51.8%) of the unigenes were annotated by either method (Additional file 3: ATC unigene annotations.txt).

The Affymetrix custom design service was used to produce a tobacco expression array based on the tobacco unigene sequences described above. Of the 40642 sequences submitted, 596 were omitted from the array as suitable probes could not be identified. Standard hybridisation and PolyA controls were included on the array together with four species-specific maintenance genes. A small proportion of the sequences were represented by 2 probe sets, resulting in a total of 43768 probe sets present on the array.

Temporal and spatial changes in tobacco gene expression

Temporal and spatial regulation of gene expression plays a critical role in shaping plant growth and development and considerable variation in gene expression at these levels has been demonstrated in several plant species [6,7,9]. Understanding where and when expression occurs provides a valuable insight into gene function

and the functioning of the plant as a whole. Using the custom designed Affymetrix expression array, this variation was captured for tobacco genes by measuring gene expression in different tissue types at various stages throughout plant development (see legend in Figure 1 for a full list of the samples). For ease of representation each of the 19 conditions in the Tobacco Expression Atlas (TobEA) was given a Tissue Type Index (TTI) number, with samples from the same tissue/organ grouped together and then ordered temporally based on developmental stage (e.g. sample 1.1 is an imbibed seed and samples 2.1 and 2.2 are roots from a seedling and mature plant respectively).

Data were pre-processed with the MAS5 algorithm to allow the future addition of further tissue types to the TobEA database [29]. Based on the MAS5 flag values, genes were considered expressed if they were called as Present or Marginal in at least four out of the five biological replicates for at least one tissue type. On this basis genes corresponding to 33458 of the 43768 probe sets on the array were expressed. Detection of expression for 76% of the probesets on the array was considered reasonable as TobEA did not comprehensively cover all of the conditions and varietal differences represented in the EST libraries used to generate the sequences that the array was designed against. The detection level is also not much lower than that shown previously in a larger study in *Arabidopsis* [6]. Of the expressed genes, 30009 showed a significant difference in expression between tissues, based on a one way ANOVA ($p < 0.05$) following the removal of non-changing genes (genes showing expression between + and -1 Log of the median across all conditions). The number of differentially expressed genes identified suggested that expression varied significantly between tobacco tissues and throughout development for the majority of genes in the plant.

Tissues/organs in other plant species can be distinguished based on their transcriptional signature [6,7]. Principal Components Analysis (PCA) was carried out on the TobEA data to determine the relationship between tobacco tissues. Figure 1 shows the individual TobEA microarrays plotted within the first three Principal Components (PCs). For each tissue type, the five biological replicate samples clustered closely together, suggesting that they were a reliable representation of the transcriptional profile for their respective tissue (Figure 1). The 19 different tissue types could be broadly subdivided into 9 clusters based on their position within PC space (Figure 1). Samples from imbibed seed (Cluster ix) showed distinct separation from the main body of samples along PC3 (Figure 1). The root (Cluster vii) and cotyledon/leaf (Clusters i, ii and iv) samples were also separated along PC3, with the remaining tissue types

- 1.1 – Imbimed seed
- 2.1 – Young root
- 2.2 – Mature root
- 3.1 – Cotyledons
- 3.2 – Whole seedling shoot
- 3.3 – Young leaf
- 3.4 – Mature leaf
- 3.5 – Early senescent leaf
- 3.6 – Mid/early senescent leaf
- 3.7 – Mid/late senescent leaf
- 3.8 – Late senescent leaf
- 3.9 – Cauline leaf
- 4.1 – Lower stem
- 4.2 – Upper stem
- 5.1 – Vegetative apex
- 5.2 – Floral apex
- 6.1 – Closed flower bud
- 6.2 – Open flower bud
- 6.3 – Open flower

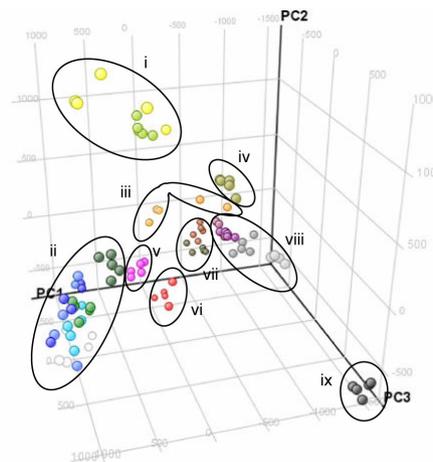


Figure 1 Tobacco organ expression trends. Principal Components Analysis (PCA) of the microarray data for the 95 samples in TobEA. PCA is a statistical method that reduces the dimensionality of multidimensional data-sets and enables the variance in gene expression for microarray samples to be represented across a set of vectors. Data points representing individual MAS5 pre-processed tobacco microarray samples are plotted within the first three PC vectors. Samples occupying similar position in PC space share similar gene expression trends. Black ovals in chart represent clusters of samples occupying similar position in PC space (see inset labels i to ix). Data points are coloured according to sample tissue of origin, with biological replicate samples coloured identically (see inset key for TTI identity).

falling between these groups (Figure 1). This suggested that PC3 was representative of some of the spatial variation in tobacco gene expression.

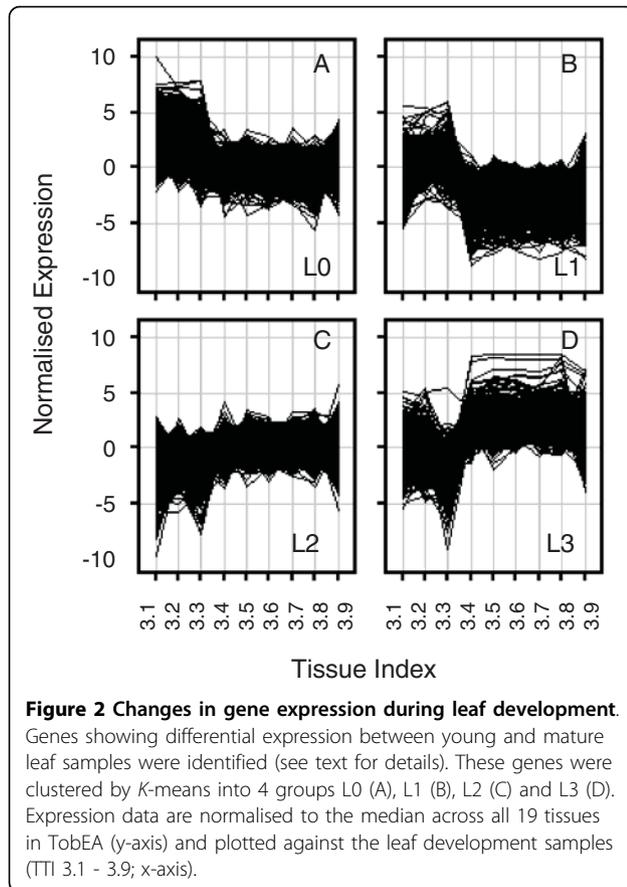
Some tissue types, such as roots (Cluster vii), could not be distinguished based on the developmental stage, however, other tissue types did show a developmental separation along PCs 1 and 2 (Figure 1). This was apparent for the transition from floral apex to floral buds in Cluster viii (Figure 1). Greater separation in PC space for a developmental series was shown by leaves. Cotyledon and young shoot samples (Cluster i) along with young leaf samples (Cluster iv) showed clear separation from mature and senescing leaf samples (Cluster ii) along PCs 1 and 2 (Figure 1). This showed a clear distinction in transcriptional signature between young and old leaves.

Gene expression changes during tobacco leaf development

To investigate the underlying cause of the separation between young and mature/senescing leaf material, genes were identified that showed differential expression

across the 'leaf' samples (TTI 3.1-3.9). Genes were called differentially expressed if they showed significant variation (One-way ANOVA $p < 0.05$) and a minimum of five-fold change in expression for at least one pairwise comparison between young (3.1 to 3.3) and mature (3.4 to 3.9) leaf samples. The 8944 genes identified on this basis were clustered by *K*-means into 4 groups as shown in Figure 2. Two basic responses were represented by the 4 clusters; genes showing relatively higher expression in younger samples (Clusters L0 and L1; Figures 2A and 2B) and genes showing relatively higher expression in older samples (Clusters L2 and L3; Figures 2C and 2D). Clusters were tested for over-representation of gene functions, based on the annotations borrowed from the best Arabidopsis BLAST hit.

When genes in clusters L0 and L1 were combined, the Gene Ontology (GO) analysis showed over-representation of multiple terms relating to photosynthesis and cell cycle, or cell growth and expansion (Additional file 4: GO analysis of young versus mature leaves.zip). A combined list of the genes in clusters L2 and L3 showed over-representation of GO terms relating to senescence



related responses, including break down and re-mobilisation of various metabolites, hormone responses and stress responses (Additional file 4: GO analysis of young versus mature leaves.zip). This suggested that there was a clear distinction between gene expression in young and old leaf material, with the former showing greater activity in photosynthesis, growth and expansion and the latter undergoing senescence.

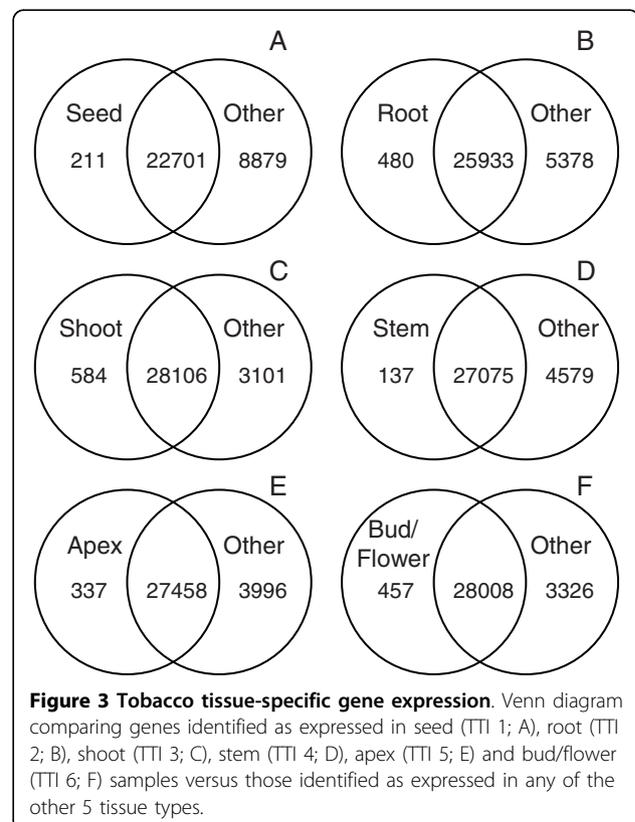
Although the four clusters were defined by two basic responses, normalisation of the data across all of the tissues in TobEA allowed further distinction of the clusters. For instance, genes in Clusters L0 and L1 both showed relatively higher expression in younger leaves than older leaves, however, in the context of other tissues in the expression atlas, Cluster L0 genes were relatively highly expressed in young leaf samples, whereas Cluster L1 genes showed relatively low expression in older leaf samples (Figure 2A and 2B). GO analysis of the genes in Cluster L0 principally showed terms associated with Photosynthesis, whereas genes in Cluster L1 showed terms related to cell cycle or cell growth and expansion (Additional file 4: GO analysis of young versus mature leaves.zip). Together, this suggested that the differences between young and old leaf tissues can be

described as increased photosynthetic activity in younger samples and reduced cell division and growth along with increased senescence in older leaf tissues. This clearly shows the benefit of considering changes in gene expression in the context of all of the other tissue types in the TobEA database and not just between a limited subset of tissue types.

Tissue specific tobacco gene expression

The above analyses demonstrated that many tobacco genes show changes in expression between different tissues. To test for genes that are only expressed in specific tissues, all of the expressed genes were identified for each tissue type. For this analysis only genes flagged as Present or Marginal in all 5 biological replicates of at least one developmental stage within a tissue type were considered expressed. Figure 3 shows a Venn diagram for each of the six tissue types represented in the TobEA database, with the genes identified as expressed in each tissue compared to a combined list of all those identified as expressed in the other tissues.

GO analysis was carried out on genes identified as expressed specifically in a single tissue type (Additional file 5: GO analysis of tissue specific genes.zip). The 211 genes identified as specifically expressed in imbibed seed samples (TTI 1.1) showed over-representation of



terms associated with seed development and nutrient storage as well as those responsive to water and high light intensity, all processes/responses that are likely to be well represented in an imbibed seed (Additional file 5: GO analysis of tissue specific genes.zip). Similarly, genes specifically expressed in root samples (TTI 2.1-2.2) showed over-representation of terms related to nutrient uptake, transport and toxin or defence responses. Those specifically expressed in leaf samples (TTI 3.1-3.9) showed over-representation of genes involved in photosynthesis (Additional file 5: GO analysis of tissue specific genes.zip). Additionally, genes specifically expressed in stem (TTI 4.1-4.2), shoot apices (TTI 5.1-5.2) and flower buds/flowers (TTI 6.1-6.3) showed over-representation of genes associated with microtubule development, implicated in the regulation of stem growth [30], Chromosome segregation/DNA metabolism and floral organ development

respectively (Additional file 5: GO analysis of tissue specific genes.zip).

Co-expression of tobacco genes

The set of 30009 genes showing differential expression across the TobEA samples were analysed for co-expression. Figure 4 shows a subset of 9 clusters from a set of 30 *K*-means clusters (See Additional file 6: Tobacco gene co-expression.ppt and Additional file 7: GO analysis of *K*-means clusters.zip for the complete set of 30 clusters). GO analysis was carried out on the 30 clusters and many of the identified biological processes could be related to the tissue types they showed expression in. For example genes in Cluster K17 showed relatively high expression in imbibed seed samples (TTI 1.1) and an over-representation of genes involved in seed development and nutrient storage (Figure 4A and Additional file 7: GO analysis of *K*-means clusters.zip). Indeed 50%

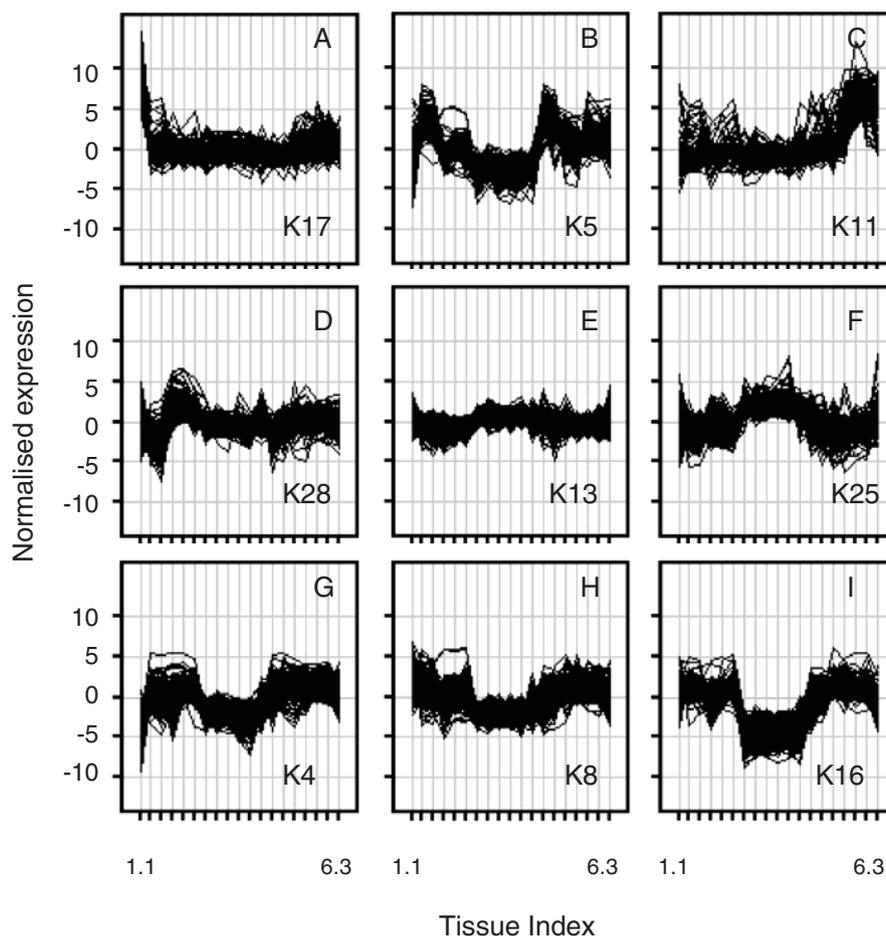


Figure 4 Tobacco gene co-expression. 30009 genes identified as differentially expressed across the TobEA samples were *K*-means clustered into 30 groups of co-expressed genes. Panels show normalised gene expression levels (y-axis) plotted across all 19 TobEA tissue types (x-axis) for clusters K17 (A), K5 (B), K11 (C), K28 (D), K13 (E), K25 (F), K4 (G), K8 (H) and K16 (I). The full set of 30 clusters can be seen in Additional file 6: Tobacco gene co-expression.ppt.

of the genes in Cluster K17 were among the 211 identified as specifically expressed in imbibed seed. Similarly genes in cluster K11 showed relatively high expression in buds and flowers (TTI 6.1-6.3), and genes in clusters K28 showed relatively high expression in young leaves (TTI 3.1-3.3) and these 2 clusters showed over-representation of genes related to floral development and Photosynthesis respectively (Figures 4C and D and Additional file 7: GO analysis of *K*-means clusters.zip). Another example of expression of genes involved in relevant processes for the respective tissue is offered by Cluster K5. These genes showed high expression in both root and stem samples and over-representation of genes involved in Xylem and Phloem development (Figure 4B and Additional file 7: GO analysis of *K*-means clusters.zip).

Clusters K13/K25 and K4/K8/K16 showed relatively high and relatively low levels of gene expression in mature and senescing leaf samples respectively (Figures 4E-I). The former set of clusters showed over-representation of genes involved in senescence related processes and the latter for genes involved in cell cycle and growth (Additional file 7: GO analysis of *K*-means clusters.zip). Together with the results for cluster K28, this supported the suggestion that transcriptional signature differences between young and old leaf tissues was based on genes involved in photosynthesis, cell division and growth, or senescence.

Analysis of tobacco transcription factor expression patterns

Transcription factors (TFs) play an important role in the regulation of gene expression. Comprehensive surveys of the TFs present in several other plant species have been undertaken [20-22]. Although complete sequence coverage is not available for the tobacco genome, the GSR sequence data generated by the TGI allowed the identification of 2513 TFs in *N. tabacum* as part of the TOBFAC database [23,24]. A BLASTN search was used to map these TFs to the tobacco unigenes assembled during this study (see methods and Additional file 8: Tobacco transcription factors.txt). The 2513 TFs were mapped to 779 unique unigenes, of which 772 were represented by at least one probe set (850 probe sets in total) on the microarray.

Of the TOBFAC TFs mapped to probe sets on the array, 87% were identified as expressed in at least one tissue. Figure 5 shows the number and percentage of tobacco microarray probe sets representing TFs for each of the TOBFAC families scored as expressed. For the majority of the TF families over 60% of their members were scored as expressed (Figure 5B). The exceptions to this were the SHI Related Sequence (SRS) and CPP families, where only 50% and 0% were expressed respectively (Figure 5B). However, it should be noted that

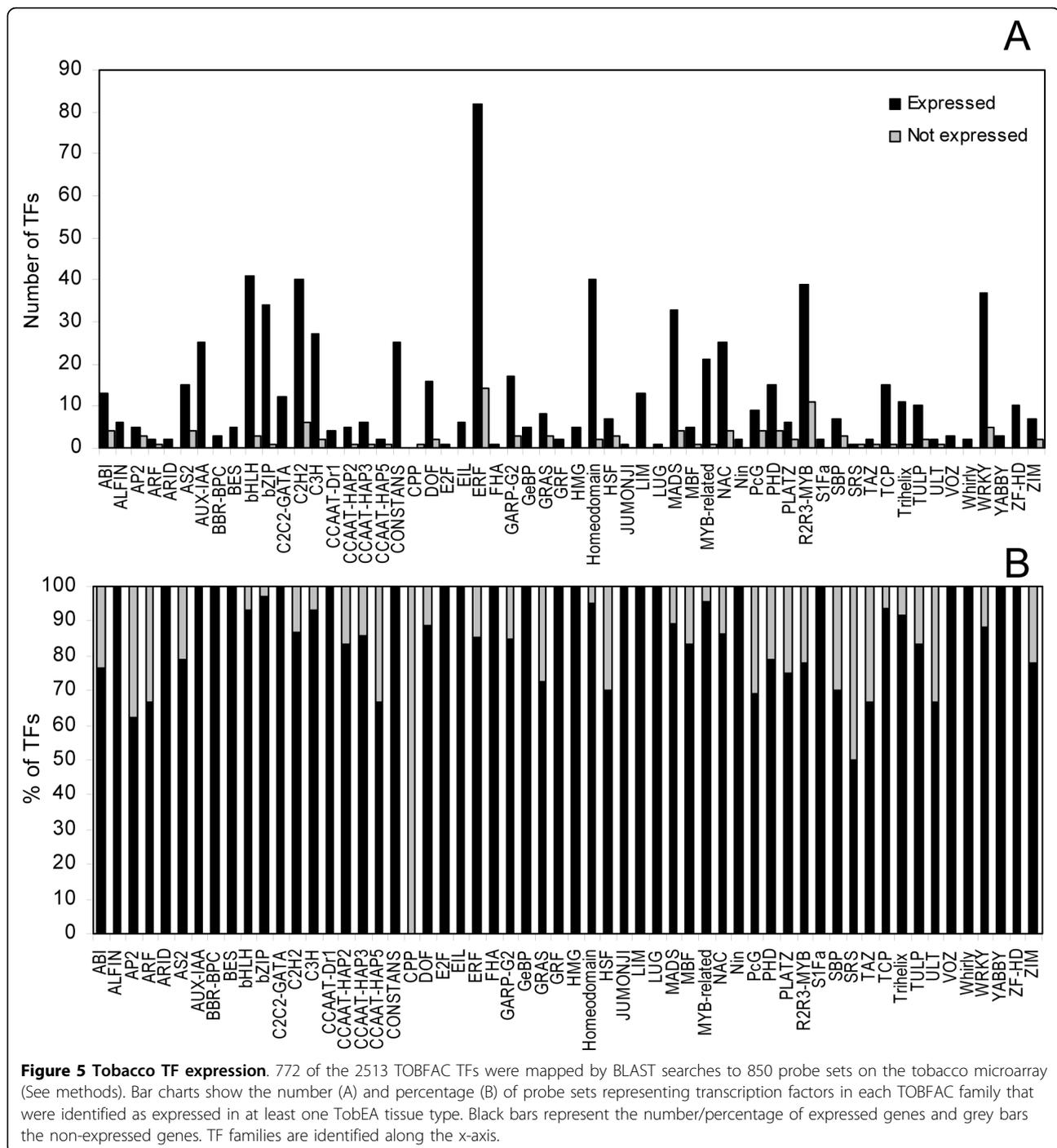
both these families were represented by two or less probe sets (Figure 5A).

The Ethylene Response Factors (ERF) are one of the largest TF families in Tobacco with over 230 members [24] and are represented by 96 probesets on the tobacco microarray. The ERF family did not show any strong trend in tissue specificity within TobEA, with a large number being expressed in all of the different tissue types (43.8% expressed in seed, 65.6% in root and between 52% and 61% in the other tissue types). The ERFs are divided into 10 subgroups, and a high percentage of group IX tobacco ERFs were previously shown to be induced in response to methyl jasmonate (MeJA) treatment [24]. The Group IX ERFs were represented by 11 probe sets on the tobacco microarray. Compared to the family as a whole, the group IX ERFs showed fewer members expressed in all other tissues (18.2% in seed, 54.5% in leaves, 45.5% in both stem and apices and 36.4% in buds and flowers) aside from roots, where 72.7% of them were scored as expressed. MeJA treatment has been shown to inhibit root growth in *Arabidopsis* [31] and tobacco [32] and it may be that the induction of these root expressed group IX ERFs by MeJA plays an important role in the inhibition of growth.

Identification of putative transcriptional networks

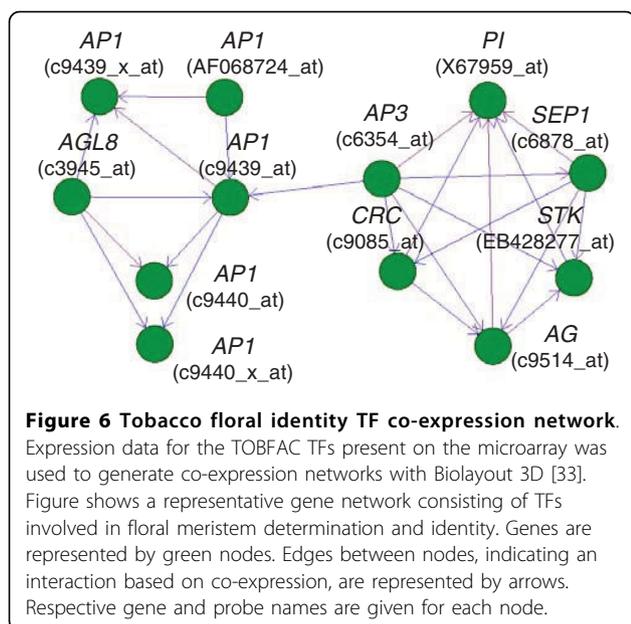
Although expression could be measured for a majority of TFs across a range of different tissues, this does not mean that their expression levels were similar between tissues. Correlation of expression level across different conditions provides a useful tool for functional genomics. The program Biolayout 3D was used to generate gene networks based on co-expression of TFs across all of the arrays in TobEA [33]. Individual probe sets for TFs are represented by nodes in the network and the edges between nodes represent correlated gene expression. Of the 850 probe sets representing TFs on the array, 297 were placed into 20 networks. Figure 6 shows an example of one of the networks generated including orthologs for 6 MADS-box TFs and one YABBY TF involved in floral meristem identity in other plant species (See Additional file 9: Floral identity gene correlation.doc for table of Pearson correlation values between nodes displayed in network).

It should be noted that an edge between two nodes does not necessarily indicate a causal interaction between these nodes as it could equally be interpreted as co-regulated expression of both nodes by a common regulator. However, several interactions suggested by the tobacco floral identity gene network were consistent with the gene regulatory network proposed for *Arabidopsis* (Figure 6; [34]). For example putative interactions between *APETALA1* (*API*) and *AP3*, between *AP3* and



AGAMOUS (AG) and both of these genes with *PISTILLATA (PI)* in tobacco are supported by results in Arabidopsis [34]. Some interactions shown in Arabidopsis were not represented in the tobacco TF network, such as a positive interaction between *API* and *PI* and a negative interaction between *AG* and *API*. Failure to see the latter interaction is not surprising as the network is based on positive correlation.

The TF network also suggests a positive interaction between *API* and *AGAMOUS LIKE 8 (AGL8)*, whereas *AGL8* is reportedly repressed by *API* in Arabidopsis (Figure 6 and [35]). This suggests a different relationship may exist between these genes in tobacco compared to Arabidopsis. Two putative tobacco *AGL8* orthologs were identified, so it was possible that only one had lost the repression by *API*. However, the other *AGL8* ortholog



was placed in a separate small network along with another *API* ortholog (as well as *AGL20*), suggesting that both of the tobacco *AGL8* orthologs were not repressed by *API* (data not shown).

The tobacco TF network appeared to consist of two distinct sub-clusters, one including *AGL8* and *API* orthologs and the other including orthologs for *AP3*, *PI*, *AG*, *CRABS CLAW* (*CRC*), *SEPALLATA1* (*SEP1*) and *SEEDSTICK* (*STK*) (Figure 6). This network structure is consistent with the former two genes being involved earlier on during floral determination, whereas the remaining genes have been implicated later in floral development and floral organ identity [36-42].

Development of web-based tools for TobEA

The tobacco microarray was designed from a unigene assembly based on 86188 sequences. Since the assembly was carried out there has been a significant increase in the sequence data available for tobacco, with over 250 k ESTs currently available at GenBank. A more recent set of 84602 unigenes was assembled from 239761 *N. tabacum* sequences by the SGN. The SGN also houses *NicotianaCyc*, a database linking the tobacco unigenes to the *MetaCyc* biochemical pathways [43]. To benefit from the greater genomic coverage and annotation information, unigenes present on the microarray were mapped to the current SGN unigene assembly (see methods). 41,648 of the probe sets on the microarray could be mapped to 31,929 of the SGN unigenes, indicating approximately 38% coverage of the SGN unigene build by the microarray. Approximately 81% of the mapped SGN unigenes were represented by a unique probe set, and over 99% by 4 probe sets or less.

Information of the temporal and spatial expression pattern of genes, along with co-expression across a broad range of conditions can be very useful in inferring function for unknown genes. In order to make the information contained within TobEA more accessible to other researchers and enable them to place information on their gene of interest within the context of its expression information, it was decided to integrate the TobEA expression data into the SGN website. Figure 7 shows some sample screen shots of the website, where users can access expression information on their gene of interest. The SGN Expression Data Module (SEDM) allows users to view the expression level of their unigene across the different samples included in TobEA (Figures 7A and 7B). The SEDM also allows users to identify other SGN unigenes that show correlated expression across the TobEA microarray data set (Figure 7C).

Discussion

Recent efforts, such as the TGI, have massively increased the sequence resources available for *N. tabacum*. However, the 58969 ESTs generated by this study still represent a significant contribution to the total number of ESTs available for tobacco and will prove very useful in applications such as the detection of varietal polymorphisms or gene model predictions. Along with 27219 other available sequences, these ESTs were assembled into a set of 40642 unigenes and used to design a custom Affymetrix expression microarray for tobacco. Affymetrix chips are a proven and widely accepted technology and the design service has been used successfully to generate a large number of custom chips for other research applications <http://www.affymetrix.com>. The resultant microarray was used to capture the temporal and spatial changes in gene expression throughout the lifecycle of the tobacco plant.

Considerable variation in gene expression was shown across the different tissues within TobEA, with the transcriptional signatures revealing relationships between samples. As observed for *Arabidopsis* the transcriptional profile of imbibed seed showed the greatest difference to all other samples causing them to be distinctly removed in principal component space [6]. PCA also showed a broad distinction between samples originating from root and shoot tissues. Studies in other plants have shown that the complement of genes expressed in a tissue is related to the biological function that tissue performs [6,7]. Analysis of the GO annotations shows that this is also true in tobacco. As well as spatial changes, temporal changes in gene expression were apparent over developmental series in the same tissues. This was more apparent for some tissues than others. For example PCA showed little separation between young and mature roots. There was, however, a tendency for stronger

Gene search results

| Organism | Locus name | Locus symbol | Allele name | Allele symbol | Allele phenotype |
|----------|---------------------|--------------|-------------|---------------|------------------|
| Tobacco | Nicotiana Agamous 1 | TOBNAG1A | | | |

Unigene Basic Information

Unigene ID: SGN-U431152
 Unigene Build: Nicotiana tabacum #2
 Date: 2008-07-17
 Organism: none
 Alternative ID: none
 mRNA sequence: Length: 1116 bp

Associated Loci (1)

| Locus symbol | Locus name | Gene activity |
|--------------|---------------------|----------------------|
| TOBNAG1A | Nicotiana Agamous 1 | floral homeotic gene |

Expression Data (1)

| Template | Experimental design | Experiments | Hybridizations |
|----------|---------------------|-------------|----------------|
| C9514_at | TobEA | 19 | 100 |

Expression Data for Template: C9514_at

Template Basic Information

Template ID: 26922
 Name: C9514_at
 Type: unigene
 SGN internal mapping: SGN-U431152
 External mapping: GenBank: L23925 @
 Platform Name: Affymetrix TobEA

Expression Data by Experiments (19)

| Experimental design name | Experiment name | Tissue | Experiment replicates | Replicates | pValue<0.05 | Mean | Median | SD | CV |
|--------------------------------|--------------------------------|------------|-----------------------|------------|-------------|-------|--------|------|----|
| TobEA | TobEA seed | Seed | 6 | 6 | 3.97 | 4.13 | 1.31 | 0.33 | |
| | TobEA cotyledons | Cotyledons | 5 | 5 | 4.75 | 4.68 | 0.92 | 0.19 | |
| TobEA young root | TobEA young root | Root | 6 | 6 | 3.84 | 4.00 | 1.28 | 0.33 | |
| | TobEA mature root | Root | 5 | 5 | 4.62 | 4.47 | 0.42 | 0.09 | |
| TobEA young shoot | TobEA young shoot | Shoot | 5 | 5 | 5.01 | 5.03 | 0.48 | 0.10 | |
| | TobEA vegetative shoot apex | Shoot | 8 | 8 | 3.83 | 3.98 | 0.80 | 0.21 | |
| TobEA floral shoot apex | TobEA floral shoot apex | Shoot | 5 | 5 | 8.75 | 8.75 | 0.30 | 0.03 | |
| | TobEA lower stem | Stem | 5 | 5 | 4.09 | 4.53 | 1.54 | 0.38 | |
| TobEA upper stem | TobEA upper stem | Stem | 5 | 5 | 4.49 | 4.58 | 0.87 | 0.19 | |
| | TobEA young leaf | Leaf | 5 | 5 | 4.68 | 4.92 | 0.88 | 0.18 | |
| TobEA mature leaf | TobEA mature leaf | Leaf | 5 | 5 | 4.38 | 4.26 | 0.75 | 0.17 | |
| | TobEA early senescent leaf | Leaf | 5 | 5 | 5.05 | 4.98 | 0.64 | 0.13 | |
| TobEA mid/early senescent leaf | TobEA mid/early senescent leaf | Leaf | 5 | 5 | 5.01 | 4.98 | 0.39 | 0.08 | |
| | TobEA mid/late senescent leaf | Leaf | 5 | 5 | 3.93 | 5.16 | 1.92 | 0.49 | |
| TobEA late senescent leaf | TobEA late senescent leaf | Leaf | 5 | 5 | 5.00 | 4.86 | 0.38 | 0.08 | |
| | TobEA cauline leaf | Leaf | 5 | 5 | 4.47 | 4.44 | 0.36 | 0.08 | |
| TobEA closed bud | TobEA closed bud | Bud | 5 | 5 | 10.86 | 10.75 | 0.26 | 0.02 | |
| | TobEA open bud | Bud | 5 | 5 | 10.89 | 10.69 | 0.37 | 0.03 | |
| TobEA flower | Flower | Flower | 5 | 5 | 10.78 | 10.84 | 0.27 | 0.02 | |

Correlation analysis (1)

Correlation Analysis ID: 1
 Methodology: Pearson Correlation An
 Description: This Pearson correlati

Correlation Type: Positive correlations (302 templates correlated)

| Template | Pearson | SGN unigene |
|-------------|----------|-------------|
| EB426081_at | 0.897872 | SGN-U444801 |
| EB424763_at | 0.882099 | SGN-U472386 |
| X67959_at | 0.880392 | SGN-U483102 |
| C6170_at | | |
| EB425199_at | | |
| U63163_at | | |
| C6354_at | | |

Expression Data for Template: X67959_at

Template Basic Information

Template ID: 43735
 Name: X67959_at
 Type: unigene
 SGN internal mapping: SGN-U483102
 External mapping: GenBank: X67959 @
 Platform Name: Affymetrix TobEA

Annotations associated to Unigene: SGN-U483102

Manual annotations
 None manual annotation was found for this unigene

Blast Annotations [Show All]

| Database | Program | Best Match | Score | E-value | Identity | Alignment Length | Frame |
|----------|---------|--------------------------------|-------|---------|----------|------------------|-------|
| genbank | blastx | gi 417063 sp Q03416 GLOB_TOBAC | 417 | 4e-115 | 100 | 627 bp (81.6%) | 1 |

Description: gi|417063|sp|Q03416|GLOB_TOBAC Floral homeotic protein GLOBOSA gi|486750|pir|I53226 homeotic p...

Figure 7 Data access using SEDM (SGN Expression Data Module). Screen shots summarising user access to expression data via the SGN Expression Data Module. Data is available to users from a direct search or via from a search of another SGN element (A). The expression data is shown in the template page with data about the probes, annotation and correlation analysis associated (B). The correlation analysis section can be used to access to other probe-sets with similar expression across the TobEA microarray dataset (C).

expression of genes related to cell division, growth and expansion (along with photosynthesis related genes in green samples) in younger tissues and stronger expression of senescence related genes in older tissues. On this basis there was a clear separation between young and old leaf samples, consistent with findings for *Arabidopsis* [6].

Consistency of the results contained within TobEA with previous results in other plants suggests that it is a high quality expression data set and a good representation of the spatial and temporal changes in gene expression throughout the lifecycle of the tobacco plant. Together with close clustering of independent replicate samples this also indicates that the microarray can be used to reliably and reproducibly measure gene expression in tobacco.

Since the study was initiated larger unigene sets have been assembled for tobacco from the increased amount of sequence data available. These new unigene assemblies likely represent greater coverage of the tobacco genome. Mapping of the microarray probe sets to the current unigene build from the SGN showed that the microarray still represents around 38% coverage of tobacco genes. Mapping of the probe sets to any future unigene assemblies will be maintained to ensure that the annotation information available for the array remains up to date.

As demonstrated by TOBFAC, the tobacco GSR sequence data provides the potential to identify genes that are not represented in the available EST libraries [23,24]. This is most likely the case for rare transcripts or those showing temporal, spatial, or alternative forms of regulation that is not captured in the conditions sampled by the current EST libraries. In future assemblies, the incorporation of sequence data such as GSRs with EST sequences has the potential to provide more comprehensive coverage of the genes present in the tobacco genome than the EST data alone.

The microarray represents a very valuable tool for research in tobacco, allowing comprehensive expression studies to be carried out in the plant. To maximise the benefit of this the array has been made publicly available for use via the Affymetrix service at NASC <http://affymetrix.arabidopsis.info>. The microarray might also provide a very useful tool to researchers in closely related species, in particular other *Nicotiana* species, via cross-species hybridisations [16,44].

TobEA provides a useful reference data set to any future experiments carried out using the microarray, allowing a more detailed characterisation of the differentially expressed genes identified. To facilitate use of the information contained in TobEA, a set of web-based tools were developed. These tools currently allow researchers to view the expression level of their gene(s) of interest across the different tissues in the atlas and

identify other genes that show highly correlated expression via the SGN website. This will enable individual users to carry out functional genomic analysis allowing the annotation of the SGN unigenes to be updated by experts from all different fields of biology via the locus editor feature of the SGN website.

As further experiments are carried out using the tobacco microarray, the data will be uploaded into the SGN expression database SEDM. The SEDM can store results for any Solanaceae species, and is capable of storing data sets from both microarray and direct sequencing studies. Enabling users to upload expression data to the SGN expression database means it has the potential to become a very useful resource for all researchers in the Solenaceae.

The value of TobEA is demonstrated by the study of specific sets of genes such as transcription factors. Due to their regulatory role, TFs represent good targets for gross modifications of biochemical or signalling pathways for various biotechnological applications [45]. Over 30% of the TFs identified by Rushton *et al.*, [23,24] in a genome scale survey of tobacco could be mapped to unigenes on the microarray. Simple correlation of TF expression levels across the tissues in TobEA enabled the construction of a gene network involved in floral determination and floral organ identity. Several of the connections observed in this network were consistent with known interactions from other plant species. Extending this type of network analysis to include genes beyond TFs could allow the identification of putative transcriptional networks for specific metabolic or signalling pathways. This analysis could identify regulatory targets to enable the manipulation of these pathways in efforts to increase or decrease the production of various secondary metabolites in tobacco.

Conclusions

An Affymetrix microarray was developed that allows reliable and reproducible measurement of gene expression in tobacco. This array is available for public use via the NASC Affymetrix service. The array was used to produce the Tobacco Expression Atlas (TobEA), which provides a snap-shot of gene expression from multiple tissues throughout the lifecycle of the tobacco plant. A set of web based tools were developed in order to facilitate public access to this data via the SGN web site <http://solgenomics.net/>. It is intended that the database will be extended in future with the addition of further expression datasets, not only from tobacco but other plants of the Solanaceae family. The SGN welcomes the donation of such data from the research community. TobEA represents a valuable resource for functional genomic analysis in plants and has potential application in research leading to biotechnological and agronomic improvements to tobacco as well as related species.

Methods

Plant materials and growth

Tobacco plants (*Nicotiana tabacum* L. cv. K326; Burley 21; Samsun NN and TN86) were grown on soil (Levington M2), or sand (for root samples) in a greenhouse under a light/dark regime of 16 h light and 8 h darkness. Plants for the EST libraries not required to produce flowers were topped after 12 weeks by removal of the apical meristem. Samples were taken from at least 5 independent replicate plants (30 plants for trichomes, and 1 g for seed) and flash frozen in liquid nitrogen. Leaf senescence was determined empirically by the level of Chlorosis displayed in mature leaves. Chlorophyll levels were measured using a CCM-200 Chlorophyll meter (Opti-sciences Inc., NH, USA), with Chlorophyll Content Index (CCI) readings of between 16 - 25 for early senescent leaves (TTI 3.5), 8 - 16 for mid-early senescent leaves (TTI 3.6), 2 - 8 for mid-late senescent leaves (TTI 3.7) and <2 for late senescent leaves (TTI 3.8). See Additional file 10: TobEA samples.xls for further detail of the growth conditions of individual array samples.

Preparation of RNA samples

Trichomes were isolated from frozen stems using the method of Aziz *et al.* [46] and RNA extracted using Tri Reagent (Sigma, Gillingham, UK). Total RNA for the remaining cDNA samples was extracted according to the method of Dean *et al.* [47] and mRNA was isolated using DynalBeads (Invitrogen, Paisley, UK). Total RNA was extracted using Trizol for the microarray samples (Invitrogen, Paisley, UK) and purified using RNeasy spin columns (Qiagen, Crawley, UK).

Sequencing of cDNA libraries

See Additional file 11: Tobacco EST generation.doc for description of the construction and processing of the cDNA libraries. cDNA clones were single pass sequenced from the 5' end using the M13 primer. All sequence information was collated and trimmed for vector sequence. Sequences from the TT1, TT2, TT3, TT4 and TT5 libraries had their 5' sequence tags removed after assignment to their respective libraries. All EST sequences were uploaded to dbEST at NCBI (Genbank accession numbers DV075759, DV157477-DV162736, DV998727-DV998768, DV998776-DV998788, DV998796-DV999999, EB424600-EB452264, EB677167-EB684271, FG633215-FG645637 and FL577778-FL577786).

Generation of tobacco unigene sequences for microarray design

ESTs were assembled into a unigene set using CAP3 [48] with parameters set as $p = 95$ and $o = 40$.

Sequences were filtered for polyA tails, low complexity, low quality and short sequences. Sequences were further processed to mask repeat, vector and organelle sequences using EGAssembler <http://egassembler.hgc.jp>. Unigenes were annotated using BLASTX based on the best hit (e -value $< 1 \times 10^{-10}$) against a database of protein sequences from *Arabidopsis thaliana* (The Arabidopsis Information Resource [TAIR] 7.0 release) and also using the program Blast2GO [49] against a database of non-redundant proteins from Genbank.

Microarray data analysis

Samples were run from 5 independent biological replicates for each tissue type in TobEA. All of the microarray data is available from ArrayExpress (E-MTAB-176). The data set includes an additional 5 microarrays (samples 41, 58, 59, 61 and 79), that were excluded from the statistical analysis of this study to maintain a balanced design, but are present in the SEDM data set. All microarray data were pre-processed using the MAS5 algorithm to facilitate the future addition of further samples to TobEA [29]. Statistical tests were carried out as described in the text using the program GeneSpring ver 10.0 (Agilent, Warriner, UK) with Benjamini and Hochberg false discovery rate multiple testing correction (MTC; [50]). GO analysis was carried out using a custom Perl script based on GO annotation from the TAIR 8 release (as of April 2009). Only genes with a significant BLASTX hit to Arabidopsis were included in a fishers exact test (with Benjamini and Hochberg MTC [50]) comparing the test list with the whole array as a background list. TF co-expression networks were produced using the tool Biolayout 3D ver 3.0 [33], with a Pearson correlation value of 0.8.

Mapping of TFs to unigenes on the microarray

TFs identified by Rushton *et al.*, were tested against a custom database of the tobacco unigenes using BLASTN (e -value $< 1 \times 10^{-10}$). Only hits with >75% ID over at least 100 bp were considered. The best hit was selected for each TF based on highest bit score, using the length of the aligned region as a deciding factor where bits scores were equal. Unigene hits for 7 of the TFs could not be distinguished based on the above measure so both unigenes were annotated with the same TF information. Similarly 13 unigenes showed equal hits against two TFs so were annotated with both.

Mapping of microarray probes to SGN Unigenes

Tobacco Affymetrix probes were mapped to SGN unigenes in a two step process. First, mapping was established between the Genbank accessions or the clone

names of the Tobacco Affymetrix probes and the database external references or clone names of the ESTs from the SGN database [26]. Then the SGN unigene members (EST) and SGN unigenes were linked to the Tobacco Affymetrix probes by joining the tables through the common member sequences. The Affymetrix chip contains a number of non-Solanaceae probes, mainly from *Arabidopsis thaliana*, representing positive and negative controls, which obviously were not included in this mapping. Of the 43,768 tobacco probes on the chip, 41,648 could be mapped to 31,929 SGN unigenes.

SGN Expression Data Module

SEDM was developed to allow users access to TobEA expression data via the SGN website. Data in SEDM are presented for 100 microarray hybridisations with the additional 5 hybridisations provided by extra replicates for some of the 19 conditions as described above. MAS5 pre-processed expression values are presented for SGN unigenes based on mappings from the probe sets as described above. Co-expressed unigenes were identified based on pair-wise Pearson correlation of all probe sets on the microarray using the `cor` function of R (version 2.7.0) as were the values presented in Additional file 9: Floral identity gene correlation.doc. Only those showing a Pearson correlation co-efficient of <-0.8 or >0.8 are presented.

Additional file 1: ATC unigene lengths. Histogram showing tobacco unigene lengths (base pairs).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S1.PPT>]

Additional file 2: ATC unigene sequences. FastA format file containing ATC unigene sequences.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S2.ZIP>]

Additional file 3: ATC unigene annotations. ATC Unigene annotation information file. Genbank IDs of unigene member sequences and information of *Arabidopsis* and nr best hit information presented for each unigene.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S3.TXT>]

Additional file 4: GO analysis of young versus mature leaves. Zip file containing lists of probes sets and GO analysis results for genes differentially expressed in young versus mature leaves (cluster L0 to L3).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S4.ZIP>]

Additional file 5: GO analysis of tissue specific genes. Zip file containing lists of probes sets and GO analysis results for genes showing tissue specific expression.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S5.ZIP>]

Additional file 6: Tobacco gene co-expression. Charts showing expression level for genes in *K*-means clusters K0-K29.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S6.PPT>]

Additional file 7: GO analysis of *K*-means clusters. Zip file containing lists of probes sets and GO analysis results for in each of the 30 *K*-means clusters (K0- K29).

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S7.ZIP>]

Additional file 8: Tobacco transcription factors. Table summarising results of mapping TOBFAC TFs to the ATC Unigenes.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S8.TXT>]

Additional file 9: Floral identity gene correlation. Table showing Pearson correlation values between nodes in the floral gene identity co-expression network.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S9.DOC>]

Additional file 10: TobEA samples. Table containing information on the biological source of samples contained in TobEA.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S10.XLS>]

Additional file 11: Tobacco EST generation. Supplementary text with more detailed description of production and processing of EST library samples.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-11-142-S11.DOC>]

Acknowledgements

The normalized EST libraries (TT1-TT5) were generated commercially by GATC biotech (Konstanz, Germany) and the virtual subtraction ESTs libraries (KP1, KR2, KR3, KL4, KL5, KN6, KT7, KF8, KG9, BL12 and TL13) by Biogemma (Paris, France). The authors would specifically like to thank François Dorlhac de Borne, who carried out the unigene assembly as part of the ESTobacco project. The microarray was produced by the Affymetrix custom design service. The authors would also like to thank Aiping Yang for RNA preparation, Robert Lister for plant husbandry and Marina Murphy for help with manuscript preparation.

Author details

¹Advanced Technologies (Cambridge) Ltd, 210 Cambridge Science Park, Milton Road, Cambridge, CB4 0WA, UK. ²Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, New York 14853-1801, USA. ³Current address: Graduate School of Life Sciences, University of Cambridge, Cambridge, CB2 1RX, UK.

Authors' contributions

KDE conceived and initiated the project, carried out bioinformatics analysis and wrote the manuscript. GWS was involved in the initiation of the project and sample processing. FA carried out the sample processing and produced the microarray data. AB and LAM mapped the microarray unigenes to the SGN unigene set and developed the web-based tools. SAC and LJ were involved in the conception and design of the EST sequencing and unigene assembly project and in communication with Affymetrix over the design of the array. All authors were involved in editing of manuscript. All authors have read and approved the final manuscript.

Received: 7 December 2009

Accepted: 26 February 2010 Published: 26 February 2010

References

1. Hoekema A, Hirsch PR, Hooykaas PJJ, Schilperoort RA: **A binary plant vector strategy based on separation of vir- and T-region of the *Agrobacterium tumefaciens* Ti-plasmid.** *Nature* 1983, **303**(5913):179-180.
2. Gerstel DU: **Segregation in new allopolyploids of *Nicotiana*.** I. Comparison of 6x (*N. tabacum* x *tomentosiformis*) and 6x (*N. tabacum* x *otophora*). *Genetics* 1960, **45**:1723-1734.
3. Goodspeed TH: **The genus *Nicotiana*.** Massachusetts: Waltham 1954.
4. Murad L, Lim KY, Christopodulou V, Matyasek R, Lichtenstein CP, Kovarik A, Leitch AR: **The origin of tobacco's T genome is traced to a particular lineage within *Nicotiana tomentosiformis* (Solanaceae).** *Am J Bot* 2002, **89**(6):921-928.
5. Leitch IJ, Hanson L, Lim KY, Kovarik A, Chase MW, Clarkson JJ, Leitch AR: **The Ups and Downs of Genome Size Evolution in Polyploid Species of *Nicotiana* (Solanaceae).** *Ann Bot* 2008, **101**(6):805-814.
6. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, Scholkopf B, Weigel D, Lohmann JU: **A gene expression map of *Arabidopsis thaliana* development.** *Nat Genet* 2005, **37**(5):501-506.
7. Druka A, Muehlbauer G, Druke I, Caldo R, Baumann U, Rostoks N, Schrieber A, Wise R, Close T, Kleinhofs A, et al: **An atlas of gene expression from seed to seed through barley development.** *Functional & Intergrative Genomics* 2006, **6**(3):202-211.
8. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W: **GENEVESTIGATOR. *Arabidopsis* Microarray Database and Analysis Toolbox.** *Plant Physiol* 2004, **136**(1):2621-2632.
9. Schreiber A, Sutton T, Caldo R, Kalashyan E, Lovell B, Mayo G, Muehlbauer G, Druka A, Waugh R, Wise R, et al: **Comparative transcriptomics in the *Triticaceae*.** *BMC Genomics* 2009, **10**(1):285.
10. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(16):6062-6067.
11. Kilian J, Whitehead D, Horak J, Wanke D, Weinl S, Batistic O, D'Angelo C, Bombberg-Bauer E, Kudla J, Harter K: **The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses.** *The Plant Journal* 2007, **50**(2):347-363.
12. Goda H, Sasaki E, Akiyama K, Maruyama-Nakashita A, Nakabayashi K, Li W, Ogawa M, Yamauchi Y, Preston J, et al: **The AtGenExpress hormone and chemical treatment data set: experimental design, data evaluation, model data analysis and data access.** *The Plant Journal* 2008, **55**(3):526-542.
13. Manfield IW, Jen CH, Pinney JW, Michalopoulos I, Bradford JR, Gilmartin PM, Westhead DR: ***Arabidopsis* Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis.** *Nucl Acids Res* 2006, **34**:W504-509.
14. Wei H, Persson S, Mehta T, Srinivasasainagendra V, Chen L, Page GP, Somerville C, Loraine A: **Transcriptional Coordination of the Metabolic Network in *Arabidopsis*.** *Plant Physiol* 2006, **142**(2):762-774.
15. Thimm O, Blasing O, Gibon Y, Nagel A, et al: **mapman: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.** *The Plant Journal* 2004, **37**(6):914-939.
16. Craigan DJ, James N, Okyere J, Higgins J, Jotham J, May S: **NASCArrays: a repository for microarray data generated by NASC's transcriptomics service.** *Nucl Acids Res* 2004, **32**:D575-577.
17. Mueller LA, Zhang P, Rhee SY: **AraCyc: A Biochemical Pathway Database for *Arabidopsis*.** *Plant Physiol* 2003, **132**(2):453-460.
18. Bassel GW, Provart NJ: **Gene Expression Analyses for Elucidating Mechanisms of Hormonal Action in Plants.** *Plant Hormones* Humana Press 2009, **495**:1-17.
19. Toufighi K, Brady SM, Austin R, Ly E, Provart NJ: **The Botany Array Resource: e-Northern, Expression Angling, and promoter analyses.** *The Plant Journal* 2005, **43**(1):153-163.
20. Riechmann JL, Heard J, Martin G, Reuber L, Jiang ZC, Keddie J, Adam L, Pineda O, Ratcliffe OJ, et al: ***Arabidopsis* Transcription Factors: Genome-Wide Comparative Analysis Among Eukaryotes.** *Science* 2000, **290**(5499):2105-2110.
21. Gao G, Zhong Y, Guo A, Zhu Q, Tang W, Zheng W, Gu X, Wei L, Luo J: **DRTF: a database of rice transcription factors.** *Bioinformatics* 2006, **22**(10):1286-1287.
22. Zhu QH, Guo AY, Gao G, Zhong YF, Xu M, Huang M, Luo J: **DPTF: a database of poplar transcription factors.** *Bioinformatics* 2007, **23**(10):1307-1308.
23. Rushton P, Bokowiec M, Laudeman T, Brannock J, Chen X, Timko M: **TOBFAC: the database of tobacco transcription factors.** *BMC Bioinformatics* 2008, **9**(1).
24. Rushton PJ, Bokowiec MT, Han S, Zhang H, Brannock JF, Chen X, Laudeman TW, Timko MP: **Tobacco Transcription Factors: Novel Insights into Transcriptional Regulation in the Solanaceae.** *Plant Physiol* 2008, **147**(1):280-295.
25. Albert R: **Network Inference, Analysis, and Modeling in Systems Biology.** *Plant Cell* 2007, **19**(11):3327-3338.
26. Mueller LA, Solow TH, Taylor N, Skwarecki B, Buels R, Binns J, Lin C, Wright MH, Ahrens R, Wang Y, et al: **The SOL Genomics Network. A Comparative Resource for Solanaceae Biology and Beyond.** *Plant Physiol* 2005, **138**(3):1310-1317.
27. Ko MSH: **An 'equalized cDNA library' by the reassociation of short double-stranded cDNAs.** *Nucl Acids Res* 1990, **18**(19):5705-5711.
28. Nelson PS, Hawkins V, Schummer M, Bumgarner R, Ng WL, Ideker T, Ferguson C, Hood L: **Negative selection: a method for obtaining low-abundance cDNAs using high-density cDNA clone arrays.** *Genetic Analysis: Biomolecular Engineering* 1999, **15**(6):209-215.
29. Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18**(12):1585-1592.
30. Shoji T, Narita NN, Hayashi K, Asada J, Hamada T, Sonobe S, Nakajima K, Hashimoto T: **Plant-Specific Microtubule-Associated Protein SPIRAL2 Is Required for Anisotropic Growth in *Arabidopsis*.** *Plant Physiol* 2004, **136**(4):3933-3944.
31. Staswick PE, Su W, Howell SH: **Methyl jasmonate inhibition of root growth and induction of a leaf protein are decreased in an *Arabidopsis thaliana* mutant.** *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(15):6837-6840.
32. Shoji T, Ogawa T, Hashimoto T: **Jasmonate-Induced Nicotine Formation in Tobacco Is Mediated by Tobacco COI1 and JAZ Genes.** *Plant Cell Physiol* 2008, **49**(7):1003-1012.
33. Freeman TC, Goldovsky L, Brosch M, van Dongen S, Mazière P, Grocock RJ, Freilich S, Thornton J, Enright AJ: **Construction, Visualisation, and Clustering of Transcription Networks from Microarray Expression Data.** *PLoS Computational Biology* 2007, **3**(10):2032-2042.
34. Espinosa-Soto C, Padilla-Longoria P, Alvarez-Buylla ER: **A Gene Regulatory Network Model for Cell-Fate Determination during *Arabidopsis thaliana* Flower Development That Is Robust and Recovers Experimental Gene Expression Profiles.** *Plant Cell* 2004, **16**(11):2923-2939.
35. Mandel MA, Yanofsky MF: **The *Arabidopsis* AGL8 MADS Box Gene Is Expressed in Inflorescence Meristems and Is Negatively Regulated by APETALA1.** *Plant Cell* 1995, **7**(11):1763-1771.
36. Yanofsky MF, Ma H, Bowman JL, Drews GN, Feldmann KA, Meyerowitz EM: **The protein encoded by the *Arabidopsis* homeotic gene *agamous* resembles transcription factors.** *Nature* 1990, **346**(6279):35-39.
37. Goto K, Meyerowitz EM: **Function and regulation of the *Arabidopsis* floral homeotic gene PISTILLATA.** *Genes & Development* 1994, **8**(13):1548-1560.
38. Rounsley SD, Ditta GS, Yanofsky MF: **Diverse Roles for MADS Box Genes in *Arabidopsis* Development.** *Plant Cell* 1995, **7**(8):1259-1269.
39. Jack T, Brockman LL, Meyerowitz EM: **The homeotic gene APETALA3 of *Arabidopsis thaliana* encodes a MADS box and is expressed in petals and stamens.** *Cell* 1992, **68**(4):683-697.
40. Hempel FD, Weigel D, Mandel MA, Ditta G, Zambryski PC, Feldman LJ, Yanofsky MF: **Floral determination and expression of floral regulatory genes in *Arabidopsis*.** *Development* 1997, **124**(19):3845-3853.
41. Bowman JL, Smyth DR: **CRABS CLAW, a gene that regulates carpel and nectary development in *Arabidopsis*, encodes a novel protein with zinc finger and helix-loop-helix domains.** *Development* 1999, **126**(11):2387-2396.
42. Ferrario S, Immink RGH, Shchennikova A, Busscher-Lange J, Angenent GC: **The MADS Box Gene FBP2 Is Required for SEPALLATA Function in *Petunia*.** *Plant Cell* 2003, **15**(4):914-925.
43. Krieger CJ, Zhang P, Mueller LA, Wang A, Paley S, Arnaud M, Pick J, Rhee SY, Karp PD: **MetaCyc: a multiorganism database of metabolic pathways and enzymes.** *Nucl Acids Res* 2004, **32**:D438-442.

44. Hammond J, Broadley M, Craigon D, Higgins J, Emmerson Z, Townsend H, White P, May S: **Using genomic DNA-based probe-selection to improve the sensitivity of high-density oligonucleotide arrays when applied to heterologous species.** *Plant Methods* 2005, **1**(1):10.
45. Century K, Reuber TL, Ratcliffe OJ: **Regulating the Regulators: The Future Prospects for Transcription-Factor-Based Agricultural Biotechnology Products.** *Plant Physiol* 2008, **147**(1):20-29.
46. Aziz N, Paiva NL, May GD, Dixon RA: **Transcriptome analysis of alfalfa glandular trichomes.** *Planta* 2005, **221**(1):23-38.
47. Dean C, Elzen van den P, Tamaki S, Dunsmuir P, Bedbrook J: **Differential expression of the eight genes of the petunia ribulose biphosphate carboxylase small subunit multi-gene family.** *EMBO Journal* 1985, **4**(12):3055-3061.
48. Huang X, Madan A: **CAP3: A DNA Sequence Assembly Program.** *Genome Res* 1999, **9**(9):868-877.
49. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**(18):3674-3676.
50. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society* 1995, **57**(1):289-300.

doi:10.1186/1471-2164-11-142

Cite this article as: Edwards *et al.*: **TobEA: an atlas of tobacco gene expression from seed to senescence.** *BMC Genomics* 2010 **11**:142.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

