# DeepSAGE—digital transcriptomics with high sensitivity, simple experimental protocol and multiplexing of samples

# Kåre L. Nielsen\*, Annabeth Laursen Høgh and Jeppe Emmersen

Department of Biotechnology, Chemistry and Environmental Engineering, Aalborg University, DK-9000 Aalborg, Denmark

Received May 10, 2006; Revised August 23, 2006; Accepted September 15, 2006

#### **ABSTRACT**

Digital transcriptomics with pyrophosphatase based ultra-high throughput DNA sequencing of di-tags provides high sensitivity and cost-effective gene expression profiling. Sample preparation and handling are greatly simplified compared to Serial Analysis of Gene Expression (SAGE). We compare DeepSAGE and LongSAGE data and demonstrate greater power of detection and multiplexing of samples derived from potato. The transcript analysis revealed a great abundance of up-regulated potato transcripts associated with stress in dormant potatoes compared to harvest. Importantly, many transcripts were detected that cannot be matched to known genes, but is likely to be part of the abiotic stress-response in potato.

#### INTRODUCTION

Transcriptomics is essential to monitoring the genomic activation of cells or organisms in response to environmental signals. Global gene expression analysis has been conducted either by hybridization with oligo nucleotide microarrays (1), or by counting of sequence tags. An advantage of microarray analysis is that once the array has been made at a high cost, many measurements can be made at a relatively low cost. However, only known genes can be spotted on the array. In contrast, sequence tag based approaches, like Serial Analysis of Gene Expression (SAGE) (2) and massive parallel signature sequencing (MPSS) (3) can measure the expression of both known and unknown genes. The MPSS technology, however, is too complex to be performed in non-specialized laboratories and very expensive. On the contrary, a SAGE experiment consists of a series of molecular biology manipulation that, in principle, can be carried out in any molecular biology laboratory with access to a 96 capillary DNA sequencer. SAGE relies on the extraction of one 14-21 nt sequence tag from each mRNA. Tags are ligated together, cloned and sequenced. In a typical sequence run of 96 samples  $\sim\!1500$  tags of corresponding mRNAs can be detected. Due to the cost of sequencing, a SAGE study typically encompasses  $50\,000$  tags and provides detailed knowledge of the 2000 most highly expressed genes in the tissue analyzed. In practice, it can be difficult to achieve enough clones of the appropriate insert length (4) to facilitate efficient detection.

Here we describe an experimentally simple method for ditag-based transcript detection, DeepSAGE, similar to the initial steps of LongSAGE (5) in conjunction with emulsion-based amplification and pyrophosphate based ultra-high throughput DNA sequencing (6). DeepSAGE allows the counting of more than 300 000 tags with less effort and cost than a typical LongSAGE study encompassing 50 000 tags. The deep sampling facilitates the measurement of rare transcripts below the detection limit of existing global transcript profiling technologies. Moreover multiple samples can be sequenced in a single run.

## **MATERIALS AND METHODS**

#### DeepSAGE sample preparation

RNA was isolated (7) from field grown potato tubers cv. Kuras at the time of harvest (HAR) and at dormancy after 60 days of storage at 10°C (DOR). Quality of RNA was verified from integrity and intensity of ribosomal RNA following 1% TAE-agarose gel electrophoresis. Fifty microgram of RNA was used to construct LongSAGE ditags as described by Saha et al. (5). Following ligation of linker-tags to form ditags, six 50 µl PCR consisting of 2.5 U Taq polymerase (Ampligon, Copenhagen, Denmark), 0.5 mM deoxynucleotide triphosphates, 1 µl 1:160 dilution of the ligation reaction, 2 μM of 5'-GCCTTGCCAGCCCGCTCAGCAAGCTTCTA-ACGATGTACGT-3' and 2 µM of either 5'-GCCTCC-CTCGCGCCATCAGAAGTGGTGCAGTACAACTAGGCT (HAR) or 5'-GCCTCCCTCGCGCCATCAGACGTGGTGC-AGTACAACTAGGCT (DOR) in 10 mM Tris-HCl, 50 mM KCl, 3 mM MgCl, 1% Triton X-100 were prepared. PCR

<sup>\*</sup>To whom correspondence should be addressed. Tel: +45 9635 8527; Fax: +45 9814 1808; Email: kln@bio.aau.dk

## Tag extraction and data analysis

Tags were extracted from sequence FASTA files containing ditags using the PERL script DeepSAGE\_extract.pl (see Supplementary Data). Linker and poly-A derived tags were removed, but duplicate ditags were not. The tags were mapped to potato tentative contiguous sequences (www.tigr.org) using Sagemap-tsv.pl. (www.bio.aau.dk/en/biotechnology/software\_applications). The resulting tabulator separated value-files were imported into Excel for further analysis. The entire dataset including tags only observed in one of the datasets only was used for the calculation of correlation coefficients. To improve interpretability however, Figures 2 and 3 were displayed in logarithmic scale thereby omitting

tags observed only in one of the datasets. Statistically significant gene expression changes were detected (8) using strict Bonferroni correction.

#### RESULTS AND DISCUSSION

mRNA from two stages of potato tuber development, at harvest (HAR) and dormancy (DOR), were extracted. Following the preparation of LongSAGE ditags (5), 100-400 50 µl PCR are usually pooled to provide enough concatemers for LongSAGE. In the present study, only six 50 µl reactions yielded more than 10 times the material used for a DeepSAGE experiment. Amplification of ditags was carried out using primers containing a sequence primer recognition site, a 3 nt sample identification key (AAG for HAR, ACG for DOR) and a sequence complementary to the linkers used in LongSAGE. Both samples yielded amplification products of 125 bp which were purified by gel electrophoresis. DNA concentration was determined and equimolar amounts of the two samples were pooled. Contrary to LongSAGE these amplified ditags were used directly for sequencing. Preparation of beads carrying sequence templates, clonal amplification in emulsion and DNA sequencing were done according to Margulies et al. (6).

A total of 224 310 sequences were obtained in a single sequence run, which included both forward and reverse sequences (Table 1). The distribution of the length between the two CATG sequences flanking the ditags (Figure 1) was found to be very similar to traditional LongSAGE. A PERL script (DeepSAGE\_extract.pl) was used to extract 314 212 tags of 19 nt (167 367 from forward sequences and 146 845 from reverse sequences). Overall, 70% of these sequences yielded a good ditag sequence. This is comparable to our experience with traditional LongSAGE, where 73% of sequenced clones contained at least one ditag.

It was reported that the pyro-sequencing employed in this study has a somewhat higher error rate than Sanger sequencing, especially in homopolymer regions of four or more (6). Therefore, we inspected our dataset for tags containing homopolymers which were truncated or elongated. Surprisingly, we only found such tags in very low abundance similar to other type of sequencing errors, even though several abundant tags contained homopolymers.

To further address sequence accuracy and the impact on tag based transcriptome analysis, the forward sequences from both runs were sorted by their identification key into 91 580 from the HAR sample and 122 100 from the DOR sample. We determined the sequence error rates using SAGEscreen

Table 1. Summary of sequencing statistics

Library	# Sequences	# Tags exti 1st run	racted 2nd run	Total	% Match to unigenes <sup>a</sup>	% Erro	r rate <sup>b</sup> Ins	Del	Total
DeepSAGE	224 310	314 212	119 835	434 047					
Forward sequences		167 367	46 313	213 680					
DOR		95 427	26 673	122 100	49.5	5.2	2.5	1.7	9.1
HAR		71 940	19 640	91 580	50.5	9.2	2.2	1.3	12.4
LongSAGE DOR	3206	53 688			52.3	15.3	0.7	0.8	16.6

<sup>&</sup>lt;sup>a</sup>Tags were matched to the 38 239 unigenes in Solanum tuberosum Gene Index release 10 (www.tigr.org).

<sup>&</sup>lt;sup>b</sup>Error rates were estimated according to Akmaev and Wang (9).

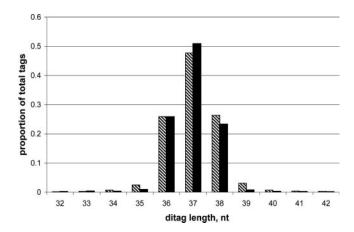
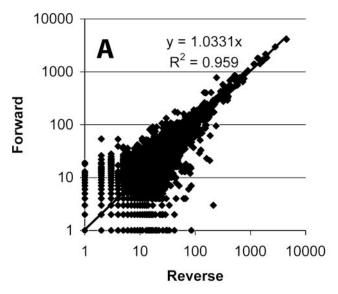


Figure 1. Distribution of ditags length in LongSAGE (solid) and DeepSAGE (hatched).

(9) for both the DeepSAGE datasets and for a LongSAGE DOR library of 53688 tags. Tags observed more than 50 times (87 229 and 141 tags for LongSAGE, DeepSAGE DOR and HAR, respectively). The results are shown in Table 1. Overall estimates of sequence error containing tags in DeepSAGE are in fact lower (9.1–12.4%) than LongSAGE (16.6%). The overall estimates are composed of lower substitution error rate in DeepSAGE compared to LongSAGE (5.2-9.2% versus 15.3% of tags) and a higher insertion (2.2-2.5% versus 0.72%) and deletion rate (1.3-1.7% versus 0.72%)0.8%) in agreement with what was previously found for ultrahigh throughput pyrophosphate sequencing (6). Presumably, the higher sequence accuracy of DeepSAGE is obtained because tag sequences are extracted from nt 33 to approximately 73 (dependent on variation in MmeI cleavage) of DNA sequences, well within the first 90 nt which are determined with the highest accuracy (6). Indeed, correlation analysis of tags extracted from forward and reverse sequences (Figure 2A) indicated good sequence fidelity and reliable tag extraction ( $R^2 = 0.96$ ). Reproducibility was confirmed by performing a second limited run yielding 119835 tags and comparing the two runs ( $R^2 = 0.96$ ) (Figure 2B).

Figure 3 shows a comparison of the numbers of LongSAGE DOR tags versus DeepSAGE DOR tags and DeepSAGE HAR tags, respectively. The distribution of DOR tags was very similar for the Long- and DeepSAGE methods ( $R^2 = 0.96$ ) showing that these measurements of the transcriptome are equivalent. Comparison of the transcriptomes at dormant and harvest were significantly different ( $R^2 = 0.33$ ) as expected. A similar correlation of  $R^2 = 0.35$  was obtained for DeepSAGE DOR versus DeepSAGE HAR (data not shown).

Little is known about the potato tubers adaptation to the abiotic stress imposed by the unnatural environment above ground during storage. Comparing the gene expression of the analyzed potato libraries, 69 genes were up-regulated and 65 genes were down-regulated (P < 0.05 with Bonferroni correction) in DeepSAGE DOR compared to DeepSAGE HAR (Supplementary Table 1). Strikingly, among the 69 up-regulated transcripts, 22 of the 42 transcripts that can be matched to a known sequence are homologues to either chaperones (4), genes involved in the ubiquitin protein



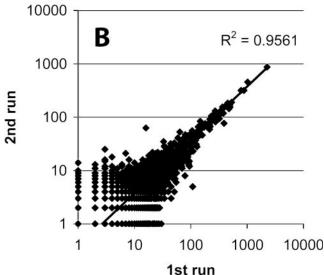
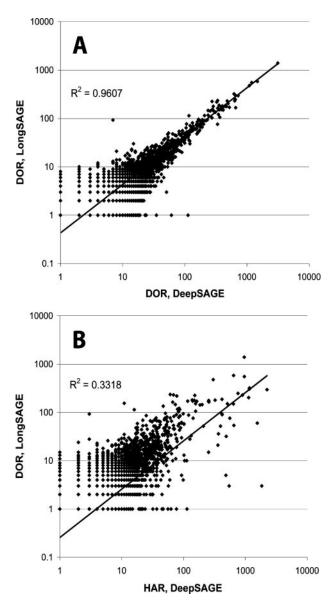


Figure 2. Correlation of tag counts extracted from (A) forward and reverse sequences, respectively. Data sets consisted of 167 159 forward sequences and 199 413 reverse sequences. Using tags observed at least once in both directions only (12 025 tags) the  $R^2 = 0.9611$ . (**B**) Counts extracted from two different sequencing runs. Data sets consisted of 96 427 tags from the first run and 26 673 tags from the second run. Using tags observed at least once in both runs only (6631 tags) the  $R^2 = 0.9609$ .

degradation pathway (5) or suggested to be otherwise stress-related (13) (see Table 2). Intriguingly, among these transcripts are three members of Ca2+ signal transduction pathways: TC112122 (Calmodulin, 4-fold up-regulated), TC126796 (Phosholipase C, 15-fold up-regulated) and TC119057 (Annexin P34, 5-fold up-regulated). Interestingly, Phospholipase C and Annexin P34 transcripts were also observed during EST sequencing in three cDNA libraries derived from abiotically stressed tissue [www.tigr.org and Ref. (10)]. In addition, the cell wall protein (Q40142) which has been shown to be involved in thickening the protective periderm layer was increased (11), consistent with the fact that potatoes thicken their skin during storage, presumably as a response to drought. In stark contrast, only 3 of the 65 down-regulated genes encode chaperones and no other



**Figure 3.** Correlation of LongSAGE and DeepSAGE DOR tags (**A**) and DeepSAGE HAR (**B**). Data sets consisted of 51 918 LongSAGE tags, 122 100 DeepSAGE DOR tags and 91 580 DeepSAGE HAR tags. The most abundant DOR tag was encountered 1397 in LongSAGE and 3145 in DeepSAGE. The least abundant tags were seen once in all data sets. Using tags observed at least once in both libraries only (8567 tags) the  $R^2$  for the comparison of DOR Deep- and LongSAGE increase to 0.9694.

stress-related transcripts were identified. It seems likely, that several of the unknown up-regulated transcripts are additional parts of the potato tuber's response to the abiotic stress induced by storage. 38 of the 65 down-regulated genes can be matched to a potato tentative consensus sequence (TC) and of these more than half (25) match storage proteins, such as patatins, metallocarboxypeptidase inhibitors or Kunitz type protease inhibitors. This indicates that loading of storage protein loading into the potato tuber, at least to some extent can take place up to the point of harvest, but is completely shut down after the tuber has been taken out of the ground. Using the LongSAGE DOR tags (see Supplementary Table 2) only 97 of these were identified as regulated by comparison to

DeepSAGE HAR. Importantly, 26 additional 'false positive' transcripts, which fail to meet the statistical significance criteria using the larger dataset were deemed differentially regulated using the smaller dataset.

Typically 225 000 sequences are obtained in a single experiment, and can be divided in forward and reverse sequences (as in this study), or be generated exclusively from one end only. In DeepSAGE, each sequence contains a ditag. Therefore, at the success rate of this study of 70%, 225 000 sequences represent 315 000 tags determined in a single sequence run. This might be compared to microarray transcriptomics, as Lu *et al.* (12) have estimated that the sensitivity towards detecting rare transcripts in an Affymetrix gene chip experiment is comparable to a SAGE study of 120 000 tags. Therefore, three samples can be multiplexed in a single run to generate this sensitivity.

To estimate the usefulness of the increased sensitivity we analyzed the expression of transcription factors, a class of genes known to be expressed at very low abundance. Mapping the DOR derived tags to the ∼190 000 tentative contiguous sequences of potato (www.tigr.org), only 67 different gene matches were identified as homologues of transcription factors among LongSAGE tags (see Supplementary Table 3 for details). Of these, 13 were observed >5 times and 8 tags were observed 3, 4 or 5 times. The majority (69%) of the 67 tags were seen only twice (17) or once (29). Because, a large proportion of the very rare tags might be tags generated by sequencing errors (13), it cannot be established with certainty whether the corresponding transcripts are present. Furthermore, reliable expression levels cannot be established for rare tags because of random sampling (13). In comparison, 94 different gene matches were found to putative transcription factors in the DeepSAGE analysis of the DOR sample. Twenty-two tags were observed >5 times, whereas 51 tags (54%), were seen twice (16) or once (36). Interestingly, 21 or almost three times as many tags were encountered 3, 4 or 5 times. This underscores that a deeper sampling indeed will detect more rare transcripts, and provides more reliable gene expression estimates.

Multiplexing of different samples or replicates of the same sample, each tagged with a unique nucleotide identification key is a further possibility of DeepSAGE as we have shown here by co-analyzing potato tuber ditag libraries at dormancy and harvest. Until now the lack of replicates has been a severe drawback of SAGE. Following sequencing, different samples are first sorted according to their identification keys and the tags are counted prior to comparison of gene expression.

The DeepSAGE protocol omits the sample consuming concatenation of ditags, the tedious clone picking and sequence template preparation, which constitute most of the experimental SAGE protocol. As an example, a single person in our laboratory has consistently spent six weeks to generate data from two LongSAGE libraries, including sequencing of concatemers. Using DeepSAGE, the same person has recently generated 19 SAGE libraries from pig mRNA in 2 weeks. Sequence template preparation and analysis was performed in 1 week.

Tag based gene expression profiling methods have been inhibited by the cost of DNA sequencing, despite their advantageous global and digital nature, but will become

Table 2. Stress genes regulated between harvest and dormant potato tubers

Tentative consensus	DeepSAGE DOR	DeepSAGE HAR	LongSAGE DOR	Gene name	References
TC111718	337	50	76	Glycine rich binding protein	(14)
TC119035	140	18	59	ADP-ribosylation factor	(15)
TC112122	122	33	52	Calmodulin	(16)
TC119041	311	47	135	Starch phosphorylase L-1	(17)
TC120303	84	11	35	Universal stress protein	(18)
TC118975	121	23	ND	S-Adenosylmethionine synthase	(19)
CN514071	99	17	36	Glycine rich RNA binding protein	(14)
TC126796	62	4	21	Phospholipase C	(20)
TC125909	1478	646	583	Glycine rich binding protein	(14)
TC125835	137	22	69	Low temp and salt responsive protein	(21)
TC126028	367	157	ND	Auxin-repressed protein	(22)
TC126037	136	37	67	Stress-related protein	à
TC119057	339	67	163	Annexin P34	(23)

<sup>&</sup>lt;sup>a</sup>This entry was associated with stress by the Li and McKersie during direct submission of Q9SW70 in SwissProt.

increasingly cost-effective as the 1000\$ genome is approached. The total cost of DeepSAGE, including labor costs, was reduced at least 10-fold compared to LongSAGE, from 25 cents to  $\sim$ 2.5 cents per tag.

#### **ACKNOWLEDGEMENTS**

The authors wish to thank Dr Michael Egholm, 454-Life Science Corp., CT, USA for help with sequencing and Professor Karen G. Welinder, Aalborg University for the critical reading of this manuscript. The work was supported by the Danish Veterinarian and Agricultural Research Council (23-02-0034) and Technical Research Council (26-00-0141). Funding to pay the Open Access publication charges for this article was provided by the Danish Research Councils.

Conflict of interest statement. None declared.

# REFERENCES

- 1. Lockhart, D.J., Dong, H.L., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., Mittmann, M., Wang, C.W., Kobayashi, M., Horton, H. et al. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol., 14, 1675-1680.
- 2. Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. Science, 270, 484–487.
- 3. Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M. et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat. Biotechnol., 18, 630-634.
- 4. Gowda, M., Jantasuriyarat, C., Dean, R.A. and Wang, G.L. (2004) Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. Plant Physiol., 134, 890-897.
- 5. Saha, S., Sparks, A.B., Rago, C., Akmaev, V., Wang, C.J., Vogelstein, B., Kinzler, K.W. and Velculescu, V.E. (2002) Using the transcriptome to annotate the genome. Nat. Biotechnol., 20, 508-512.
- 6. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, 376-380.
- 7. Scott, D.L., Clark, C.W., Deahl, K.L. and Prakash, C.S. (1998) Isolation of functional RNA from periderm tissue of potato tubers and sweet potato storage roots. Plant Mol. Biol. Rep., 16, 3-8.
- 8. Audic, S. and Claverie, J.M. (1997) The significance of digital gene expression profiles. Genome Res., 7, 986-995.

- 9. Akmaev, V.R. and Wang, C.J. (2004) Correction of sequence-based artifacts in serial analysis of gene expression. Bioinformatics, 20, 1254-1263.
- 10. Rensink, W., Hart, A., Liu, J., Ouyang, S., Zismann, V. and Buell, C.R. (2005) Analyzing the potato abiotic stress transcriptome using expressed sequence tags. Genome, 48, 598-605.
- 11. Domingo, C., Gomez, M.D., Canas, L., Hernandez-Yago, J., Conejero, V. and Vera, P. (1994) A novel extracellular matrix protein from tomato associated with lignified secondary cell walls. Plant Cell, 6,
- 12. Lu,J., Lal,A., Merriman,B., Nelson,S. and Riggins,G. (2004) A comparison of gene expression profiles produced by SAGE, long SAGE, and oligonucleotide chips. Genomics, 84, 631-636.
- 13. Anisimov, S.V. and Sharov, A.A. (2004) Incidence of 'quasi-ditags' in catalogs generated by Serial Analysis of Gene Expression (SAGE). BMC. Bioinformatics, 5, 152.
- 14. Kim, Y.O. and Kang, H. (2006) The role of a zinc finger-containing glycine-rich RNA-binding protein during the cold adaptation process in Arabidopsis thaliana. Plant Cell Physiol., 47, 793-798.
- 15. Sieburth, L.E., Muday, G.K., King, E.J., Benton, G., Kim, S., Metcalf, K.E., Meyers, L., Seamen, E. and Van Norman, J.M. (2006) SCARFACE encodes an ARF-GAP that is required for normal auxin efflux and vein patterning in Arabidopsis. Plant Cell, 18, 1396-1411.
- 16. Sathyanarayanan, P.V. and Poovaiah, B.W. (2004) Decoding Ca2+ signals in plants. CRC Crit Rev. Plant Sci., 23, 1-11.
- 17. Zeeman, S.C., Thorneycroft, D., Schupp, N., Chapple, A., Weck, M., Dunstan, H., Haldimann, P., Bechtold, N., Smith, A.M. and Smith, S.M. (2004) Plastidial alpha-glucan phosphorylase is not required for starch degradation in Arabidopsis leaves but has a role in the tolerance of abiotic stress. Plant Physiol., 135, 849-858.
- 18. Kerk, D., Bulgrien, J., Smith, D.W. and Gribskov, M. (2003) Arabidopsis proteins containing similarity to the universal stress protein domain of bacteria. Plant Physiol., 131, 1209-1219.
- 19. Sanchez-Aguayo, I., Rodriguez-Galan, J.M., Garcia, R., Torreblanca, J. and Pardo, J.M. (2004) Salt stress enhances xylem development and expression of S-adenosyl-L-methionine synthase in lignifying tissues of tomato plants. Planta, 220, 278-285.
- 20. Lin, W.H., Ye, R., Ma, H., Xu, Z.H. and Xue, H.W. (2004) DNA chip-based expression profile analysis indicates involvement of the phosphatidylinositol signaling pathway in multiple plant responses to hormone and abiotic treatments. Cell Res., 14, 34-45.
- 21. Nakane, E., Kawakita, K., Doke, N. and Yoshioka, H. (2003) Elicitation of primary and secondary metabolism during defense in the potato. J. Gen. Plant Pathol., 69, 378-384.
- 22. Hwang, E.W., Kim, K.A., Park, S.C., Jeong, M.J., Byun, M.O. and Kwon, H.B. (2005) Expression profiles of hot pepper (Capsicum annuum) genes under cold stress conditions. J. Biosci.,
- 23. Clark, G.B., Thompson, G. and Roux, S.J. (2001) Signal transduction mechanisms in plants: an overview. Curr. Sci., 80, 170-177.