METHODS IN MOLECULAR BIOLOGY™ 387

Serial Analysis of Gene Expression (SAGE)

Methods and Protocols

Edited by
Kåre Lehmann Nielsen

Humana Press

METHODS IN MOLECULAR BIOLOGY™

Serial Analysis of Gene Expression (SAGE)

Methods and Protocols

Edited by

Kåre Lehmann Nielsen

Department of Life Sciences, Aalborg University, Aalborg, Denmark

HUMANA PRESS TOTOWA, NEW JERSEY

Editor
Kåre Lehmann Nielsen
Department of Life Sciences
Aalborg University
Aalborg
Denmark

ISBN: 978-1-58829-676-4 e-ISBN: 978-1-59745-454-4

Library of Congress Control Number: 2007927139

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Serial Analysis of Gene Expression (SAGE)

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

387. **Serial Analysis of Gene Expression (SAGE):** *Methods and Protocols*, edited by *Kåre Lehmann Nielsen, 2008*
386. **Peptide Characterization and Application Protocols**, edited by *Gregg B. Fields, 2007*
385. **Microchip-Based Assay Systems:** *Methods and Applications*, edited by *Pierre N. Floriano, 2007*
384. **Capillary Electrophoresis:** *Methods and Protocols*, edited by *Philippe Schmitt-Kopplin, 2007*
383. **Cancer Genomics and Proteomics:** *Methods and Protocols*, edited by *Paul B. Fisher, 2007*
382. **Microarrays, Second Edition:** *Volume 2, Applications and Data Analysis*, edited by *Jang B. Rampal, 2007*
381. **Microarrays, Second Edition:** *Volume 1, Synthesis Methods*, edited by *Jang B. Rampal, 2007*
380. **Immunological Tolerance:** *Methods and Protocols*, edited by *Paul J. Fairchild, 2007*
379. **Glycovirology Protocols**, edited by *Richard J. Sugrue, 2007*
378. **Monoclonal Antibodies:** *Methods and Protocols*, edited by *Maher Albitar, 2007*
377. **Microarray Data Analysis:** *Methods and Applications*, edited by *Michael J. Korenberg, 2007*
376. **Linkage Disequilibrium and Association Mapping:** *Analysis and Application*, edited by *Andrew R. Collins, 2007*
375. **In Vitro Transcription and Translation Protocols:** *Second Edition*, edited by *Guido Grandi, 2007*
374. **Biological Applications of Quantum Dots**, edited by *Marcel Bruchez and Charles Z. Hotz, 2007*
373. **Pyrosequencing®Protocols**, edited by *Sharon Marsh, 2007*
372. **Mitochondrial Genomics and Proteomics Protocols**, edited by *Dario Leister and Johannes Herrmann, 2007*
371. **Biological Aging:** *Methods and Protocols*, edited by *Trygve O. Tollefsbol, 2007*
370. **Adhesion Protein Protocols**, *Second Edition*, edited by *Amanda S. Coutts, 2007*
369. **Electron Microscopy:** *Methods and Protocols, Second Edition*, edited by *John Kuo, 2007*
368. **Cryopreservation and Freeze-Drying Protocols**, *Second Edition*, edited by *John G. Day and Glyn Stacey, 2007*
367. **Mass Spectrometry Data Analysis in Proteomics**, edited by *Rune Matthiesen, 2007*
366. **Cardiac Gene Expression:** *Methods and Protocols*, edited by *Jun Zhang and Gregg Rokosh, 2007*
365. **Protein Phosphatase Protocols:** edited by *Greg Moorhead, 2007*
364. **Macromolecular Crystallography Protocols:** *Volume 2, Structure Determination*, edited by *Sylvie Doublié, 2007*
363. **Macromolecular Crystallography Protocols:** *Volume 1, Preparation and Crystallization of Macromolecules*, edited by *Sylvie Doublié, 2007*
362. **Circadian Rhythms:** *Methods and Protocols*, edited by *Ezio Rosato, 2007*
361. **Target Discovery and Validation Reviews and Protocols:** *Emerging Molecular Targets and Treatment Options, Volume 2*, edited by *Mouldy Sioud, 2007*
360. **Target Discovery and Validation Reviews and Protocols:** *Emerging Strategies for Targets and Biomarker Discovery, Volume 1*, edited by *Mouldy Sioud, 2007*
359. **Quantitative Proteomics by Mass Spectrometry**, edited by *Salvatore Sechi, 2007*
358. **Metabolomics:** *Methods and Protocols*, edited by *Wolfram Weckwerth, 2007*
357. **Cardiovascular Proteomics:** *Methods and Protocols*, edited by *Fernando Vivanco, 2006*
356. **High-Content Screening:** *A Powerful Approach to Systems Cell Biology and Drug Discovery*, edited by*D. Lansing Taylor, Jeffrey Haskins, and Ken Guiliano, and 2007*
355. **Plant Proteomics:** *Methods and Protocols*, edited by *Hervé Thiellement, Michel Zivy, Catherine Damerval, and Valerie Mechin, 2007*
354. **Plant–Pathogen Interactions:** *Methods and Protocols*, edited by *Pamela C. Ronald, 2006*
353. **Protocols for Nucleic Acid Analysis by Nonradioactive Probes**, *Second Edition*, edited by *Elena Hilario and John Mackay, 2006*
352. **Protein Engineering Protocols**, edited by *Kristian Müller and Katja Arndt, 2006*
351. ***C. elegans:*** *Methods and Applications*, edited by *Kevin Strange, 2006*
350. **Protein Folding Protocols**, edited by *Yawen Bai and Ruth Nussinov 2007*
349. **YAC Protocols**, *Second Edition*, edited by *Alasdair MacKenzie, 2006*
348. **Nuclear Transfer Protocols:** *Cell Reprogramming and Transgenesis*, edited by *Paul J. Verma and Alan Trounson, 2006*
347. **Glycobiology Protocols**, edited by *Inka Brockhausen, 2006*
346. ***Dictyostelium discoideum* Protocols**, edited by *Ludwig Eichinger and Francisco Rivero, 2006*
345. **Diagnostic Bacteriology Protocols**, *Second Edition*, edited by *Louise O'Connor, 2006*
344. **Agrobacterium Protocols**, *Second Edition: Volume 2*, edited by *Kan Wang, 2006*
343. **Agrobacterium Protocols**, *Second Edition: Volume 1*, edited by *Kan Wang, 2006*
342. **MicroRNA Protocols**, *edited by Shao-Yao Ying, 2006*

Preface

All living things carry their genetic information in genes, usually in the form of DNA. The activity of these genes is regulated to meet the requirement by the organism itself or as a response to external abiotic factors such as light, heat, and temperature, but also to biotic factors such as infection by pathogens. Genes are transcribed into mRNAs, which in turn are translated into proteins and catalytically active enzymes. Regulation of this system is primarily obtained by controlling the amount of mRNA that is produced from each gene and the turnover of the corresponding protein. The mRNA population is often referred to as the transcriptome and the protein population as the proteome. The complexity of the system is enormous; all higher organisms, from higher plants to humans, tend to have a similar number of genes, i.e., approx 24,000. In order to understand the genetics that underlie biological change such as development, disease, crop yield, or resistance, it is necessary to perform comparative transcriptomics to understand how the genes are regulated in response to these changes.

Several methods for gene expression profiling exist, such as Northern blotting, Differential Display, EST sequencing, DNA microarrays, and Serial Analysis of Gene Expression (SAGE). The choice of method depends on the need for sensitivity and specificity and whether the methods allow monitoring of genes previously characterized. The dominant method for global gene expression profiling today is DNA microarrays. An array may consist of up to 100,000 unique single-stranded DNA molecules attached to a glass slide in an ordered fashion. 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, so it requires a detailed knowledge of the genetic background.

SAGE, on the other hand, can measure the expression of both known and unknown genes. This method relies on the extraction of a unique 14–21 nt sequence (tag) from each mRNA. These tags are ligated together end to end and sequenced. In a typical sequence run of 96 samples approx 1600 tags and, therefore, mRNAs, can be detected. A SAGE study encompasses 50,000 tags and provides detailed knowledge of the 2000 most highly expressed genes in the tissue analyzed. Another application of SAGE is to discover new genes.

Unknown tags obtained through SAGE analysis of a sample can be efficiently used as gene-specific primers in Rapid Amplification of cDNA Ends (RACE) reactions to generate full-length transcripts that can be cloned and sequenced. In principle, a SAGE experiment consists of a series of molecular biology manipulations that can be carried out in any molecular biology laboratory with access to a 96 capillary DNA sequencer. In practice, however, it has proven difficult to achieve enough clones of the appropriate insert length to facilitate efficient detection, and many laboratories have found SAGE a difficult, time consuming, and expensive method.

The aims of *Serial Analysis of Gene Expression (SAGE): Methods and Protocols* are twofold: (i) To enable users, inexperienced with SAGE and having only limited experience in standard molecular biology techniques, to conduct SAGE experiments by providing detailed, commented, tried-and-tested experimental protocols of SAGE and derived methods from experienced researchers across the world. (ii) To facilitate the analysis and comparison of data from SAGE experiments in a correct and efficient way. To achieve this, this book is divided into two parts. Part 1 discusses the experimental procedures of SAGE and related methods such as aRNA-LongSAGE, SuperSAGE, DeepSAGE, and GMAT, and Part 2 discusses the correct extraction and filtering of tags, the analysis of ditag populations, and the performing of statistically correct comparisons of gene expression profiles.

Tag-based gene expression profiling methods, such as SAGE, have been inhibited by the cost of DNA sequencing despite their advantageous global and digital nature. But sequence-based gene expression profiling approaches will become increasingly cost-effective as we approach the $1000 genome with emerging, much cheaper DNA sequencing technologies. It is the hope that *Serial Analysis of Gene Expression (SAGE): Methods and Protocols* may help many laboratories to their first successful experience with tag-based sequencing methods and obtain comprehensive, useful, and interpretable data.

Kåre Lehmann Nielsen

Contributors

VIATCHESLAV R. AKMAEV • *Bioinformatics, Genzyme Corporation, Framingham, MA*
FRANK BAAS • *Academic Medical Center, Amsterdam, The Netherlands*
LI CAI • *University of California at Berkeley, Berkeley, CA*
JEPPE EMMERSEN • *Department of Biochemistry, Chemistry, and Environmental Engineering, University of Aalborg, Aalborg, Denmark*
XIJIN GE • *Evanston Northwestern Healtcare Research Institute, Evanston, IL*
MALALI GOWDA • *Ohio State University, Columbus, OH*
ANNA M. HEIDENBLUT • *Departmental of Internal Medicine, Knappschaftskrankenhaus, Ruhr-University, Bochum, Germany*
HAIYAN HUANG • *Department of Statistics, University of California at Berkeley, Berkeley, CA*
ANNABETH LAURSEN HØGH • *Department of Biochemistry, Chemistry, and Environmental Engineering, University of Aalborg, Aalborg, Denmark*
ANTOINE VAN KAMPEN • *Academic Medical Center, Amsterdam, The Netherlands*
MARCEL KOOL • *Academic Medical Center, Amsterdam, The Netherlands*
DETLEV H. KRÜGER • *Humboldt University, Berlin, Germany*
HIDEO MATSUMURA • *Iwate Biotechnology Research Center, Iwate, Japan*
KÅRE LEHMANN NIELSEN • *Department of Life Sciences, University of Aalborg, Aalborg, Denmark*
MONIKA REUTER • *Humboldt University, Berlin, Germany*
TAE-YOUNG ROH • *National Institute of Health, Bethesda, MD*
JAN M. RUIJTER • *Academic Medical Center, Amsterdam, The Netherlands*
FRED VAN RUISSEN • *Department of Neurogenetics, Academic Medical Center, Amsterdam, The Netherlands*
GERBEN J. SCHAAF • *Department of Human Genetics, Academic Medical Center, Amsterdam, The Netherlands*
RYOHEI TERAUCHI • *Iwate Biotechnology Research Center, Iwate, Japan*
GUO-LIANG WANG • *Department of Plant Pathology, Ohio State University, Columbus, OH*
SAN MING WANG • *Evanstan Northwestern Healtcare Research Institute, Evanston, IL*

PETER WINTER • *University of Frankfurt, Frankfurt am Main, Germany*
WING HUNG WONG • *University of California at Berkeley, Berkeley, CA*
KEJI ZHAO • *Laboratory of Molecular Immunology, National Institute of Health, Bethesda, MD*

Contents

1

Experimental Procedures

1

SAGE and LongSAGE

Annabeth Laursen Høgh and Kåre Lehmann Nielsen

Summary

Serial analysis of gene expression (SAGE) is a high-throughput method for global gene expression analysis that allows the quantitative and simultaneous analysis of a large number of transcripts. SAGE is a digital method and its sensitivity depends only on the number of tags sequenced. Furthermore, SAGE is a powerful tool for finding novel genes that are expressed under certain conditions or in certain tissues. SAGE has been widely used in fields as diverse as cancer research and the development and study of microorganisms. The SAGE method is a series of routine molecular biology procedure and can, at least in principle, be carried out in any laboratory. However, the number of consecutive steps is quite large and in practice, SAGE has been difficult to carry out on a routine basis.

Key Words: Serial analysis of gene expression; SAGE; LongSAGE; global transcriptome profiling.

1. Introduction

Serial analysis of gene expression (SAGE) is a high-throughput method for global gene expression analysis that was introduced by Velculescu et al. in 1995 ***(1)***. SAGE is based on two principles. First, a short nucleotide sequence (tag) from a unique position contains sufficient information to uniquely identify a transcript. Second, these sequence tags can be linked together to form long serial molecules (concatemers) that can be cloned and sequenced ***(1)***. To obtain the tags, mRNA is synthesized into complementary DNA (cDNA) using biotinylated Oligo(dT) **(Fig. 1)**. Double-stranded cDNA is cleaved with a frequent

From: *Methods in Molecular Biology, vol. 387: Serial Analysis of Gene Expression (SAGE): Methods and Protocols*
Edited by: K. L. Nielsen © Humana Press, Totowa, NJ

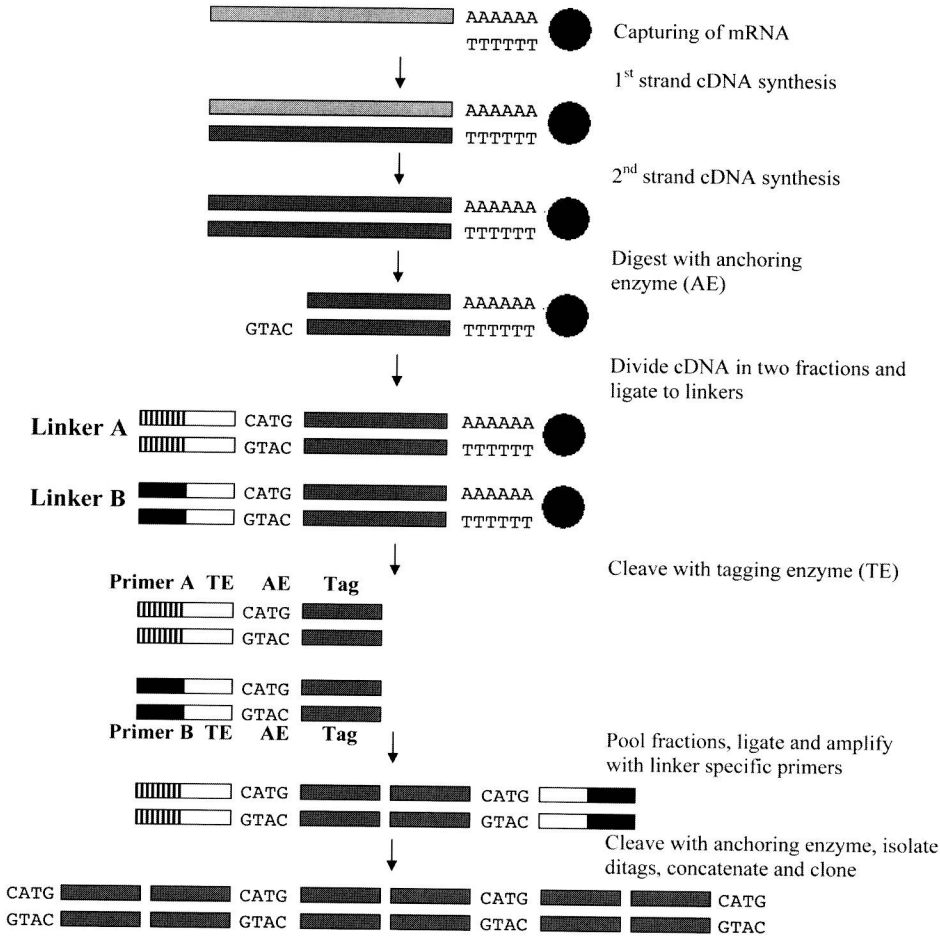

Fig. 1. Schematic overview of serial analysis of gene expression. mRNA is extracted and transcribed into double-stranded complementary (c)DNA on Oligo(dT) streptavidin magnetic beads. cDNA is digested by the anchoring enzyme. The digested cDNA is divided into two fractions, and ligated to different linkers (Linker A and Linker B). The tags are released from the streptavidin magnetic beads by digestion of the tagging enzyme. The linker containing tags are pooled and ligated to form ditags. Following amplification, linkers are removed by digestion of the anchoring enzyme. Ditags are isolated, ligated to form concatemers, cloned, and sequenced.

cutting anchoring enzyme, e.g., *Nla*III, that recognizes the sequence CATG. The 3′-most cDNA fragments are retained using magnetic streptavidin beads. Subsequently, the sample is divided into two fractions, and two different linkers are ligated to the fragments. The linkers contain a restriction site for the tagging enzyme (a type IIS restriction endonuclease, e.g., BsmFI), the anchoring enzyme, and a priming site for PCR. The tagging enzyme cleaves at a defined distance up to 20 bp away from its recognition site, and releases the tags from the magnetic streptavidin beads. The two fractions are pooled, and two sets of linker tag molecules are ligated together to form linker-ditag-linker molecules that can be amplified by PCR using linker-specific primers. Ditags are liberated by digesting with the anchoring enzyme, isolated, and ligated to form concatemers, which are cloned and sequenced ***(1)***. The number of times a particular tag is observed is proportional to the expression level of the corresponding gene. Dinel et al. ***(2)*** have shown that the SAGE method has very good reproducibility, and that the reproducibility, precision, and sensitivity of SAGE are indeed increased by increasing the number of sequenced tags. Furthermore, no *a priori* knowledge of the genes to be identified is required, and the sequence tags can be used to expand sequence information by rapid amplification of cDNA ends (RACE) using cDNA as template ***(3)***.

The original SAGE method was modified into LongSAGE by Saha et al. ***(4)***, generating 21-bp tags instead of 14-bp tags by using another tagging enzyme (*Mme*I instead of BsmFI). The 21-bp tag contains the restriction site of the anchoring enzyme (e.g., CATG) followed by a unique 17-bp tag. In theory, a sequence of 17 bp can distinguish among 17,179,869,184 transcripts (4^{17}) compared to a sequence of 10 bp, which can distinguish among 1,048,576 transcripts (4^{10}). Detailed studies using real sequences show that in practice, SAGE can uniquely identify 94.1 % of *Drosophila melanogaster* genes and 87.6 % of the *Caenorhabditis elegans* genes, whereas LongSAGE uniquely identifies 97.3 % of *D. melanogaster* genes and 93.5 % of the *C. elegans* genes ***(5)***.

2. Materials

2.1. RNA Extraction

1. Liquid nitrogen.
2. Diethylpyrocarbonate (DEPC) water: add 0.75 mL DEPC to 500 mL Milli Q water. Shake well, leave the bottle in a fume cupboard overnight, and autoclave. Store at room temperature.
3. Extraction Buffer: 100 m*M* LiCl, 100 m*M* Tris-HCl pH 8.5, 10 m*M* ethylenediamine tetraacetic acid (EDTA), 1 % sodium dodecyl sulfate (SDS), 15 m*M* dithiothreitol (DTT), in DEPC water.

4. Phenol pH 4.5 (Sigma-Aldrich, St. Louis, MO).
5. Chloroform:isoamyl alcohol (24:1) (Sigma-Aldrich, St. Louis, MO).
6. Phenol:chloroform:isoamyl alcohol (PCI) (25:24:1) (Sigma-Aldrich, St. Louis, MO).

2.2. mRNA Binding to Magnetic Beads

1. Dynabeads Oligo$(dT)_{25}$ (Dynal Biotech Asa, Oslo, Norway).
2. Lysis Buffer: 100 m*M* Tris-HCl, pH 7.5, 500 m*M* LiCl, 10 m*M* EDTA, 1 % lithium dodecyl sulfate, 5 m*M* DTT (Invitrogen, Carlsbad, CA).
3. Wash Buffer A: 10 m*M* Tris-HCl, pH 7.5, 0.15 *M* LiCl, 1 m*M* EDTA, 0.1 % lithium dodecyl sulfate, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
4. Wash Buffer B: 10 m*M* Tris-HCl pH 7.5, 150 m*M* LiCl, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
5. 5X First Strand Buffer: 250 m*M* Tris-HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$ (Invitrogen, Carlsbad, CA).

2.3. cDNA Synthesis

1. DEPC water.
2. dNTP mix, 25 m*M* each (Fermentas, Burlington, Canada).
3. 5X First Strand Buffer: 250 m*M* Tris-HCl, pH 8.3, 375 m*M* KCl, 15 m*M* $MgCl_2$ (Invitrogen, Carlsbad, CA).
4. 0.1 *M* DTT (Invitrogen, Carlsbad, CA).
5. SuperScript™ II Reverse Transcriptase (200 U/μL) (Invitrogen, Carlsbad, CA).
6. 5X Second Strand Buffer: 100 m*M* Tris-HCl, pH 6.9, 450 m*M* KCl, 23 m*M* $MgCl_2$, 0.075 m*M* β-NAD^+, 50 m*M* $(NH_4)_2SO_4$ (Invitrogen, Carlsbad, CA).
7. RNase inhibitor (40 U/μL) (New England Biolabs, Ipswich, MA).
8. *Escherichia coli* DNA ligase (10 U/μL) (Invitrogen, Carlsbad, CA).
9. *E. coli* DNA polymerase (10 U/μL) (Invitrogen, Carlsbad, CA).
10. *E. coli* RNase H (5 U/μL) (Fermentas, Burlington, Canada).
11. 0.5 *M* EDTA (Bie & Berntsen A-S, Rødovre, Denmark).
12. Wash Buffer C: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 1 % SDS, 10 μg/mL glycogen (Fermentas, Burlington, Canada).
13. Wash Buffer D: 5 m*M* Tris-HCl, pH 7.5, 0.5 m*M* EDTA, 1 *M* NaCl, 200 μg/mL bovine serum albumin (BSA) (New England Biolabs, Ipswich, MA).
14. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs, Ipswich, MA).

2.4. Cleavage of cDNA With the Anchoring Enzyme NlaIII

1. LoTE: 3 m*M* Tris-HCl, pH 7.5, 0.2 m*M* EDTA, pH 7.5.
2. 100X BSA (New England Biolabs, Ipswich, MA).
3. 10X NEB Buffer 4: 200 m*M* Tris-acetate, pH 7.9, 100 m*M* magnesium acetate, 500 m*M* potassium acetate, 10 m*M* DTT (New England Biolabs, Ipswich, MA).