High-throughput localisation of novel plant proteins using virus-based vectors

N. Escobar, S. Haupt, S. Chapman, G. Pogue* and K.J. Oparka

The *Arabidopsis* genome project has generated a large number of sequences that encode for proteins of unknown function. In the post-genomics era, characterising the location and function of these proteins in plant cells will become a major challenge. In order to ascribe subcellular targeting information to novel proteins, we have developed a high-throughput screening procedure involving viral vectors based on tobacco mosaic virus (TMV). In this system, random cDNA inserts produced from *Nicotiana* root mRNA were inserted *en masse* into TMV vectors containing the gene for green fluorescent protein (GFP) to allow subsequent infection and fluorescence screening of leaves infected with the libraries. Each infection site that arose on an infected leaf expressed a unique cDNA-GFP fusion, allowing the subcellular 'address' of a protein to be ascribed at high resolution under the confocal laser scanning microscope (CLSM).

Library construction To prevent masking of the targeting sequences that frequently occur at the ends of proteins, methods were developed to express fusions to both the amino- and carboxy-termini of *gfp*. For expression of protein sequences fused to the amino-terminus of *gfp* it was necessary to produce partial cDNAs that contained the initiating methionine codon, found near the 5' end of the mRNAs, but lacked the termination codon, found near the 3' ends. Therefore, to create a library of cDNAs fused to the 5' end of *gfp*, first-strand cDNA synthesis was performed with a random primer and second-strand synthesis was performed using a method that is dependent upon the cap structure at the 5' end of mRNAs. The cDNAs produced were subjected to size fractionation, so that small inserts containing little targeting information would not be included in the library, and inserted into a TMV vector just 5' of *gfp*. The natural initiation codon of *gfp* had been removed so that any cDNAs that were not fused in frame with *gfp* would not produce fluorescent protein on inoculation of plants. First-strand cDNA synthesis was carried out with an oligo-dT primer, complementary to the 3' ends of the mRNAs, to generate a library of cDNAs fused to the 3' end of *gfp*. After standard second-strand synthesis, the cDNA was normalized to decrease the representation of cDNAs derived from abundant mRNAs. Subsequently the cDNA was size selected and inserted 3' of *gfp* in a TMV vector lacking the stop codon of *gfp*.

Screening of infection sites Following inoculation of the cDNA-GFP libraries onto leaves, individual leaves were taped down onto large glass slides and the leaf surface scanned under the CLSM. To avoid the use of coverslips, water-dipping lenses were employed to facilitate high-throughput screening. To image the leaf surface, a water droplet was suspended between the objective and leaf epidermis, and then 'dragged' across the leaf surface between adjacent infection sites. In the case of the 3'cDNA-GFP library, approximately 90% of the infection sites displayed a cytosolic protein localisation; that is, the majority of proteins were not targeted to discrete organelles. However, in the case of the 5'cDNA-GFP library, over 50% of the GFP-tagged proteins were targeted to discrete organelles. This result was expected as many proteins exhibit discrete targeting signals at their amino terminus. An example of the screening procedure is shown in Figure 1. Following the detection of a novel subcellular localisation, the infection sites were excised and the virus passaged onto new leaves to confirm the protein localisation, to remove any contaminating viruses, and to produce material for recloning of the cDNA inserts. Reverse transcription of RNA extracted from the passaged tissue was primed with an oligonucleotide complementary to TMV sequence downstream of the cDNA sequences. The reverse transcribed products were amplified through PCR using *gfp*- and TMV-specific primers that flanked the cDNA inserts, prior to cloning of the amplified products and determination of the nucleotide sequences of the inserts. Sequencing of the cDNA inserts revealed that many proteins were targeted to the predicted organelle, based on the known functions of the proteins encoded by these sequences.

Plasmodesmatal proteins As there are no known genes that encode for specific plasmodesmatal proteins, we were curious to determine whether the screen would reveal novel proteins that are intergral to, or interact with, components of plasmodesmata. In an initial screen of over 15,000 infection sites, we detected 12 unique proteins that accumulated specifically

*Large Scale Biology Corporation

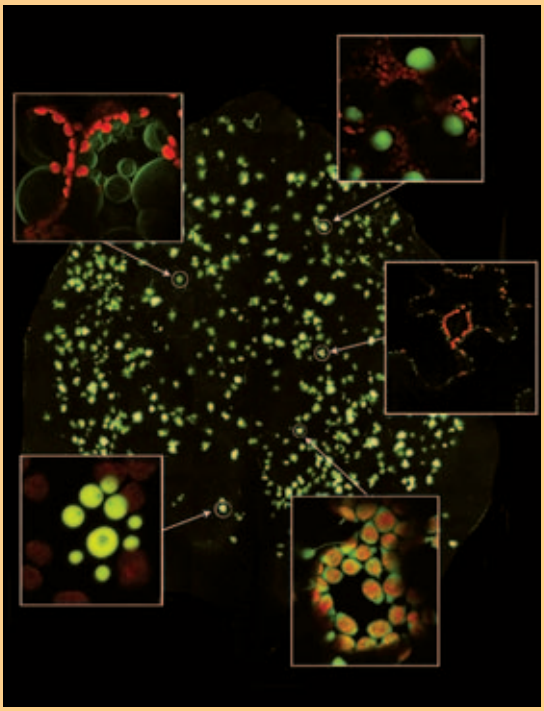

Figure 1 Virus-based screening of cDNA-GFP fusion libraries. Each fluorescent infection site that arises on the leaf surface contains a unique cDNA-GFP fusion protein. The infection sites are imaged using a confocal laser scanning microscope fitted with water-dipping lenses. Examples of protein localisations are shown in the inset images. Following detection of a novel protein, the infection sites are excised for subsequent determination of the cDNA sequence.

within plasmodesmata. The sequences encoding these fusion proteins were recloned, sequenced and introduced back into TMV vectors to confirm their subcellular localisation. To establish that the GFP-fusion proteins were unique to plasmodesmata, immunogold labelling was performed, using antibodies against the GFP moiety of the fusion protein. An example of a plasmodesmal-specific protein is shown in figure 2. A major goal of the cell-to-cell communication programme is to study the structure and function of plasmodesmata in relation to plant development and defence. The discovery of unique plasmodesmatal proteins therefore represents a breakthrough in this field, and the functions of such proteins are the subject of further intensive study in our laboratory.

Nuclear proteins During the course of screening fluorescent infection sites, we identified over 150 proteins that showed a unique localisation to the nucleus. Many of these proteins were predicted to be nuclear based on

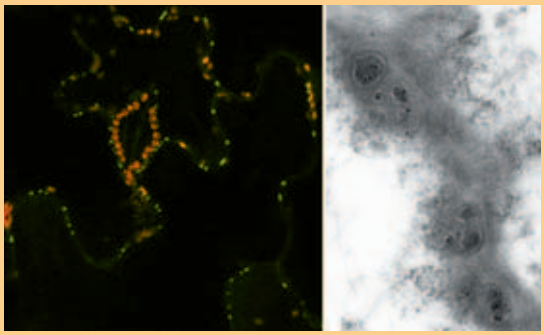

Figure 2. Virus-based screening has identified 12 novel plasmodesmatal proteins. These cDNA-GFP fusions show a strong localisation to regions of cell wall containing plasmodesmata. Immunogold labelling of the GFP moiety has confirmed that the cDNAs are located in the plasmodesmal pore.

sequence homology with known nuclear proteins. However, some unidentified proteins revealed unique structures associated with the nuclear envelope or nucleoplasm. For example, one unknown protein decorated the nuclear envelope, and was possibly associated with nuclear pore complexes (Fig. 3A). Another unidentified protein formed discrete sub-nuclear structures that were arranged helically as hollow spheres in non-DNA containing regions of the nucleus (Fig. 3B).

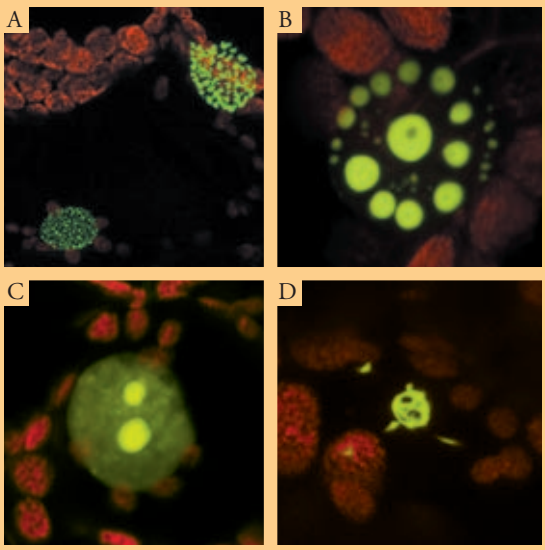

Figure 3 Punctate labelling of the nuclear membrane, possibly associated with nuclear pore complexes. B. An unidentified protein associates with sub-nuclear spherical bodies. C. Localisation of a cDNA-GFP fusion to the nucleolus. D. An unidentified filamentous protein appears to be wrapped around the nucleolus.

Yet other novel proteins were present within the nucleolus (Fig. 3C) or were closely associated with the nucleolus. For example, the protein shown in Fig. 3D appears to be wrapped around the nucleolus in the form of filamentous structures. The further isolation and characterisation of novel plant nuclear proteins forms part of an ongoing collaboration between the Cell-to-Cell Communication and Gene Expression programmes at SCRI.

Chloroplast proteins Although many chloroplast proteins were of predicted location and function, some unknown proteins labelled specific structures within chloroplasts. For example, the filamentous structures shown within chloroplasts in Figure 4A contain a protein of unknown function. Some GFP-fusion proteins labelled the thylakoid membranes while others labelled chloroplast protrusions known as 'stromules', which have been shown to interconnect chloroplasts throughout the cell. One of the stromule-labelling sequences was found to encode an aspartate aminotransferase (Fig.4B)

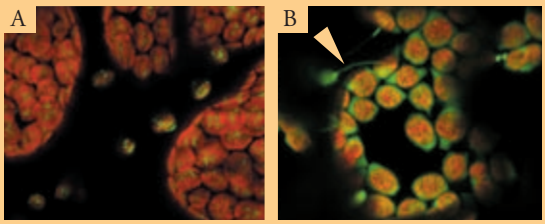

Figure 4. Detection of chloroplast proteins. A. Needle-like structures identified within the chloroplast matrix contain an unidentified protein. B. A cDNA-GFP that labels stromules, hollow extensions of the chloroplast membrane (dart), was shown to have sequence homology with an aspartate aminotransferase.

Cell wall proteins Some of the proteins identified through the cDNA-GFP screen were detected in the apoplast. Interestingly, secreted GFP is not normally stable in the apoplast of tobacco leaves, and undergoes rapid degradation by proteases. Thus, some sequences appeared to confer apoplastic stability to GFP. One of the identified apoplastic sequences encoded for a RALF (*R*apid *AL*kalisation *F*actor) peptide, a recently discovered peptide hormone involved in defence signalling. Other apoplastic proteins remain unidentified. For example, the GFP-fusion protein depicted in Figure 5 was stable in the apoplast and formed discrete protein 'bridges' between adjacent mesophyll cells.

Future utility of viral-based cDNA-GFP libraries The screening of cDNA-GFP libraries is continuing to identify unique proteins with novel subcellular localizations. While many of the sequences identified by the screen encode for proteins of predicted subcellular 'address', such GFP-fusion proteins are nonetheless likely to find considerable utility as fluorescent tags for different organelles within the cell. In the case of sequences with strong homology to unknown proteins from the *Arabidopsis* or *Nicotiana* databases, further characterisation of the function of these proteins is required. In this respect, ascribing a subcellular localisation to such proteins is a useful starting point in identifying their function. The viral-based cDNA-GFP screen is likely to find greatest utility when focussed on a single organelle of interest. For example, the screen has identified putative plasmodesmatal proteins that will form the basis of further functional characterisation of these proteins. During the course of screening the cDNA-GFP libraries, we also noticed that some proteins are expressed specifically in some cell types but not others. For example, the cell-wall wall protein shown in Figure 5 was found extensively between palisade mesophyll cells but not between spongy mesophyll cells. Other proteins were found exlusively in epidermal cells but not mesophyll cells. Thus, the screening procedure outlined here might find utility in identifying proteins that are unique to specific tissues in the plant. In the functional genomics era, ascribing subcellular localisations to proteins will prove to be a useful tool among an armoury of approaches currently being used to assess the functions of unknown genes in plants.

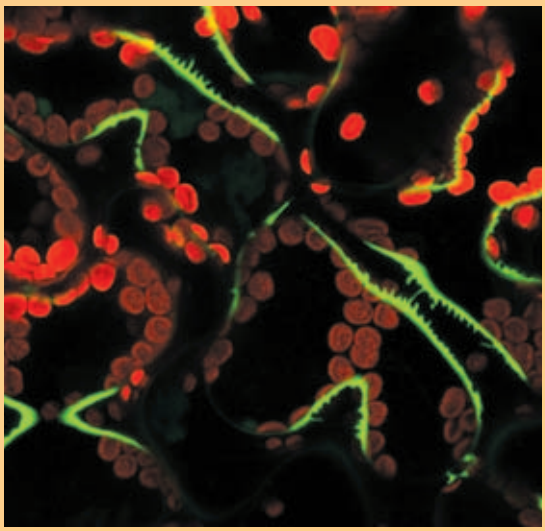

Figure 5. A novel cDNA-GFP fusion labels discrete protein 'bridges' between adjacent palisade mesophyll cells.