Serial Analysis of Gene Expression (SAGE) applied to the study of *Brassica napus* seed development

Christian Obermeier, Bashir Hosseini, Rod Snowdon

Justus Liebig University, Department of Plant Breeding, Research Centre for BioSystems, Land Use and Nutrition, Heinrich-Buff-Ring 26-32, D-35392 Giessen, Germany Email: christian.obermeier@agrar.uni-giessen.de

Abstract

Serial analysis of gene expression (SAGE) is a high-throughput sequencing-based genomic technique that allows identification and quantification of tissue-specific gene expression based on the cloning of short sequence tags (13-26 bp) derived from expressed poly A+ transcripts. A modification of the original protocol, Robust-LongSAGE, which generates 21 bp tags, was used and optimised to enable efficient cloning and transcript identification from the large *Brassica napus* genome. Two libraries were produced and analysed from total RNA extracted from seeds harvested at 23 and 35 days after pollination (DAP). The tag-matching efficiency was restricted by the current limited availability of brassica genomic data and EST annotation quality, which, however, is expected to increase rapidly in the near future. Tags expressed differentially between 23 and 35 DAP that were successfully matched to genes present in databases include genes involved in storage protein accumulation, fatty acid and protein metabolism, photosynthesis, development and secondary compound metabolism. Differentiation between closely related target sequences was possible within the conserved napin storage protein gene family, which was highly up-regulated at 35 DAP. About 7 % of the total counts of all tags derived from genes of the brassica napin gene family matched in antisense orientation within the napin gene coding regions, suggesting an involvement of poly A+ containing antisense RNAs in regulation of napin gene expression during *B. napus* seed development.

Key words: Oilseed rape, *Brassica napus*, seed development, SAGE, expression profiling, napin, antisense RNA

Introduction

Serial analysis of gene expression (SAGE) is a high-throughput technique to simultaneously measure the levels of genes expressed in a given tissues. The technique is based on the excision of short tags from poly A+ RNAs and end-to-end ligation of ditags to form high molecular weight concatemers. This allows cost-effective high-throughput cloning and sequencing of concatemers. Matching of tags to genomic sequences is a critical step in SAGE data analysis. The efficiency of tag-to-gene assignments is dependent on genome structure, transcriptome size and tag length (Saha et al. 2002). SAGE was first used for quantification of gene expression in yeast using 13-15 bp tags (Velculescu et al. 1995). Modifications of the original SAGE protocols producing 21 bp tags (LongSAGE, Saha et al. 2002) and 26 bp tags (SuperSAGE , Matsumura et al. 2003) have been developed to enable more efficient and unambiguous tag-to-gene assignment in higher organisms with more complex transcriptomes. SAGE is commonly used in animal genomics, but to date has been used in only a limited number of plant species and tissues (for examples see references in Ibrahim et al. 2005). The intention of the present study is to adapt the SAGE technique for analysis of global gene expression in *Brassica napus* and other similarly complex polyploid plant genomes. Of particular interest is the usefulness for identification of genes that are differentially expressed during seed development of *B. napus* and might be associated with synthesis of commercially valuable seed compounds.

Material and Methods

Seeds from an inbred line of *B. napus* cv. Express were harvested at two developmental stages (23 and 35 days after pollination; DAP) and stored at -80 C. Total RNA was extracted from seeds using TRIzol reagent according to manufacturer's instructions (Invitrogen, Carlsbad, CA, USA). SAGE libraries containing 21 bp tags were produced from total RNA extracted from seeds. Production of libraries was performed using the I-SAGE Long kit (Invitrogen, Carlsbad, CA, USA) and the manufacturer's protocol was optimised and modified mainly based on the Robust-LongSAGE protocol (Gowda et al. 2004). A series of libraries was produced using different concatemer and *Nla*III enzyme concentrations for partial concatemer digestion before ligation and transformation of *E. coli*. Plasmid DNAs extracted from 100 clones were tested for each library by standard PCR. One library for each time point with the largest average insert sizes and least percentages of empty clones was selected for sequencing. For each library 2,750 clones were randomly selected, plasmid DNA purified and sequenced according to standard protocols by SeqWright (Houston, TX, USA). Data were processed using the SAGE2000 software provided by K.W. Kinzler (John Hopkins University, MD, USA) and the DiscoverySpace 4 software developed by R. Varhol (British Colombia Cancer Agency, BC, Canada).

Results

The quantity and concentrations of PCR-amplified ditag products and enzymes used throughout the LongSAGE cloning procedure were found to be critical for efficient cloning of high molecular weight fragments. The original protocol of the I-SAGE Long kit (Invitrogen) proved to be inappropriate for expression profiling in an organism with a large transcriptome. A number of modifications were introduced into the protocol to optimise the procedure. Modifications were mainly based on Gowda et al. (2004). In particular, cost-effective library production was achieved by increasing the ligation efficiency of vector plasmid with high molecular weight concatemers through previous partial digestion of concatemers in a test series for production of five LongSAGE libraries. Presumably this step may reduce the concentration of circularized and non-clonable concatemer molecules (Crawford et al. 2005). Another major factor reducing cost-efficient sequencing was found to be that depending on pZErO vector plasmid (Invitrogen) and partially digested concatemer concentrations during ligation reactions a large portion of colonies contained empty vector plasmids (20-50 %), raising total sequencing costs. Although the vector plasmid pZErO was designed to prevent cloning of empty vector by a LacZa-ccdB fusion cDNA insert that, when expressed, kills vector-only containing bacteria, we found in agreement with other authors that the pZErO vector can efficiently re-ligate without containing inserts (Angelastro et al. 2002). For this reason 4-5 libraries were produced for every time-point and one library was selected for mass sequencing after screening 100 plasmids each for large insert size averages and low percentage of empty vector plasmid.

To date 53,319 and 14,978 tags were obtained from two libraries produced from seeds harvested at 23 and 35 DAP, respectively (Table 1). Within these two libraries 77 % and 81 % of tags were found to be singletons. To eliminate potential sequencing errors only tags detected two or more times were considered reliable and included in further analysis. From 7,866 unique tags sampled twice or more 268 (4 %) were differentially expressed at $P < 0.01$ in seeds between 23 and 35 DAP.

Table 1. Summary of current LongSAGE data sampling process, count distribution and number of differentially expressed tags between two libraries extracted from time points 23 and 35 days after pollination of developing *B. napus* seeds. Statistical analysis is based on the formula of Audic & Claverie (1997).

23 DAP libary	35 DAP library	Tag counts
53,319	14,978	Current total
31,416 (59 %)	9,562 (64 %)	Different (Unique)
24,165 (77 %)	7,780 (81 %)	Not accepted: 1 count (singletons)
7,251 (23 %)	1,782 (19 %)	Accepted: 2 or more counts:
5,934 (81.8 %)	1,569 (88.0 %)	2-5
1,168 (16.1 %)	183 (10.2 %)	6-10
141 (1.9 %)	26 (1.5 %)	21-99
8 (0.1 %)	4 (0.2 %)	>100
834 of 7,866 (11 %)		Differentially expressed at P<0.05
268 of 7,866 (4 %)		Differentially expressed at P<0.01

From 268 tags differentially expressed at P<0.01 87 % were successfully matched to expressed sequence tags (ESTs) from public databases. These included 75 % matching to ESTs produced from brassica seeds, 2 % matching to ESTs produced from brassica tissues other than seeds, 2 % matching to ESTs produced from Arabidopsis tissues and 8 % matching to ESTs from other plant species. A limited number of these Brassica seed-specific EST hits were found to be linked to well-annotated genes, resulting in only 27 % of differentially expressed tags matching to plant genes from the EMBL database (Fig. 1). Brassica genes up- or down-regulated between 23 DAP and 35 DAP and identified from the EMBL database include genes involved in storage protein accumulation, fatty acid and protein metabolism, photosynthesis, development and secondary compound metabolism.

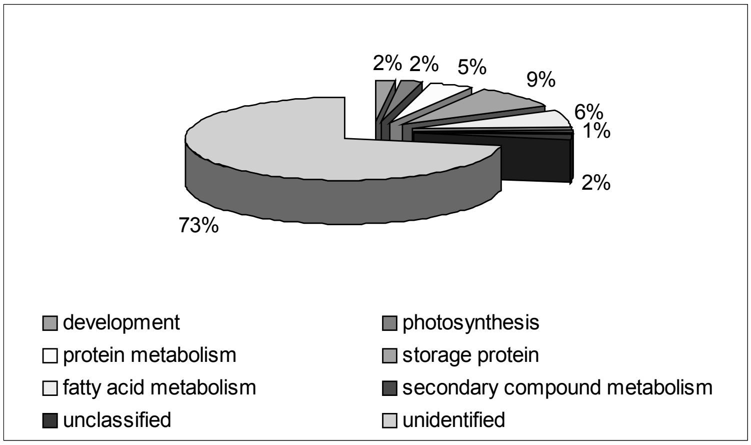

Fig. 1. Tag-to-gene matching success for 268 21-bp tags differentially expressed at P<0.01between 23 and 35 days after pollination in *B. napus* seeds and broad functional categorisation. Tag-to-gene matching was performed using NCBI-blast2 and the genomic DNA database at EMBL.

The largest portion (9 %) of differentially expressed genes is represented by storage protein genes (Fig. 1). At 35 DAP tags derived from genes of the large napin storage protein family were found to be highly abundant and the efficiency of tag-to-gene matching was investigated in more detail for this gene family. Napin transcripts were 110-fold up-regulated at 35 DAP amounting to a total number of 1,104 napin-specific tag molecules. These represent 33 % of all differentially expressed tag molecules at 35 DAP. In total 17 different tag species were matched to known brassica napin transcript and genomic DNA sequences. Several of them were located within conserved regions of napin genes detecting several members of the napin gene family, whereas others were located within highly diverse regions of the napin gene family and thus enabled differentiation between the expression profiles for different members of the napin gene family within the *B. napus* genome. About 7 % of the total counts for all matched napin-specific tags were found in antisense orientation located within coding regions of napin genes.

Discussion

Cost-effective production and sequencing of LongSAGE libraries for an organism with a large transcriptome like *B. napus* was found to require optimisation of cloning procedures. This is in agreement with other reports describing limitations and recent improvements of commercially available SAGE cloning kits (Kenzelmann & Mühlemann 1999, Angelastro et al. 2002, Gowda et al. 2004, Crawford et al. 2005). For high-efficiency cloning and cost-efficient sequencing of SAGE libraries, quality testing of libraries was found to be a crucial step and cloning strategies may still be improved in the future to avoid high ratios of colonies containing empty vectors and to avoid time-consuming PCR or plasmid screening procedures.

Production of two LongSAGE libaries from 23 and 35 DAP revealed a large percentage of singleton tags (about 80 %), indicating that the *B. napus* seed transcriptome at these time-points is much larger than the current sampling size of about 50,000 and 15,000, respectively. The size distribution of transcripts and the high proportion of unique tags showing more than 80 % low-abundant tags (2-5 counts) correlate well with the frequency distribution of SAGE profiles observed for the rice seed transcriptome, reflecting an unusually high diversity of gene expression in highly specialized developing seed tissues (Gibbings et al. 2003). Although the most dramatic biochemical and morphological changes are known to occur during seed maturation between 23 and 35 DAP (Thomas 1993), surprisingly only 4 % of tag species were found to be differentially expressed between these time points, suggesting that post-transcriptional gene regulation may play an important role in seed development.

The tag-to-gene matching success with EST data of 75% for the differentially expressed tags suggests that the *B. napus* seed transcriptome is well covered within public databases. However, the low percentage of successful tag-to-gene matches with genomic data indicates that the availability and annotation quality of brassica EST and genomic data, or the lack of integrated bioinformatic mining tools, are currently major factors limiting the efficiency of tag-to-gene matching in brassicas. Tag-to-gene matching efficiency ratios using EST and genomic data also indicate that SAGE provides a tool for quantification, identification and annotation of previously undescribed brassica genes.

The broad functional categories observed for the identified genes differentially expressed between 23 and 35 DAP are consistent with previous seed expression profiling studies in arabidopsis and brassica using EST sequencing and microarray hybridisation technology (Girke et al. 2002). Transcripts coding for storage protein genes, e.g. for members of the napin storage protein gene family, were found to be highly regulated as has been described before (Thomas 1993). LongSAGE proved to be promising for differentiation of gene family members in complex genomes, as suggested by Saha et al. 2002, when matching specificity was analysed for the napin gene family of *B. napus* (Gehrig et al. 1996). Although it still needs to be demonstrated that the detection of tags matching in antisense orientation for the napin gene family is not due to artifacts during reverse transcription, this result is consistent with SAGE expression profiling in rice seeds (Gibbings et al. 2003). Similarly the transcript complexity and general relevance of antisense RNA within the *Arabidopsis thaliana* transcriptome has been demonstrated by Massively Parallel Signature Sequencing (MPSS, Meyers et al. 2004). Data assembled so far from *B. napus* seed tissues reveal that LongSAGE can be efficiently used for quantitative analysis of gene expression, detection of new transcripts including potential antisense transcripts and dissection of potential regulatory patterns in seed developmental processes.

Conclusions

Construction of SAGE libraries for quantitative global analysis of gene expression in developing seeds of oilseed rape and preliminary data analyses revealed that complex genomes with limited genetic resources like the polyploid *B. napus* genome can be efficiently studied using LongSAGE. The resolution is dependent on transcriptome size, tag size, tag sampling size, tag location, accuracy of bioinformatic tools and publicly available EST and genomic resources, which are constantly increasing for non-model organisms including brassicas.

Acknowledgements

The authors thank the German Research Foundation (DFG, grant no. SN14/7) for funding.

References

Angelastro, J.M., Ryu, E.J., Törõcsik, B., Fiske, B.K., Greene, L.A. (2002). Blue-white selection step enhances the yield of SAGE concatemers. BioTechniques **32**, 484-486.

Audic, S., Claverie, J.M. (1997). The significane of digital gene expression profiles. Genome Res. **7**, 986-995.

Crawford, A.C., White, J., Bundock, P., Cordeiro, G., McIntosh, S., Pacey-Miller, T., Rooke, L., Henry, R.J. (2005). Consistent production of cost-effective

LongSAGE libraries. Plant Molecular Biology Reporter **23**, 139-143.

Gehrig, P.M., Krzyzaniak, A., Barciszewski, J., Biemann, K. (1996). Mass spectrometric amino acid sequencing of a mixture of seed storage proteins (napin) from *Brassica napus*, products of a multigene family. Proceedings of the National Academy of Sciences **93**, 3647-3652.

Gibbings, J.G., Cook, B.P., Dufault, M.R., Madden, S.L.,Khuri, S., Turnbull, C.J., Dunnwell, J.M. (2003). Global transcript analysis of rice leaf and seed using SAGE technology. Plant Biotechnology Journal **1**, 271-285.

Girke, T., Todd, J., Ruuska, S., White, J., Benning, C., Ohlrogge J. (2000). Microarray analysis of developing arabidopsis seeds. Plant Physiology **124**, 1570-1581.

Gowda, M., Jantasuriyarat, C., Dean, R.A., Wang, G.-L. (2004). Robust-LongSAGE (RL-SAGE): A substantially improved LongSAGE method for gene discovery and transcriptome analysis. Plant Physiology **134**, 890-897.

Ibrahim, A.F.M., Hedley, P.E., Cardle, L., Kruger, W., Marshall, D.F., Muehlbauer, G.J., Waugh, R. (2005). A comparative analysis of transcript abundance using SAGE and Affymetrix assays. Functional and Integrative Genomics **5**, 163-174.

Kenzelmann, M., Mühlemann, K. (1999). Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. Nucleic Acids Research **27**, 917-918.

Matsumura, H., Reich, S., Ito, A., Saithoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Krüger, D.H., Terauchi, R. (2003). Gene expression analysis of host-pathogen interactions by SuperSAGE. Proceedings of the National Academy of Sciences **100**, 15718-15723.

Meyers, B.C., Vu, T.H., Tej, S.S., Ghazal, H., Matvienko, M., Agrawal, V., Ning, J., Haudenschild, C.D. (2004). Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. Nature Biotechnology **22**, 1006-1011.

Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W., Velculescu, V.E. (2002). Using the transcriptome to annotate the genome. Nature Biotechnology **20**, 508-512.

Thomas, T.L. (1993). Gene expression during plant embryogenesis and germination: An overview. The Plant Cell **5**,1401-1410.

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W. (1995). Serial analysis of gene expression. Science **270**, 484-487.