Genome Sequencer System Rapid and Accurate Pyrosequencing of Serial Analysis of Gene Expression Ditags

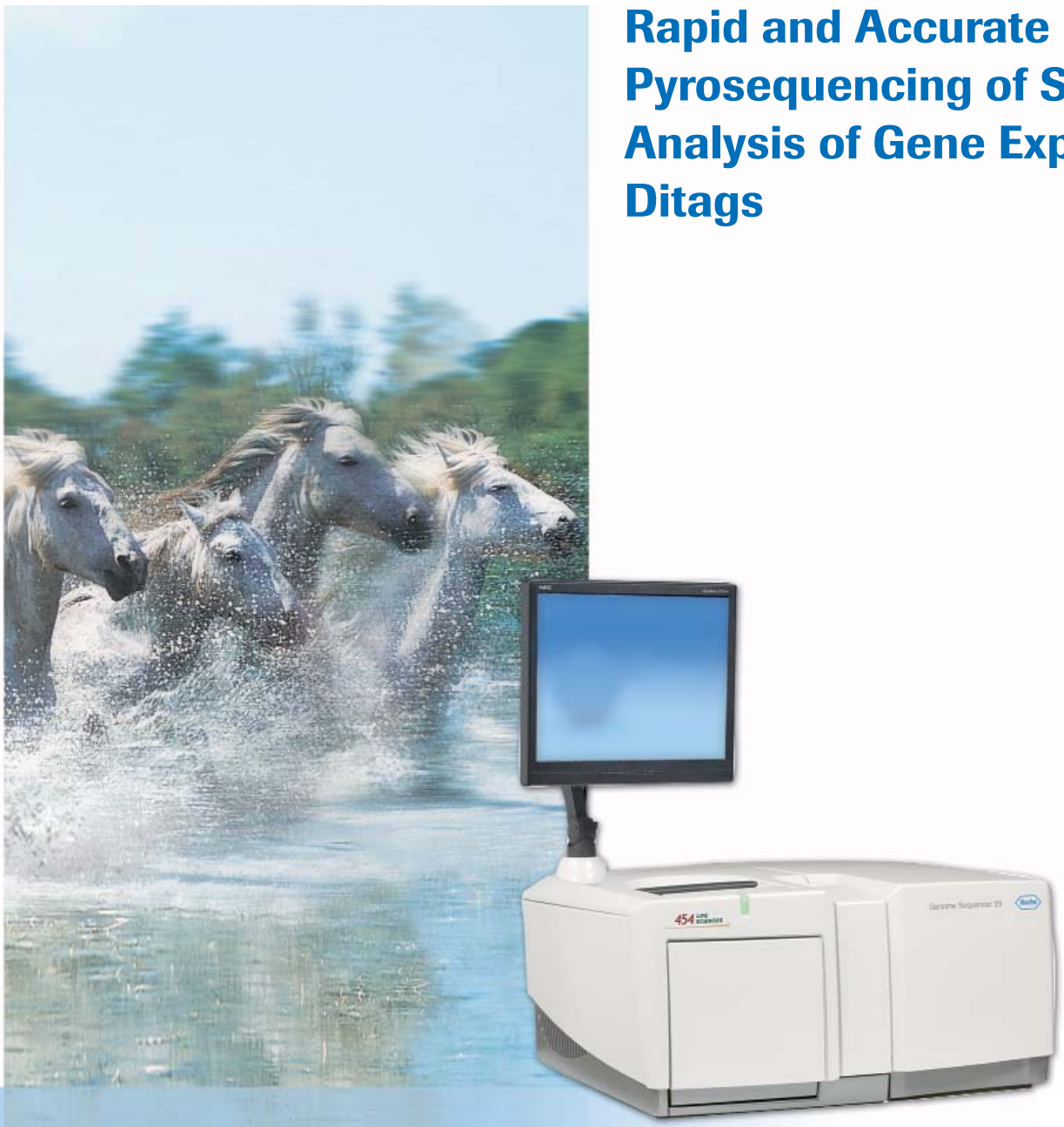
Ronan Quéré, Laurent Manchon, Fabien Pierrat, Ulricke Ludewig, Georg Nesch, Bruno Frey, Thérèse Commes, Jacques Marti, David Piquemal

Roche

Diagnostics

Genome Sequencer System

Application Note No. 4 / February 2007



Rapid and Accurate Pyrosequencing of Serial Analysis of Gene Expression Ditags

www.roche-applied-science.com

454 LIFE SCIENCES

Rapid and Accurate Pyrosequencing of Serial Analysis of Gene Expression Ditags

Ronan Quéré[1,3], Laurent Manchon[1], Fabien Pierrat[1], Ulricke Ludewig[2], Georg Nesch[2], Bruno Frey[2], Thérèse Commes[3], Jacques Marti[3] and David Piquemal[1] Corresponding author: Ronan Quéré, Skuld Tech, 2040 avenue du Père Soulas, 34080 Montpellier, France, Email: info@skuldtech.com

Introduction

A high-throughput sequencing method has been developed and offers an efficient approach to rapidly sequence DNA. This pyrosequencing method is suitable for directly sequencing ditags for Serial Analysis of Gene Expression (SAGE), and is well adapted to reduce both time and cost of SAGE library construction. This strategy avoids the traditional concatemer construction step, which often requires extensive technical knowledge and experience for high-quality fragment construction. Directly sequencing SAGE ditags may therefore allow more laboratories to become involved in SAGE library construction in the future.

Serial Analysis of Gene Expression (SAGE) remains a relevant technique that allows an accurate quantitative and qualitative analysis of cell transcription in a variety of physiological and pathological conditions.[1] Through a series of enzymatic manipulations, the SAGE method reduces cDNA molecules to tags of 14 bp (short SAGE) or 21 bp (long SAGE). Each tag represents one mRNA molecule. Tags are ligated to form concatemers that are cloned and sequenced. Comparing the sequence information from the tags with the GenBank database provides qualitative information about transcribed genes. The frequency of a specific tag within the SAGE tag population correlates with its relative abundance in the cell and gives quantitative information about expressed genes.

The SAGE technique is based on routine molecular biology methods. Nevertheless, only a handful of large genomic centers worldwide have the resources and technical expertise to construct large numbers of SAGE libraries. The most widely used strategy for SAGE library sequencing consists of constructing concatemers by ligation of SAGE tags. Concatemers are cloned into vectors, then transformed in bacterial clones. With this method of SAGE library construction, clones composed of large inserts need to be sequenced using Sanger technology. Since ligation of ditags yields concatemers of various sizes, the efficiency of the SAGE protocol is limited by a small average size of cloned concatemers. When using the original SAGE protocol, both purification and concatenation of SAGE ditags are critical for optimal performance.

A high-throughput sequencing method - the massively parallel picoliter-scale process provided by the Genome Sequencer System - offers an efficient approach to rapidly sequence DNA.[2] To date, the Genome Sequencer System has been successfully utilized in an increasing number of *de novo* sequencing projects, including sequencing the genomes of several bacteria and the mitochondrial genome of an extinct species of mammoth, as well as exploring the sequence diversity present in environmental samples.[2-6]

Here, we report a further significant reduction in time and cost for SAGE library construction through the successful use of this newly available pyrosequencing method. This method reads short lengths of DNA, averaging 80-120 bases. In this application note, we describe how this technology is well adapted for directly sequencing SAGE ditags and is therefore ideally suited for rapid SAGE library construction.

1 Skuld Tech, 2040 avenue du Père Soulas, 34080 Montpellier, France
2 Roche Diagnostics, R&D, Roche Applied Science, 82377 Penzberg, Germany
3 Groupe d'Etude des Transcriptomes, Institut de génétique humaine, UPR CNRS 1142, 141 rue de la Cardonille, 34395 Montpellier, France

Materials and Methods

Materials

Equipment:
Genome Sequencer 20 Instrument (Software 1.0.53)

Reagents:
GS DNA Library Preparation Kit, GS emPCR Kit I (Shotgun), GS 20 Sequencing Kit (40x75 or 70x75), GS PicoTiterPlate Kit (40x75 or 70x75).

A detailed list of all necessary equipment and reagents is provided in the Genome Sequencer User's Manuals and Guides.

Methods

For complete details on how to prepare a sstDNA library from low molecular weight DNA please refer to the GS DNA Library Preparation Kit User's Manual.
Steps therein that are specific to Low Molecular Weight DNA sample preparation are marked in blue.

Preparation of genomic SAGE libraries
Libraries were created using the SAGE construction procedure with the *Sau*3AI anchoring enzyme (GATC)[7, 16]. Ditags were enriched via the polymerase chain reaction (PCR), with primers matching SAGE linkers. Twenty to thirty PCRs were performed and pooled to obtain one microgram of 110 bp ditags, which was directly sequenced using the Genome Sequencer Instrument. Samples of DNA ditags were processed according to the standard DNA Library Preparation procedure with the GS emPCR Kit I (Shotgun) and GS PicoTiterPlate Kit.
Sequence data obtained with the Genome Sequencer Instrument utilizing SAGE ditags was also compared to the traditional Sanger sequencing method utilizing concatemers. Concatemers were constructed according to state-of-the-art methods, as previously described.[1] Concatemers were cloned into vectors and sequenced using BigDye terminator sequencing chemistry on ABI automated sequencers (Applied Biosystems).

Procedure

Bioinformatics

Automatic tag extraction and tag-to-gene mapping were performed with software dedicated to SAGE data mining (Skuld-Tech).[7,8] Tag prediction: virtual tags were extracted from the representative sequences associated with each UniGene cluster (release #191), downloaded from the UniGene FTP site at NCBI (http://www.ncbi.nlm.nih.gov/). Differential gene expression analysis: Biotag software (Skuld-Tech) was used for automatic comparison of expression profiles.

Application on a leukemic cell model

RARα (*Retinoic Acid Receptor Alpha*) belongs to the nuclear receptor superfamily and functions as a ligand-dependent transcription factor, establishing the granulocytic lineage in hematopoiesis. In acute promyelocytic leukemia (APL), the t(15;17) translocation and the subsequent expression of PML (*Promyelocytic Leukemia Protein*) and RARα fusion protein are responsible for blocking at the promyelocytic stage. The PML portion of PML/RARα protein prevents the RARα transcriptional activity. Only a pharmacological dose of *all-trans*-Retinoic Acid (atRA) is able to re-establish the differentiation program. The NB4 cell line is a representative model for the APL differentiation with retinoids. To examine changes in global gene expression mediated by the atRA treatment following the differentiation of APL cells, two SAGE libraries were constructed; one on the proliferative NB4 cells and the second on cells induced to differentiate by a 48-hour exposure to 1 µM atRA.

Results

Increased rapidity of SAGE library construction

While the special design of the pyrosequencing technology relies on small DNA fragment sequences, rather than insert-clone libraries, we developed an application to directly sequence SAGE ditags. In contrast to the original SAGE library procedure based on concatemer preparation, this application could potentially allow one individual to prepare and sequence several SAGE libraries in a few days (**Figure 1**).

Increased SAGE data consolidation

Performance in SAGE library construction was explored, comparing the traditional methodology based on concatemer construction to direct sequencing of SAGE ditags. Change in global gene expression mediated by the atRA treatment was examined by comparing the proliferative leukemic NB4 cells (NB4-UnT) to cells induced to differentiate by a 48-hour exposure to 1 μM atRA (NB4-Diff). To compare outcome efficiency of both protocols, we deliberately started with an equal quantity of SAGE ditags, according to the flow chart procedure described (**Figure 1**). In the original SAGE protocol, the efficiency of library construction is limited by the numerous gel purification steps required, which profoundly increase the molecular loss of ditag fragments. Moreover, since ligation of ditags yields concatemers of various sizes, concatenation is often critical for optimal sequencing performance.

Because the massively parallel picoliter-scale process can be used to directly sequence SAGE ditag fragments, the method allows a researcher to avoid the multiple polyacrylamide gel purification steps, preventing material loss and DNA contamination. At the same time, this strategy avoids concatemer construction, which often requires extensive technical knowledge and experience for high-quality fragment construction. While direct sequencing of SAGE ditags avoids both material loss and laborious technical steps, the massively parallel picoliter-scale process also offers a larger-scale sequencing method (**Figure 2**). Large-scale sequencing of SAGE ditags resulted in the identification of a variety of important regulated markers. The number of regulated genes identified with high statistical values ($P<0.001$) was increased more than 75-fold, while the sequencing level of a SAGE library was increased 6-fold (20,000 versus 120,000 tags). In conclusion, the method is ideally suited for increasing SAGE data consolidation, which is important for identifying many differentially regulated genes when comparing two physiological conditions. We observed a relevant correlation ($R>0.95$) between data obtained with the standard Sanger approach and data obtained by directly sequencing SAGE ditags with the Genome Sequencer System (**Figure 3**). Concerning the percentage of expressed genes covered by both methods, 99.7% of SAGE tags identified with the Sanger method were detected by performing sequencing reactions with the Genome Sequencer System.

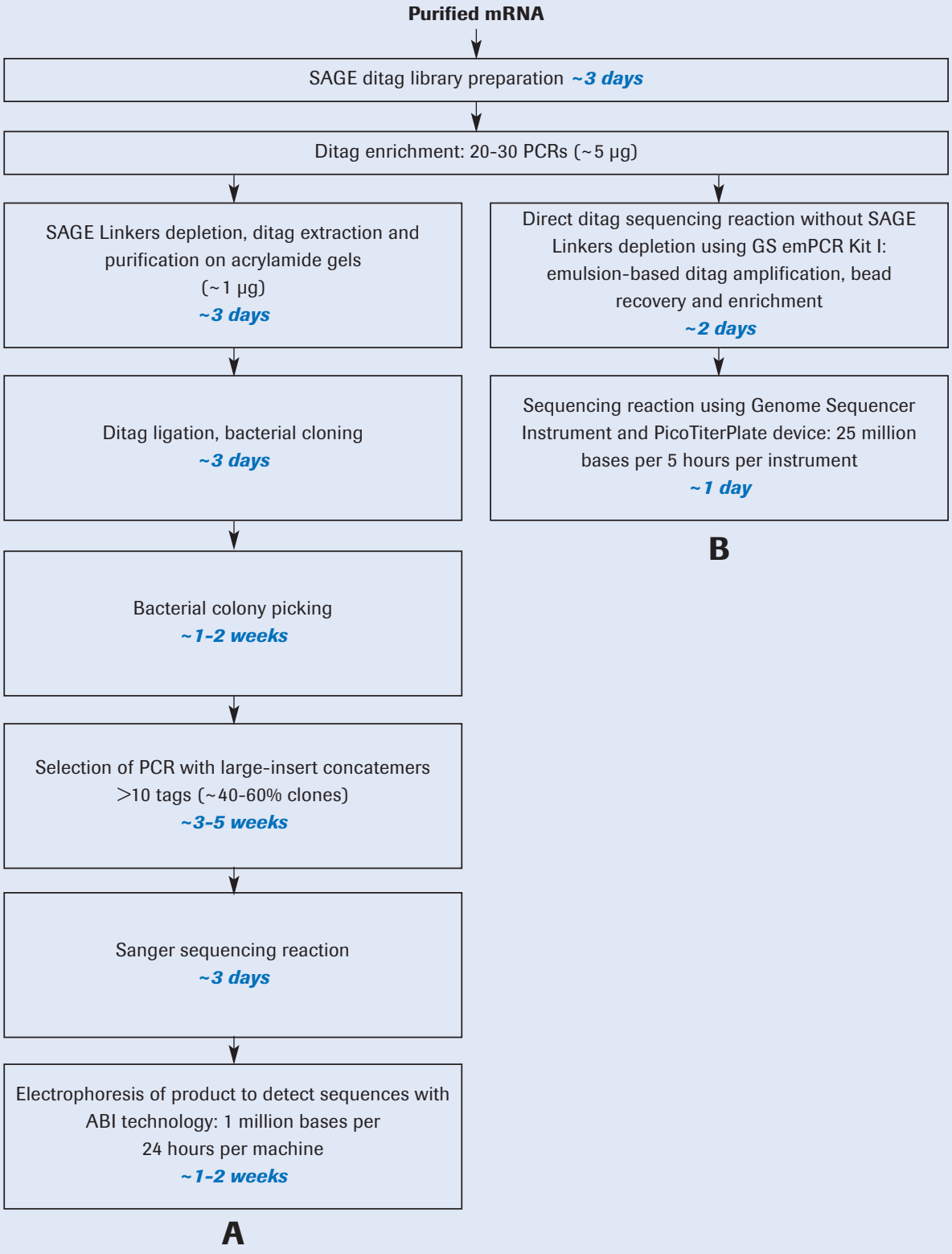


Figure 1: Increasing the speed of sequencing. Flow diagrams for the traditional Sanger DNA sequencing of SAGE concatemers (**A)** and the massively parallel picoliter-scale direct sequencing of SAGE ditags using the Genome Sequencer System. (**B**).

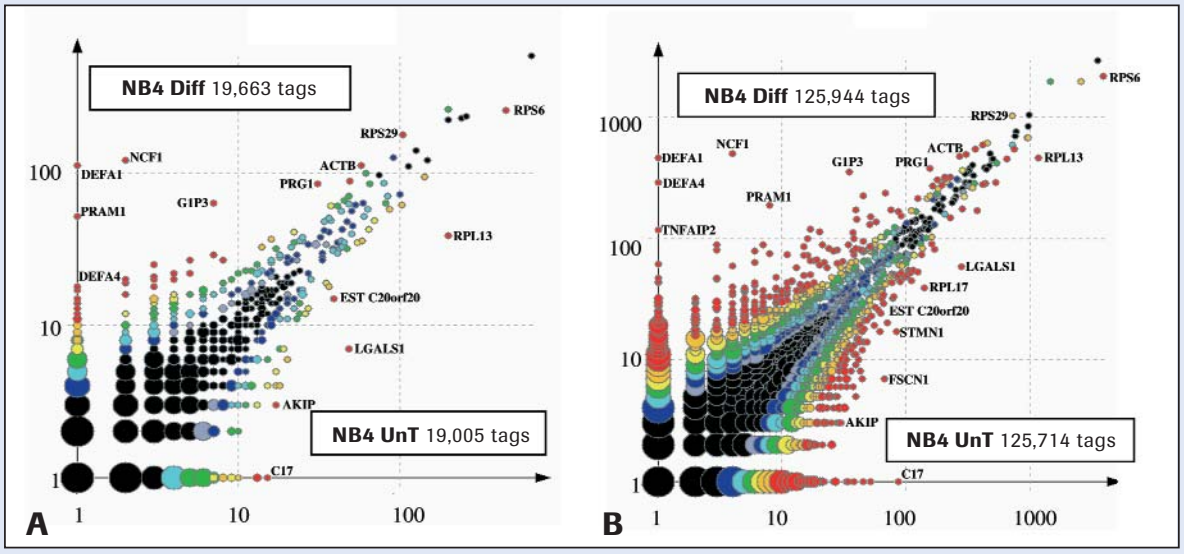


Figure 2: Increased performance of SAGE library construction. Starting with an equal quantity of SAGE ditags, performance of the traditional Sanger DNA sequencing of SAGE concatemers (**A**) was compared to the massively parallel picoliter-scale direct sequencing of SAGE ditags (**B**). To calculate the probability (*P)* that variations observed in paired comparisons occurred by chance, the statistical value of SAGE data (*P value*) was calculated as a function of tag counts. Color plots and lines: $P < 0.001$ (red), $0.001 < P < 0.005$ (orange), $0.005 < P < 0.01$ (yellow), $0.01 < P < 0.03$ (green), $0.03 < P < 0.06$ (azure), $0.06 < P < 0.09$ (blue), $0.09 < P < 0.11$ (grey), and $P > 0.11$ (black).

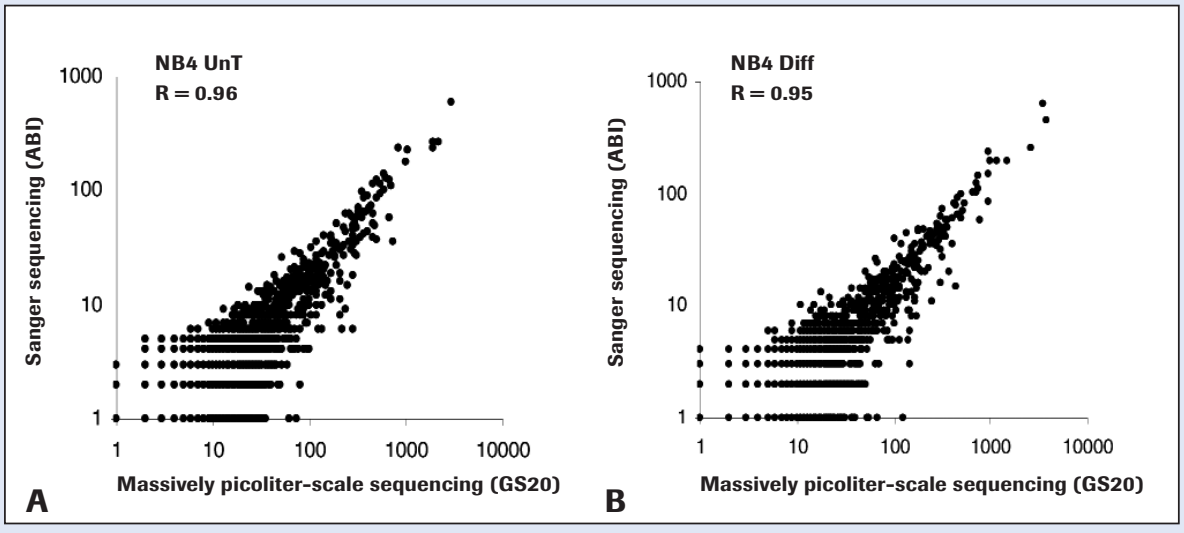


Figure 3: Correlation between tag frequencies obtained with both sequencing methods. Results are shown for the proliferative NB4 cells (**A**) and the differentiated cell libraries (**B**).

Discussion

SAGE is becoming a widely used gene expression profiling method for the study of development, cancer, and other human diseases. Investigators using SAGE rely heavily on the quantitative aspect of this method for cataloging gene expression and comparing multiple SAGE libraries. However, since ligation of ditags yields concatemers of various sizes, the efficiency of the SAGE protocol is limited by a small average size of cloned concatemers. Difficulties in generating high-quality concatemer fragments often interfere with the successful performance of the SAGE technique. To eliminate this problem, multiple modifications of the technique have been proposed that improve the standard SAGE protocol. Improvements concerning additional technical issues that compromise the efficiency of the method include the following: a purification step before digestion of the ditags may increase the yield of digested ditags for concatenation;[9] a heating step introduced before gel electrophoresis may prevent aggregation of small concatemers and migration with large ones;[10] removal of contaminating linker molecules with biotinylated PCR primers may enable removal of the unwanted linkers before concatenation.[11] In addition, the quality of SAGE ditags may be improved by column filtration or subsequent polyacrylamide gel separation.[12-15]

Compared to traditional Sanger-based shotgun sequencing, the Genome Sequencer System implements several novel technologies that enable relatively rapid and inexpensive pyrosequencing on a massive scale. The system's innovative features include an emulsion-based method to amplify and directly sequence random DNA ditag fragments. The tedious technical steps described above can be avoided by using the Genome Sequencer System. Cloning template DNA into bacterial vectors is not necessary, and SAGE library sequencing can be achieved more rapidly, within three days of SAGE library construction. Moreover, SAGE library sequences can be obtained on the Genome Sequencer Instrument in a single five-hour run, with a few days of template preparation of SAGE ditags. In the original SAGE protocol, both purification and concatenation of SAGE ditags are critical for optimal performance. In this application note, we demonstrate that the Genome Sequencer Instrument is ideal for direct and high-level sequencing of SAGE ditags.

While construction of multiple SAGE libraries is complex, many scientists prefer methods based on DNA chip technologies (depending on time and cost of experimentation) to study a variety of physiological and pathological conditions.

The Genome Sequencer System offers many advantages over traditional sequencing methods, and provides an excellent means for efficient, high-throughput direct sequencing of ditags for Serial Analysis of Gene Expression (SAGE), supporting continued growth in genomic science and SAGE library construction.

References

1. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science. 1995;270:484-487.
2. Margulies M, Egholm M, Altman WE, *et al.*, Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437:376-380.
3. Edwards RA, Rodriguez-Brito B, Wegley L, *et al.*, Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics. 2006;7:57.
4. Goldberg SM, Johnson J, Busam D, *et al.*, A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc Natl Acad Sci U S A. 2006;103:11240-11245.
5. Poinar HN, Schwarz C, Qi J, *et al.*, Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science. 2006;311:392-394.
6. Sogin ML, Morrison HG, Huber JA, *et al.*, Microbial diversity in the deep sea and the underexplored "rare biosphere". Proc Natl Acad Sci U S A. 2006;103:12115-12120.
7. Piquemal D, Commes T, Manchon L, *et al.*, Transcriptome analysis of monocytic leukemia cell differentiation. Genomics. 2002;80:361-371.
8. Quere R, Manchon L, Lejeune M, *et al.*, Mining SAGE data allows large-scale, sensitive screening of antisense transcript expression. Nucleic Acids Res. 2004;32:e163.
9. Angelastro JM, Klimaschewski LP, Vitolo OV. Improved NlaIII digestion of PAGE-purified 102 bp ditags by addition of a single purification step in both the SAGE and microSAGE protocols. Nucleic Acids Res. 2000;28:E62.
10. Kenzelmann M, Muhlemann K. Substantially enhanced cloning efficiency of SAGE (Serial Analysis of Gene Expression) by adding a heating step to the original protocol. Nucleic Acids Res. 1999;27:917-918.
11. Powell J. Enhanced concatemer cloning-a modification to the SAGE (Serial Analysis of Gene Expression) technique. Nucleic Acids Res. 1998;26:3445-3446.
12. Damgaard Nielsen M, Millichip M, Josefsen K. High-performance liquid chromatography purification of 26-bp serial analysis of gene expression ditags results in higher yields, longer concatemers, and substantial time savings. Anal Biochem. 2003;313:128-132.
13. Du Z, Scott AD, May GD. Amplification of high-quantity serial analysis of gene expression ditags and improvement of concatemer cloning efficiency. Biotechniques. 2003;35:66-67, 70-62.
14. Lee S, Chen J, Zhou G, Wang SM. Generation of high-quantity and quality tag/ditag cDNAs for SAGE analysis. Biotechniques. 2001;31:348-350, 352-344.
15. Mathupala SP, Sloan AE. "In-gel" purified ditags direct synthesis of highly efficient SAGE Libraries. BMC Genomics. 2002;3:20.
16. Virlon B, Cheval L, Buhler JM, Billon E, Doucet A and Elalouf, JM. Serial microanalysis of renal transcriptomes. Proc Natl Acad Sci U S A 1999; 96:15286-15291.

For more information, visit
www.genome-sequencing.com

04960602001 ① 0107

454 LIFE SCIENCES

Roche Diagnostics

Roche Diagnostics GmbH
Roche Applied Science
68298 Mannheim
Germany
www.roche-applied-science.com