Serial Analysis of Gene Expression (SAGE): Experimental Method and Data Analysis

This unit provides a protocol for performing serial analysis of gene expression (SAGE). SAGE involves the generation of short fragments of DNA, or tags, from a defined point in the sequence of all cDNAs in the sample analyzed. This short tag, because of its presence in a defined point in the sequence, is typically sufficient to uniquely identify every transcript in the sample. SAGE allows one to generate a comprehensive profile of gene expression in any sample desired from as few as 100,000 cells or as little as 1 μg total RNA. SAGE also allows an investigator to readily and reliably compare data to those produced by other laboratories, making the SAGE data set increasingly useful as more data are generated and shared.

Serial analysis of gene expression (SAGE), as described in the main method (see Basic Protocol 1), involves the generation of an oligonucleotide library, with each 14-bp SAGE tag representative of a discrete cDNA. Sometimes, the gene that the SAGE tag represents cannot be readily identified. Thus, a second method (see Basic Protocol 2) describes reverse cloning the 3′ end of the cognate cDNA for an unknown SAGE tag. Three additional protocols for verifying cDNA by PCR (see Support Protocol 1), optimizing ditag PCR (see Support Protocol 2), and annealing linkers (see Support Protocol 3), are also given. Finally, protocols for use of publicly available cluster analysis software designed for analysis of SAGE data are described in Basic Protocol 3.

BASIC PROTOCOL 1

MicroSAGE

SAGE library construction involves anchoring mRNA molecules via their poly(A) tails to magnetic beads. (MicroSAGE, which is described here, differs from conventional SAGE in that this anchoring at the 3′ end takes place prior to cDNA synthesis rather than after cDNA synthesis.) cDNA synthesis is then conducted, and the cDNAs are cleaved with *Nla*III to completion. This results in the loss of all cDNA sequence 5′ to the cleavage site, and ensures that only the 3′-most *Nla*III site is exposed at the 3′ end of the cDNA. The cDNA sample is then divided into two equal pools and two sets of linkers (which contain a *Bsm*FI site, PCR primer sites, and modified 3′ bases to prevent ligation to each other) are then added by ligation. *Bsm*FI is a type IIS restriction enzyme, with a cut site 15 bp 3′ of the recognition site. The resulting cDNAs are then digested with *Bsm*FI, which results in the release of the linker, the *Nla*III site, and 10 to 11 bp 3′ of the *Nla*III site. The resulting "tags" are then blunt-ended with the Klenow fragment of DNA polymerase I, and the two separate pools of tags are ligated together via blunt-end ligation to form "ditags." These are then amplified via the PCR primer sites incorporated into the linkers and then recleaved with *Nla*III. These cleaved ditags are purified and ligated together to form concatemers of tags, which are then subcloned into plasmid vectors to create a SAGE library. Individual clones are then sequenced, and analyzed via SAGE analysis software. SAGE software identifies and discards any sets of duplicate ditags (i.e., a given combination of any two individual tags) to control for PCR amplification bias. It can also be used to prepare a tag report, listing all tags and their abundance in a given library, or a tag comparison file, listing the tag abundances in a number of different libraries.

An overview of the microSAGE protocol is shown in Figure 11.7.1.

Contributed by Seth Blackshaw, Brad St. Croix, Kornelia Polyak, Jae Bum Kim, and Li Cai

Current Protocols in Human Genetics (2007) 11.7.1-11.7.38

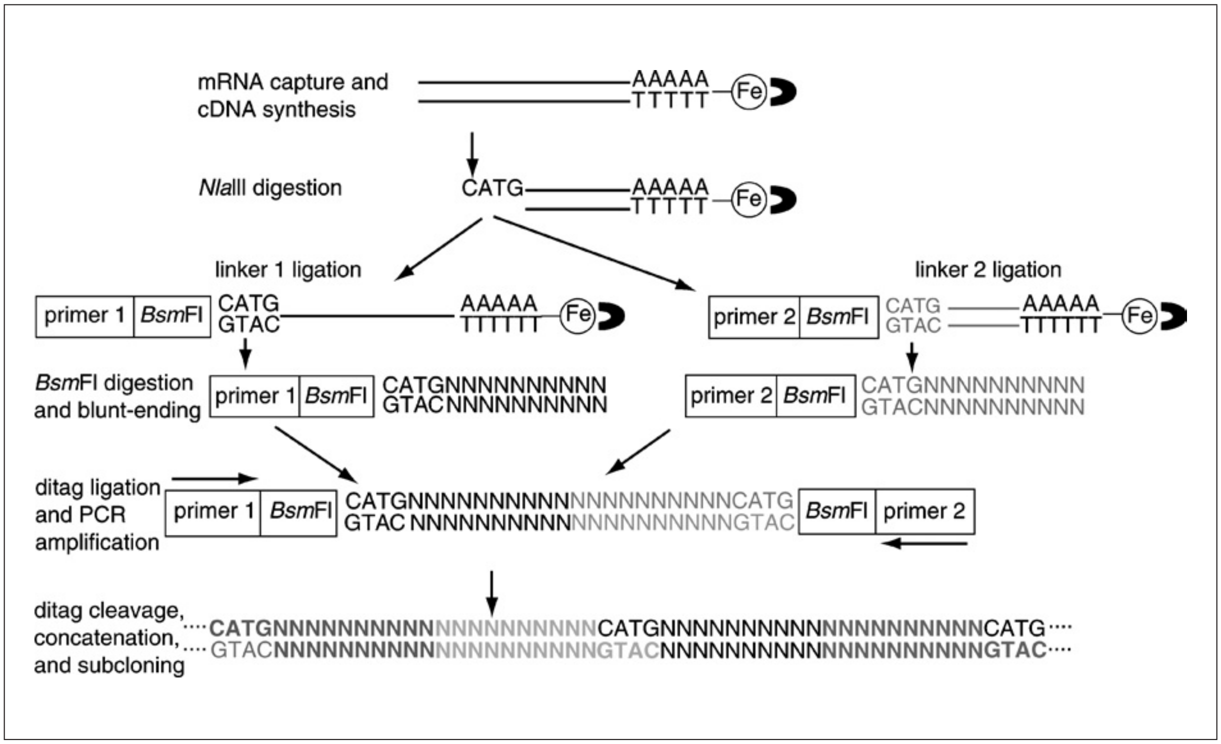

mRNA capture and cDNA synthesis
AAAAA
TTTTT
Fe
*Nla*III digestion
CATG
linker 1 ligation
linker 2 ligation
primer 1
*Bsm*FI
primer 2
CATG
GTAC
*Bsm*FI digestion and blunt-ending
CATGNNNNNNNNNN
GTACNNNNNNNNNN
ditag ligation and PCR amplification
CATGNNNNNNNNNNNNNNNNNNNNCATG
GTAC NNNNNNNNNNNNNNNNNNNNGTAC
ditag cleavage, concatenation, and subcloning
....CATGNNNNNNNNNNNNNNNNNNNNCATGNNNNNNNNNNNNNNNNNNNNCATG....
....GTACNNNNNNNNNNNNNNNNNNNNGTACNNNNNNNNNNNNNNNNNNNNGTAC....

Figure 11.7.1 The steps of a SAGE experiment.

Materials

Dynabeads mRNA DIRECT kit (Dynal Biotech):
- Dynabeads oligo $(dT)_{25}$
- Lysis/binding buffer
- Washing buffer A: add 1 µl 20 mg/ml molecular-biology-grade glycogen (Roche Diagnostics) per milliliter
- Washing buffer B

Cells or tissue of interest

SuperScript Choice System cDNA synthesis kit (Life Technologies):
- 5× first-strand buffer
- DEPC-treated double-distilled water (DEPC ddH_2O)
- 1× first-strand buffer: dilute from 5× stock in DEPC ddH_2O
- 0.1 M DTT
- 10 mM dNTP
- 200 U/µl SuperScript II reverse transcriptase
- 5× second-strand buffer
- 10 U/µl *E. coli* DNA ligase
- 10 U/µl *E. coli* DNA polymerase I
- 2 U/µl *E. coli* RNase H
- 1× and 5× T4 DNA ligase buffer
- 1 U/µl T4 DNA ligase

0.5 M EDTA, pH 8.0 (*APPENDIX 2D*)

1× BW buffer (see recipe)/2× BSA (New England Biolabs)/0.1% (w/v) SDS

1× BW buffer/2× BSA

1× NEBuffer 4 (New England Biolabs)/2× BSA

LoTE buffer (see recipe)

100× BSA (New England Biolabs)

10 U/µl *Nla*III and 10× NEBuffer 4 (New England Biolabs): store at −80°C

1× BW buffer/2× BSA/1% (v/v) Tween 20

Annealed linkers (see Support Protocol 3):

5 U/μl (high-concentration) T4 DNA ligase (Life Technologies)
2 U/μl *Bsm*FI (New England Biolabs)
PC8 (see recipe)
SeeDNA (Amersham Pharmacia Biotech)
3:1 solution of 20 mg/ml glycogen/SeeDNA (optional)
3 M sodium acetate (*APPENDIX 2D*)
70% and 100% ethanol
Klenow fragment of DNA polymerase I and 10× buffer (Amersham Pharmacia Biotech) or Roche Buffer H
3 mM Tris·Cl, pH 7.5 (*APPENDIX 2D*)
10× SAGE PCR amplification buffer (see recipe)
DMSO (Sigma)
PCR primers (see recipe):
350 ng/μl primers 1 and 2
350 ng/μl M13 forward and reverse primers
5 U/μl Platinum *Taq* DNA polymerase (Life Technologies)
20 mg/ml glycogen (Roche Diagnostics)
7.5 M ammonium acetate (Sigma)
Dry ice/methanol bath
5× loading buffer: 50 mM EDTA/50 mM Tris·Cl, pH 8.0 (*APPENDIX 2D*) containing 50% (v/v) glycerol
20% (w/v) polyacrylamide/TBE minigels (Novex)
20-bp DNA ladder (GenSura)
10,000× SYBR Green I (Roche Diagnostics)
1× TBE (*APPENDIX 2D*)
1-kb DNA ladder
pZErO-1 plasmid (Invitrogen)
*Sph*I and NEBuffer 2 (New England Biolabs)
TE buffer, pH 8.0 (*APPENDIX 2D*)
SOC medium (*APPENDIX 2D*)
0.01 ng/μl pUC19 control DNA
DH10B Electromax competent cells, −70°C (Life Technologies)
LB medium (*APPENDIX 2D*; optional)
LB plates with 100 μg/ml ampicillin (*APPENDIX 2D*)
10-cm zeocin-containing low-salt LB plate (see recipe)
10:1 U/μl *Taq*/*Pfu* polymerase (Stratagene)
Exonuclease I (USB)
Shrimp alkaline phosphatase (USB)
50 mM Tris·Cl, pH 8.0 (*APPENDIX 2D*)

0.5-, 1.5-, 2.0-ml RNase-free No-stick siliconized microcentrifuge tubes (Ambion)
Magnetic rack for 1.5-ml microcentrifuge tubes (Dynal Biotech)
Tissue homogenizer (e.g., Polytron PT1200, Brinkmann Instruments)
23-G needles and 1-ml syringes
200-μl aerosol-barrier pipet tips
16° and 65°C water baths, heat blocks, or equivalent
96-well PCR plates
50-ml conical tubes
Tabletop centrifuge with swinging-bucket rotor
Gel-loading tips
UV box and SYBR green or UV filter
0.5-ml microcentrifuge tubes with ∼0.5-mm holes in the bottom: pierce from the inside out with a 21-G needle
Spin-X centrifuge-tube filters (Costar)

Long wavelength UV source
0.1-mm disposable micro-electroporation cuvettes (BioRad)
Bio-Rad gene pulser electroporator or equivalent
15-ml culture tubes

Additional reagents and equipment for determining integrity of cDNA by PCR (see Support Protocol 1), optimizing ditag PCR conditions (see Support Protocol 2), agarose gel electrophoresis (*UNIT 2.7*), ethanol precipitation (*APPENDIX 3C*), polyacrylamide gel electrophoresis (*CPMB UNIT 2.7*) and direct sequencing of PCR products (*CPMB UNIT 15.2*)

NOTE: Prepare Dynabeads, washing solutions, and 5× first-strand mix before thawing and collecting cells.

Prepare mRNA and synthesize cDNA

1. Thoroughly resuspend Dynabeads oligo $(dT)_{25}$, transfer 100 µl to a 1.5-ml RNase-free siliconized No-stick microcentrifuge tube, and place on a magnetic rack. After ~30 sec remove supernatant.

   This volume of beads is much more than needed, but permits easy handling.

   When removing the supernatant, always place the pipet tip at the opposite side of the tube, push the pipet tip to the bottom, and pipet very slowly, so as not to disturb the beads.

2. Resuspend beads in 500 µl lysis/binding buffer by "flicking" the tube or by gently vortexing. Leave beads in buffer until ready to add them to the cell lysate (step 4).

   In this and all subsequent washing steps, add solution to the tube while keeping it on the magnetic rack in order to minimize "drying out" of the beads. Next, close the cap, remove the tube from the magnet, and resuspend the beads. Place back on the magnetic rack for ~30 sec to collect beads at the bottom before removing wash.

3. Lyse 100,000 to 1,000,000 cells (or 2 to 10 mg tissue) in 1 ml lysis/binding buffer in a 2-ml microcentrifuge tube with a tissue homogenizer for 1 min.

   Before using the homogenizer, clean it thoroughly, rinse with 100% ethanol, and pulse in 1 liter DEPC ddH$_2$O.

   If necessary, remove any cellular debris that remains following homogenization by microcentrifuging 1 min at maximum speed.

4. Immediately shear genomic DNA by pressing lysed cells through a 23-G needle attached to a 1-ml syringe into the tube containing prewashed Dynabeads (step 2), from which the buffer has been removed. Incubate 3 to 5 min at room temperature with constant agitation by hand.

   Alternatively, total RNA previously isolated and stored at −80°C may be used. Total RNA (1 to 10 µg in 500 µl of lysis/binding buffer) may be added and incubated 3 to 5 min, room temperature, with constant agitation by hand. It is best to run some of the RNA on a denaturing gel to check for degradation. Visualization of sharp 28S and 18S ribosomal bands should be seen.

5. Place the tube on a magnetic rack for 2 min, then remove the supernatant.

   This supernatant can be used for a genomic DNA prep if desired.

6. Wash beads by pipetting up and down several times with a 200-µl aerosol-barrier pipet tip in the following sequence:

   Twice with 1 ml washing buffer A
   Once with 1 ml washing buffer B
   Four times with 1× first-strand buffer.

   Pipetting the beads is more efficient than flicking the tubes.

7. Resuspend beads in the following first-strand synthesis mix:

   54 μl DEPC ddH_2O
   18 μl 5× first-strand buffer
   9 μl 0.1 M DTT
   4.5 μl 10 mM dNTP.

   Heat tube 2 min at 37°C, then add 3 μl of 200 U/μl SuperScript II reverse transcriptase. Incubate 1 hr at 37°C, mixing beads every 10 min by hand. Terminate reaction by placing tube on ice.

8. Add the following components of the second-strand synthesis to the first-strand reaction in the order shown on ice:

   227 μl ddH_2O, prechilled
   150 μl 5× second-strand buffer
   15 μl 10 mM dNTP
   3 μl 10 U/μl *E. coli* DNA ligase
   12 μl 10 U/μl *E. coli* DNA polymerase I
   3 μl 2 U/μl *E. coli* RNase H.

   Incubate 2 hr at 16°C, mixing beads every 10 min by hand.

9. After incubation, place tubes on ice and terminate reaction by adding 100 μl of 0.5 M EDTA, pH 8.0.

10. Wash beads one time with 0.5 ml of 1× BW buffer/2× BSA/0.1% (w/v) SDS.

    The BSA appears to reduce the stickiness of the beads and improves the efficiency of the washes and the quality of the library. Extra washes with SDS can cause beads to clump severely.

11. Wash beads three times, each in 500 μl of 1× BW buffer/2× BSA. Resuspend beads in 500 μl of 1× BW buffer/2× BSA and heat 20 min at 75°C.

    This heating step is crucial as it inactivates the nuclease activity of PolI.

12. Wash three times in 500 μl of 1× BW buffer/2× BSA. Wash twice with 200 μl of 1× NEBuffer 4/2× BSA, transferring to new tubes after the first wash in NEBuffer 4/BSA and saving 5 μl of the last bead suspension.

13. Check the integrity of cDNA by PCR with primers for genes known to be in the cDNA being used for library construction using the saved 5-μl aliquot (see Support Protocol 1).

Cleave cDNA with anchoring enzyme (NlaIII) and ligate linkers to cDNA

14. Resuspend beads in following mix:

    171 μl LoTE buffer
    4 μl 100× BSA
    20 μl 10× NEBuffer 4
    5 μl 10 U/μl *Nla*III.

    Incubate 1 hr at 37°C.

15. After incubation, place on a magnetic rack ∼30 sec, then wash beads with the following solutions by pipetting up and down several times with a 200-μl aerosol-barrier pipet tip:

Twice with 500 μl 1× BW/2× BSA/1% Tween 20
Four times with 500 μl 1× BW/2× BSA
Twice with 1× T4 DNA ligase buffer.

After final resuspension in ligase buffer, transfer 100 μl of each sample into two new 1.5-ml siliconized microcentrifuge tubes.

16. Remove last wash and resuspend beads with the following:

    5 μl LoTE buffer (both tubes)
    2 μl 5× T4 DNA ligase buffer (both tubes)
    3 μl 2 ng/μl annealed linkers 1A and 1B (only in tube 1)
    3 μl 2 ng/μl annealed linkers 2A and 2B (only in tube 2).

17. Heat tubes 2 min at 50°C then let sit for 5 to 15 min at room temperature. Add 1 μl of 5 U/μl (high-concentration) T4 DNA ligase to each tube and incubate 2 hr at 16°C. Mix beads intermittently.

Release cDNA-tags using tagging enzyme BsmFI

18. After ligation, place on a magnetic rack ∼30 sec, then wash each sample two times with 500 μl of 1× BW/2× BSA/0.1% SDS each, pooling tube 1 and tube 2 together after first wash in order to minimize loss in subsequent steps.

19. Wash four times with 500 μl of 1× BW/2× BSA each and twice with 200 μl of 1× NEBuffer 4/2× BSA (transfer to new tubes after first wash in NEBuffer 4/BSA).

20. Preheat the following mix 2 min at 65°C:

    170 μl LoTE buffer
    20 μl 10× NEBuffer 4
    4 μl 100× BSA
    2 μl 2 U/μl *Bsm*FI.

    Resuspend beads in the mixture and incubate 1 hr at 65°C, mixing intermittently.

21. After incubation, microcentrifuge 2 min at maximum speed, then transfer supernatant to a new 1.5-ml microcentrifuge tube. Wash beads once with 40 μl LoTE buffer, then resuspend to a final volume of 240 μl with LoTE buffer.

    IMPORTANT NOTE: *From this point on, do not use siliconized tubes.*

22. Extract with 240 μl PC8 and ethanol precipitate with SeeDNA using the following procedure:

    a. Add 4 μl SeeDNA. Alternatively, use 4 μl of a 3:1 solution of 20 mg/ml glycogen/SeeDNA mix.
    b. Add 0.1 vol of 3 M sodium acetate (24 μl) and mix briefly.
    c. Add 2 vol of 100% ethanol (480 μl) and vortex briefly.
    d. Incubate 2 min at room temperature.
    e. Microcentrifuge 5 min at maximum speed.
    f. Wash two times with 70% ethanol and microcentrifuge again after last wash. Carefully remove residual liquid with a pipet tip and resuspend pellet in 10 μl LoTE buffer.

SeeDNA is a brightly colored carrier molecule that allows easy visualization and maximal recovery of alcohol-precipitated DNA or RNA. The glycogen/SeeDNA mixture may be used to reduce cost.

One may pause the protocol here and store the pellet overnight at −20°C.

Perform blunt-end digestion on released tags

23. Add the following mix to tags:

30.5 µl ddH_2O
5 µl 10× Klenow buffer (or Roche Buffer H)
2.5 µl 10 mM dNTPs
1 µl 100× BSA
1 µl Klenow fragment of DNA polymerase I.

Incubate 30 min at 37°C then add 190 µl LoTE buffer (240 µl final volume).

24. Extract with an equal volume of PC8 (240 µl). Transfer 200 µl into a ligase "+" tube and the remaining 40 µl into a ligase "−" tube.

25. Ethanol precipitate with 2 µl SeeDNA, 0.1 vol of 3 M sodium acetate, and 2 vol of 100% ethanol. Wash two times with 70% ethanol and centrifuge again after last wash. Carefully remove residual liquid with a pipet tip and air-dry 5 to 10 min. Resuspend pellet in 2 µl LoTE buffer.

Do not overdry because DNA will be lost.

Ligate tags to form ditags

26. Prepare 2× ligase "+" mix as follows:

2.5 µl 3 mM Tris·Cl, pH 7.5
3.0 µl 5× T4 DNA ligase buffer
2.0 µl 5 U/µl (high-concentration) T4 DNA ligase.

Prepare a 2× ligase "−" mix with 4.5 µl of 3 mM Tris·Cl, pH 7.5 and 3.0 µl of 5× T4 DNA ligase buffer. Add 2 µl of appropriate mix to +/− ligase samples and incubate in a thermal cycler overnight (8 to 12 hr) at 16°C.

The sample may dry out in a water bath (in 4°C cold room), thus incubation in a PCR machine/thermal cycler is preferable.

27. After ligation, add 98 µl LoTE buffer, optimize PCR conditions (see Support Protocol 2), and proceed to large-scale PCR amplification.

Samples may be stored >1 year at −20°C.

Perform large-scale PCR amplification of ditags

28. Prepare a reaction mastermix for large-scale PCR (two to three 96-well PCR plates containing 50 µl reaction per well) using the following recipe for one reaction as a guide:

5 µl 10× SAGE PCR amplification buffer
3 µl DMSO
4.0 to 10 µl 10 mM dNTPs
1 µl 350 ng/µl PCR primer 1
1 µl 350 ng/µl PCR primer 2
Adjust volume to 49 µl with ddH_2O
0.7 µl 5 U/µl Platinum *Taq* DNA polymerase.

Aliquot 49 µl of reaction mix to each well, then add 1 µl template at appropriate dilution (see Support Protocol 2).

The authors usually use a 300-reaction PCR premix that is dispensed into 96-well plates at 50-µl per well.

The volume of dNTPs to use is determined through optimization (see Support Protocol 2).

Platinum Taq DNA polymerase is used because it allows for a room-temperature hot start reaction (the Taq DNA polymerase is complexed with an anti-Taq antibody that denatures when heated to 94°C).

29. Carry out the amplifications in a thermal cycler with the following parameters:

1 cycle:	2 min	94°C	(denaturation)
26 to 32 cycles:	30 sec	94°C	(denaturation)
	1 min	55°C	(annealing)
	1 min	70°C	(extension)
1 cycle:	5 min	70°C	(final product extension)

The number of cycles to use is determined through optimization (see Support Protocol 2).

If a thermal cycler with heated lid is not available, oil can be used to prevent evaporation (see CPMB UNIT 15.1).

The ligase "−" sample should be amplified for 35 cycles.

Do not substitute conventional hot-start PCR for use of Platinum Taq DNA polymerase. The authors have found that yields are much lower if this is done. There is no need to refrigerate the PCR mix while setting up the reactions.

Isolate ditags

30. Pool PCR reactions into a 50-ml conical tube, adjusting volume to 11.5 ml with LoTE buffer, then extract with an equal volume of PC8.

31. Precipitate with ethanol as follows:

11.5 ml samples
10 µl SeeDNA
100 µl 20 mg/ml glycogen
5.1 ml 7.5 M ammonium acetate
38.3 ml 100% ethanol.

Place in a dry ice/methanol bath for 15 min. Thaw 2 min at room temperature to fully melt the solution.

32. Vortex briefly and centrifuge 30 min in a tabletop centrifuge with swinging-bucket rotor at ∼3000 × *g* (4000 rpm), room temperature.

33. Wash with 5 ml of 70% ethanol, vortex, and centrifuge an additional 5 min at ∼3000 × *g*, room temperature.

34. Resuspend pellet in 216 µl LoTE buffer and add 54 µl of 5× loading buffer (270 µl total).

35. Using gel-loading pipet tips, load 10 µl sample into each of 27 lanes on each of three prepoured 20% polyacrylamide/TBE minigels. Include 10 µl of a 20-bp ladder on each gel as a marker.

It is critical not to overload the gel wells, as this can lead to linker contamination and poor separation of products.

36. Electrophorese 90 min at 160 V.

    The optimal distance for electrophoresis is ~1 cm above the bottom of the gel. The idea is to obtain maximum separation of the 102- (ditags) and 80-bp bands (linker-linker dimers) without allowing product to get too close to the edge of the gel. Depending on the apparatus and batch of TBE buffer, varying the electrophoresis time might be necessary.

37. Stain 15 min in a foil-wrapped container on a platform shaker using 2 to 5 µl of 10,000× SYBR Green I in 50 ml of 1× TBE buffer. Visualize on a UV box using a SYBR green or UV filter.

    Alternatively, use long-wavelength UV. Amplified ditags should run at 102 bp while a background band (linker-linker dimers) runs at ~80 bp.

38. Cut out only amplified ditags from the gel, and place three cut-out bands in 0.5-ml microcentrifuge tubes (nine tubes total) which have an ~0.5-mm diameter hole in the bottom.

39. Place the 0.5-ml microcentrifuge tubes in 2.0-ml siliconized microcentrifuge tubes and microcentrifuge 4 min at maximum speed.

    This serves to break up the acrylamide gel into small fragments at the bottom of the 2.0-ml microcentrifuge tube.

40. Discard 0.5-ml microcentrifuge tubes. Add 250 µl LoTE buffer and 50 µl of 7.5 M ammonium acetate to each 2.0-ml microcentrifuge tube.

    At this point, the 2.0-ml microcentrifuge tubes can remain overnight at 4°C.

41. Vortex each tube, and incubate 15 min at 65°C. Add 5 µl LoTE buffer to the membrane of each of 18 Spin-X centrifuge-tube filters.

42. Transfer contents of each tube to two Spin-X centrifuge tube filters (i.e., nine tubes transferred to 18 Spin-X centrifuge tube filters). Microcentrifuge each SpinX filter for 5 min at maximum speed. Consolidate sets of two eluates (300 µl total) and transfer to 1.5-ml microcentrifuge tubes.

    Sometimes, purified 102-bp bands do not recut well with NlaIII, which seems to be related to imperfect purification from the gel. If this is a problem, run 300 µl eluate through a Qiaquick gel extraction protocol (Qiagen). Bring the volume of the extract back up to 300 µl to proceed.

43. Ethanol precipitate eluates by adding the following:

    300 µl sample
    0.5 µl SeeDNA
    1.5 µl glycogen
    133 µl 7.5 M ammonium acetate
    1000 µl 100% ethanol.

    Vortex and place in a dry ice/methanol bath for 15 min. Warm 2 min at room temperature until solution has melted, then microcentrifuge 15 min at 4°C.

44. Microcentrifuge 15 min at maximum speed. Wash two times with 75% ethanol. Resuspend each DNA tube in 10 µl LoTE buffer. Pool samples into one microcentrifuge tube (90 µl total).

    The total amount of DNA at this stage should be 10 to 20 µg.

45. Digest PCR products with *Nla*III by adding the following:

    90 μl PCR products in LoTE buffer
    226 μl LoTE buffer
    40 μl 10× NEBuffer 4
    4 μl 100× BSA
    40 μl 10 U/μl *Nla*III.

    Incubate 1 hr at 37°C.

Purify the ditags

46. Extract with an equal volume of PC8. Pool aqueous phases and transfer into 1.5-ml microcentrifuge tubes. Ethanol precipitate in dry ice as follows:

    200 μl sample
    66 μl 7.5 M ammonium acetate
    3 μl SeeDNA
    825 μl 100% ethanol.

    Vortex and place in dry ice/methanol bath for 15 min.

47. Warm 2 min at room temperature until solution has melted, then microcentrifuge 15 min at 4°C.

48. Wash once with cold 75% ethanol, removing ethanol traces with a gel-loading pipet tip. Resuspend pellet in 40 μl LoTE buffer. On ice, add 10 μl of 5× loading buffer (50 μl total).

49. Load this sample into four lanes of a 20% polyacrylamide/TBE gel, load the 20-bp ladder into a separate lane, and run ∼2.5 hr at 160 V. Stain as described in step 37.

    Optimal electrophoresis time may vary somewhat. Be careful not to run the gel too long.

50. Cut out the 24- to 26-bp band from four lanes under long-wavelength UV illumination, and place two cut-out bands in each of two 0.5-ml microcentrifuge tubes which have an ∼0.5-mm diameter hole in the bottom.

51. Microcentrifuge as described in step 39.

52. Discard the 0.5-ml microcentrifuge tubes. Add 250 μl LoTE buffer and 50 μl of 7.5 M ammonium acetate to each of the 2.0-ml microcentrifuge tubes. Vortex the tubes, and incubate 1 hr at 37°C.

    IMPORTANT NOTE: *Do not incubate at 65°C. This will cause the 26-bp ditags to denature. Longer incubations (even overnight) can be performed, but do not appear to result in significantly higher yields.*

53. Use four Spin-X centrifuge-tube filters to isolate eluate as described in step 42. Ethanol precipitate in three tubes (200 μl each) with the following:

    200 μl sample
    66 μl 7.5 M ammonium acetate
    2 μl SeeDNA
    3 μl glycogen
    825 μl 100% ethanol.

    Incubate 10 min in a dry ice/methanol bath, then microcentrifuge 15 min at 4°C.

54. Wash two times with cold 75% ethanol each. Resuspend each DNA sample on ice in 2.5 μl cold LoTE buffer and pool (7.5 μl total).

    It is critical to keep the purified ditags on ice until the ligation buffer is added. High A and T content ditags can denature at room temperature, even in LoTE buffer.

Ligate ditags to form concatemers

55. Mix the following:

    7 µl pooled purified ditags
    2 µl 5× T4 DNA ligase buffer
    1 µl 5 U/µl (high-concentration) T4 DNA ligase.

    Incubate 1 to 3 hr at 16°C.

    Do not ligate overnight, as this will result in long concatemers that are difficult to clone. The authors usually ligate for 2 hr with good results.

    The length of ligation time depends on the quantity and purity of the ditags. Typically, several hundred nanograms of ditags are isolated and produce large concatemers when the ligation reaction is performed for 1 to 3 hr at 16°C (lower quantities or less-pure ditags will require longer ligations).

56. After completing ligation, add 2.5 µl of 5× loading buffer to the ligation reaction. Heat samples 5 min at 65°C and immediately place on ice.

    The heating step melts annealed sticky ends and is critical for obtaining a good yield of clonable concatemers.

57. Separate concatemers on a 10% to 12% polyacrylamide/TBE gel (*CPMB UNIT 2.7*). Load 1-kb DNA marker in first lane, leave one empty lane, and then load the entire concatenated sample into the third well. Run samples 45 min at 200 V.

58. Stain and visualize as described in step 37. Isolate regions of interest.

    Concatemers will form a smear on the gel with a range from ~100 bp to several kilobases.

    The authors usually isolate regions 600 to 1200 bp and 1200 to 2500 bp. These size ranges clone efficiently and yield ample sequencing information.

59. Place each gel piece into 0.5-ml microcentrifuge tubes which have an ~0.5-mm-diameter hole in the bottom.

60. Microcentrifuge as described in step 39.

61. Discard the 0.5-ml microcentrifuge tubes. Add 300 µl LoTE buffer to the gel pieces in the 2.0-ml microcentrifuge tubes. Vortex each tube, and incubate 15 min at 65°C.

    If desired, this incubation can be extended to overnight, but yields are not significantly increased.

    Note that ammonium acetate is not required for high-molecular-weight molecules.

62. Add 5 µl LoTE to the membrane of each four Spin-X microcentrifuge-tube filters. Transfer contents of each tube to two Spin-X microcentrifuge-tube filters (four total). Microcentrifuge each Spin-X tube 5 min at maximum speed.

63. Pool eluates from two Spin-X centrifuge tube filters into one 1.5-ml microcentrifuge tube and ethanol precipitate by adding the following:

    300 µl eluate
    2 µl SeeDNA
    133 µl 7.5 M ammonium acetate
    1000 µl 100% ethanol.

    Glycogen can be substituted for SeeDNA, but the authors obtained better results when only SeeDNA was used.

64. Microcentrifuge 15 min at maximum speed. Wash two times with 70% ethanol and air dry 5 min. Resuspend purified concatemer DNA in 6 µl LoTE buffer.

Ligate the concatemers into vector

65. Digest 1 μg pZErO-1 plasmid with *Sph*I in a total volume of 10 μl by adding the following:

    1 μl pZErO-1 plasmid
    7 μl ddH_2O
    1 μl 10× NEBuffer 2
    1 μl 10 U/μl *Sph*I.

    Incubate 15 to 30 min at 37°C, then heat inactivate 10 min at 65°C. Do not digest >30 min.

    Concatemers can be cloned and sequenced in a vector of choice. The authors currently clone concatemers into a SphI-cleaved pZErO-1.

66. Check for complete digestion on an agarose gel (*UNIT 2.7*). Dilute the cut vector with 90 μl TE buffer, pH 8.0, then extract with equal volume of PC8. Ethanol precipitate (*APPENDIX 3C*), wash two times with 70% ethanol, and resuspend in 40 μl water or TE buffer (∼25 ng/μl of vector).

    The authors recommend using the linearized DNA immediately, but it may be stored for up to 2 weeks at −20°C with decreased ligation efficiency. Ligation efficiency varies beyond 2-week storage. A 2-to-5 fold increase in background is observed upon prolonged storage, due to self-ligation—i.e., no insert.

67. Mix the following ligation reaction and set up a duplicate reaction for a control with no concatemer:

    6 μl purified concatemer (step 64; none in control)
    1.5 μl dH_2O (7.5 μl in control)
    1 μl 25 ng/μl pZErO plasmid cut with *Sph*I
    1 μl 10× T4 DNA ligase buffer
    1.0 μl 1 U/μl T4 DNA ligase.

    Incubate 2 hr at 16°C.

    Consider using 3 μl concatemers and save the rest for backup.

    The manufacturer of pZErO plasmid warns that there is increased background at incubations >1 hr, which may result in breakthrough by spontaneous mutations in the ccdB death gene.

68. Bring sample volume to 200 μl with LoTE buffer. Extract with an equal volume PC8, then ethanol precipitate by mixing the following:

    200 μl sample
    133 μl 7.5 M ammonium acetate
    2 μl SeeDNA
    777 μl 100% ethanol.

69. Wash four times with 70% ethanol. Microcentrifuge briefly at maximum speed, remove 70% ethanol, and air dry 5 min. Resuspend in 10 μl LoTE buffer.

    Excess salt can cause arcing during electroporation and kill the cells.

Transfect DNA by electroporation

70. Place an appropriate number of 0.1-mm microelectroporation cuvettes and 1.5-ml microcentrifuge tubes on ice.

71. Place 1 ml SOC medium in an appropriate number of 15-ml culture tubes at room temperature.

72. Add 1 µl DNA from step 69 to 1.5-ml microcentrifuge tubes on ice. To determine transformation efficiency, add 1 µl of 0.01 ng/µl pUC19 control DNA to a tube labelled "control."

Use 1 µl of the DNA for this transfection. The remainder of the sample is stored at −20°C.

73. Remove DH10B Electromax competent cells from −70°C and thaw on wet ice. When cells are thawed, mix cells by tapping gently.

74. Add 40 µl competent cells to each chilled 1.5-ml microcentrifuge tube containing DNA. Refreeze any unused cells in a dry ice/methanol bath for 5 min before returning to −70°C.

75. Pipet 40 µl of the cell/DNA mixture into a prechilled disposable microelectroporation cuvette (step 70). Perform electroporation with the Bio-Rad gene pulser electroporator at 100 Ω/25 µF/1.8 kV.

76. Transfer electroporated cells into a 15-ml culture tube and immediately add 1.0 ml SOC medium at room temperature. Shake 15 min at 225 rpm, 37°C.

The incubation time is short because, in theory, the postelectroporation incubation period is required for expression of the antibiotic resistance gene, hence increasing transformation efficiency. However, given that the doubling time of the bacteria is ∼20 min, it is possible that the transformed bacteria may double during the incubation period, potentially skewing the library's representation of tags. With 15 min incubation prior to plating, the authors found the transformation efficiency to be 1.0×10^{10} cfu/µg pUC19, respectable when compared with the 1-hr incubation recommended by the manufacturer that resulted in 1.5×10^{10} cfu/µg pUC19.

77. Spread 100 µl of a 1:100 dilution of control cells (pUC19) in SOC or LB medium on LB plates containing 100 µg/ml ampicillin.

78. Plate 1/10 transfected bacteria onto each of ten 10-cm zeocin-containing low-salt LB plates. Incubate and analyze 12 to 16 hr later.

Insert-containing clones should have hundreds to thousands of colonies while no-insert control plates should have zero to tens of colonies.

Save all ten plates for each concatemer ligation reaction since, if insert size appears appropriate, these may be used for sequencing described below.

Check insert size by PCR

79. Prepare a reaction mastermix using the following recipe for one reaction as a guide:

2.5 µl 10× SAGE PCR amplification buffer
1.25 µl DMSO
1.25 µl 10 mM dNTP
0.5 µl 350 ng/µl M13 forward PCR primer
0.5 µl 350 ng/µl M13 reverse PCR primer
18.5 µl ddH_2O
0.5 µl 10:1 U/µl *Taq:Pfu* DNA polymerase.

Pipet 25 µl master mix to wells of 96-well PCR plates.

Any thermostable polymerase can be used (with the appropriate buffer), but the Taq:Pfu mixture works well.

80. For each reaction, use a sterile toothpick or pipet tip to gently touch colony and then dip tip with a twirl into PCR mix.

81. Carry out the amplifications in a thermal cycler with the following parameters:

1 cycle:	2 min	95°C	(denaturation)
25 cycles:	30 sec	95°C	(denaturation)
	1 min	56°C	(annealing)
	2 min	72°C	(extension)
1 cycle:	5 min	70°C	(final product extension).

For Taq DNA polymerase-based PCR amplifications, an extension time of 0.5 to 1.0 min/kb of template amplified is sufficient, but in contrast, Pfu-based PCR amplifications require a minimum extension time of 1 to 2 min/kb of amplified template to achieve similar target synthesis.

82. Analyze on a 1.5% agarose gel at ∼150 V (UNIT 2.7).

For large-scale screening, use multichannel pipettors with an Owl Centipede 50-well horizontal electrophoresis system. The tips of the multichannel pipettors fit into every second well of the 50-slot comb used on the Owl Centipede rigs. Consequently, to maintain a sequential loading order for each 96-well plate, the authors prepare a separate 96-well loading plate with sample loading dye.

The authors typically get 85% to 95% of clones with inserts, of which >95% are >400 bp long. Libraries of this quality can be sequenced directly without gel screening and sorting.

Purify template and sequence amplification product

83. Use 2 µl PCR product (the exact amount will depend on the sequencing protocol and should be optimized) for clean-up using the following:

0.1 µl exonuclease I
0.1 µl shrimp alkaline phosphatase
1.8 µl 50 mM Tris·Cl, pH 8.0.

Add 2 µl clean-up mix to 2 µl DNA.

The exonuclease I degrades unincorporated primers while the alkaline phosphatase degrades unincorporated free nucleotides.

84. Perform reactions in 96-well plates on a thermal cycler, incubating 15 min at 37°C, then 15 min at 80°C. Add ddH_2O to 15 µl. Sequence PCR products directly (CPMB UNIT 15.2).

Use as little as 2 µl of diluted product for the sequencing reaction—optimize according to protocol. The authors run reactions on an ABI 3700 96 capillary machine, though any sequencing system may be used.

85. Download SAGE analysis software from SAGEnet (see Internet Resources) and follow easy-to-use instructions.

SUPPORT PROTOCOL 1

VERIFYING cDNA PRODUCTION BY PCR ANALYSIS

The PCR primers used to test efficiency of the reverse-transcription will depend on the species and tissue type from which the library is constructed. Working in mouse, the authors typically test a ubiquitously-expressed mRNA (RPS17) and a more tissue-restricted mRNA. Design primers to be 18 to 22 bp in length and have a T_m of 55° to 60°C. T_m for the two primers should not differ by more than 1° to 2°C. The PCR product should be 300 to 700 bp in length, with a 5′ end not more than 1 kb from the 3′ end of the mRNA. The following describes the authors' method; however, conditions will have to be optimized for each primer set (see CPMB UNIT 15.1).

Materials *(also see Basic Protocol 1)*

350 ng/μl 5′ and 3′ primers (e.g., Integrated DNA Technology)
Bead suspension (see Basic Protocol 1, step 13)

Additional reagents and equipment for agarose gel electrophoresis (*UNIT 2.7*)

1. Prepare the following PCR mixture:

   5 μl 10× SAGE PCR buffer
   3 μl DMSO
   4 μl 10 mM dNTP mix
   0.5 μl 350 ng/μl 5′ primer
   0.5 μl 350 ng/μl 3′ primer
   31.3 μl ddH_2O
   0.7 μl 5 U/μl *Taq* DNA polymerase
   5 μl bead suspension.

   It is possible to test smaller aliquots of bead suspension depending on the abundance of the template.

2. Perform PCR using the following program:

Initial step:	2 min	95°C	(denaturation)
30 cycles:	30 sec	95°C	(denaturation)
	1 min	53°–58°C	(annealing)
	1 min	72°C	(extension)
Final step:	5 min	70°C	(final extension).

   Annealing temperature should be 2° to 3°C lower than the lowest predicted T_m for the primers.

3. Analyze 5 μl of each PCR product on a 1.5% agarose gel in TAE buffer and visualize bands by ethidium bromide staining (*UNIT 2.7*).

SUPPORT PROTOCOL 2

OPTIMIZING DITAG PCR AMPLIFICATION

The following protocol gives a method for optimizing ditag PCR by varying template concentration, nucleotide concentration, and number of cycles. The optimal template concentration to use is the one which gives a high yield of the 102-bp band with the least concentration of template. A clear plateau in yield should be seen with high concentrations of template. The optimal concentration of nucleotide is simply that which gives the highest yield of the 102-bp band. If none of the PCR reactions give high yields of the 102-bp band, repeat the protocol, but run one tube for 30 cycles and one for 32 cycles. The authors have found that the optimal concentration of nucleotide can vary from batch to batch and supplier to supplier, so repeated optimization may be required.

See Basic Protocol 1 for materials.

1. Prepare serial dilutions of LoTE diluted ditag reaction (see Basic Protocol 1, step 27) at 1:3, 1:9, 1:27, 1:81, and 1:243 in LoTE buffer using 10 μl reaction and 20 μl LoTE buffer (30 μl total) at each step.

2. Prepare the following PCR reaction mixture:

   5 μl 10× SAGE PCR amplification buffer
   3 μl DMSO
   1 μl 350 ng/μl PCR primer 1
   1 μl 350 ng/μl PCR primer 2
   28.3 μl ddH_2O
   0.7 μl 5 U/μl Platinum *Taq* DNA polymerase.

3. Prepare six tubes each containing 1 μl of either stock (see Basic Protocol 1, step 27) or diluted ditag reaction (1:3, 1:9, 1:27, 1:81, or 1:243). In duplicate, add 4, 7, or 10 μl of 10 mM dNTP mix (i.e., prepare two tubes of each dilution and nucleotide concentration pair). Add sufficient double-distilled water to bring the total volume to 11 μl.

4. Perform PCR as described (see Basic Protocol 1, step 29), using 26 cycles for one of the duplicate tubes and 28 for the other.

5. Remove 10 μl from each reaction and run on a prepoured 20% polyacrylamide/TBE gel, using a 20-bp ladder as a marker (10 μl of 1:5 dilution of the marker stock solution; see Basic Protocol 1, steps 35 and 36). Stain gel and visualize as described (see Basic Protocol 1, step 37).

   The amplified ditags should be 102 bp in size. A background band of equal or lower intensity (due to linker-linker dimers) occurs at ∼80 bp. All other background bands should be of substantially lower intensity.

   The ligase "−" samples should not contain any amplified product of the size of the ditags even at 35 cycles.

BASIC PROTOCOL 2

REVERSE CLONING UNKNOWN SAGE TAGS (rSAGE)

SAGE is a technique that allows a generally unbiased evaluation of cellular mRNAs on a genome-wide scale, thus providing a generally more quantitative analysis than subtractive cloning or microarray approaches. Furthermore, the sequencing of 14-bp SAGE tags has a significantly higher throughput than conventional expressed sequence tag (EST) approaches; however, the cDNA that a SAGE tag represents may not be readily identifiable due to the lack of an appropriate anchored cDNA sequence or multiple potential tag to gene matches. This protocol describes an approach, reverse-SAGE (rSAGE), by

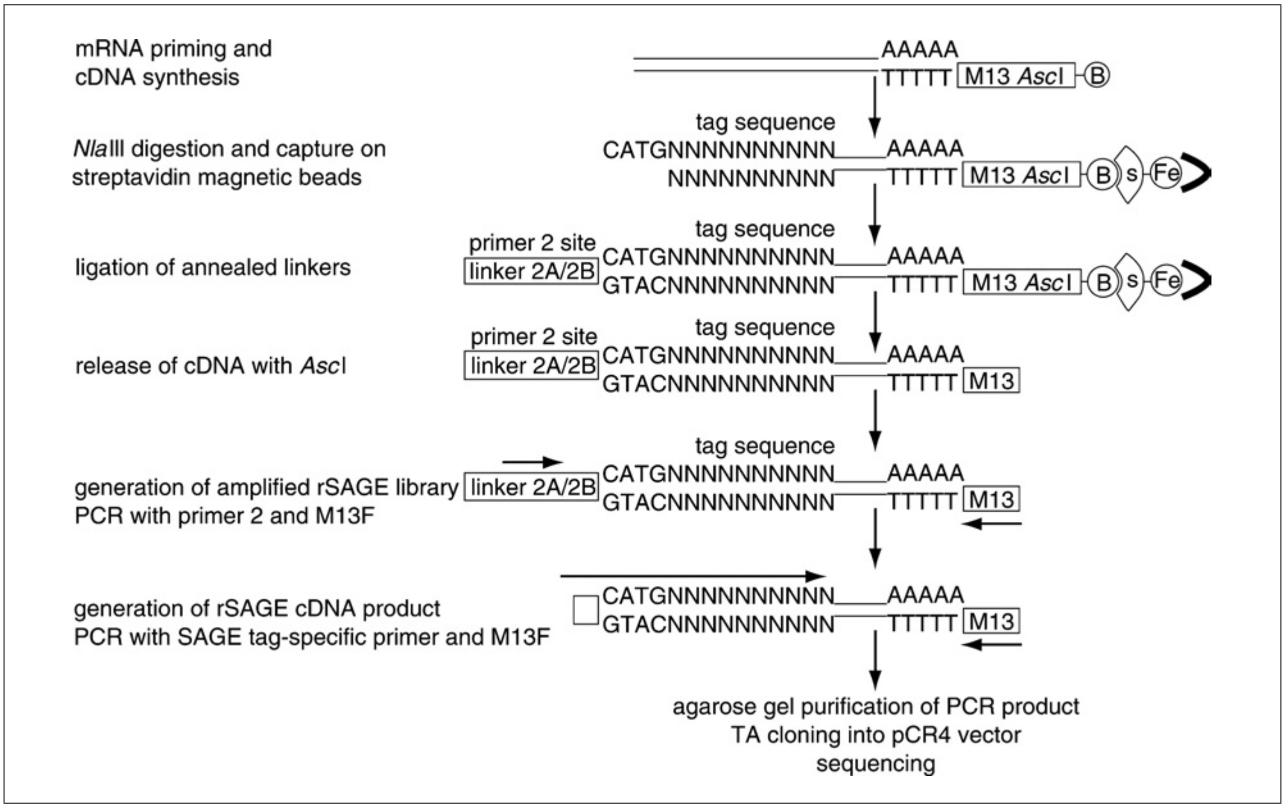


Figure 11.7.2 Steps of an rSAGE experiment.

which the native 3′ sequence can be cloned from cDNA utilizing a variation of the original SAGE protocol and PCR primers based upon sequences in the SAGE tag. The advantage of this protocol is that the unknown gene is cloned using 3′ cDNA fragments that are the most 3′ sequences containing the anchoring enzyme recognition sequence. This approach provides increased specificity of cloning the appropriate cognate cDNA from an anonymous SAGE tag.

Figure 11.7.2 summarizes this procedure. The starting material is total RNA that expresses the target gene, and as a result, the anonymous SAGE tag. Double-stranded cDNA is synthesized by mRNA priming with a biotinylated poly-dT oligonucleotide that also encodes an M13 forward priming site and an *Asc*I restriction site. The anchoring enzyme, *Nla*III, is used to cleave the cDNA and produce 3′ cDNA fragments with *Nla*III cohesive overhangs. These 3′ cDNA fragments are captured onto magnetic streptavidin Dynabeads and subsequently purified. The *Nla*III overhangs are then ligated with annealed linkers, 2A/2B, that encode a priming site for PCR primer 2, which is used for subsequent amplification. The cDNA is then released from the Dynabeads by digestion with *Asc*I restriction endonuclease. The resulting cDNA library is then amplified using PCR primer 2 and M13 forward primer (M13F). A specific rSAGE PCR product is then generated using a SAGE tag–specific primer with M13F. The SAGE tag–specific PCR product is then agarose gel purified and subsequently TA cloned into a sequencing vector.

Materials

SuperScript Choice System cDNA synthesis kit (Life Technologies):
 DEPC ddH_2O
 5× first-strand buffer
 0.1 M DTT
 10 mM dNTP
 200 U/μl SuperScript II reverse transcriptase
 5× second-strand buffer
 10 U/μl *E. coli* DNA ligase
 10 U/μl *E. coli* DNA polymerase I
 2 U/μl *E. coli* RNase H
 5 U/μl T4 DNA polymerase
 1× and 5× T4 DNA ligase buffer
1 μg/μl gel-purified BRS1 primer (see recipe)
0.5 M EDTA, pH 7.5 (APPENDIX 2D)
PC8 (see recipe)
SeeDNA (Amersham Pharmacia Biotech)
7.5 M ammonium acetate (Sigma)
70% and 100% ethanol
LoTE buffer (see recipe)
100× BSA (New England Biolabs)
10 U/μl *Nla*III and 10× NEBuffer 4 (New England Biolabs)
Streptavidin Dynabeads (Dynal)
1× BW buffer (see recipe)
Annealed linkers (see Support Protocol 1)
5 U/μl (high-concentration) T4 DNA ligase (Life Technologies)
1× BW buffer/1× BSA
1× NEBuffer 4/1× BSA
100× BSA
10 U/μl *Asc*I (New England Biolabs)
10× SAGE PCR buffer (see recipe)
DMSO
PCR primers (see recipe):

350 ng/μl M13 forward primer
350 ng/μl primer 2
5 U/μl Platinum Taq DNA polymerase (Life Technologies)
4% to 20% TBE acrylamide gel (Novex)
1-kb ladder
1× SYBR green I (Roche Diagnostics) in TBE buffer (*APPENDIX 2D*)
5 M betaine: prepare monohydrate salt (Sigma) in PCR-grade ddH_2O
SAGE tag–specific primer (see recipe)
Qiaquick gel-extraction kit (Qiagen):
Qiaquick columns
EB Buffer
TOPO TA Cloning Kit with pCR2.1 vector (Invitrogen) *or*
TOPO TA Cloning Kit for Sequencing with pCR4-TOPO vector (Invitrogen)

16°, 50°, and 70°C water baths, heat blocks, or equivalent
1.5-ml No-stick siliconized microcentrifuge tubes (Ambion)
Magnetic rack for 1.5-ml microcentrifuge tubes(Dynal)
1.5-ml nonsiliconized nuclease-free microcentrifuge tubes

Additional reagents and equipment for preparing total RNA (*CPMB UNIT 4.2*) agarose gel electrophoresis (*UNIT 2.7*) and sequencing (*CPMB UNIT 7.4A*)

Synthesize cDNA

1. Prepare total RNA in DEPC ddH_2O using standard methods (e.g., *CPMB UNIT 4.2*).

   Trizol (Sigma) is the preferred method in the authors' laboratory. The same RNA with which the original SAGE library was generated would be ideal (see Basic Protocol 1, steps 3 and 4).

   It is advisable to also generate a control rSAGE library that will not express the genes of interest. As PCR cloning from the rSAGE library might generate more than one clonable band, PCR of a control rSAGE library would allow the researcher to discriminate and identify the likely rSAGE product representing the gene of interest.

2. Add 2 μl of 1 μg/μl gel-purified BRS1 primer to a nonsiliconized 1.5-ml microcentrifuge tube. Add 6 μl total RNA (5 to 10 μg total) and mix.

3. Heat mixture to 70°C for 10 min and quick chill on ice. Microcentrifuge briefly at room temperature. Prepare first-strand-synthesis mix as shown below:

   8 μl BRS1 primer/RNA
   4 μl 5× first-strand buffer
   2 μl of 0.1 M DTT
   1 μl of 10 mM dNTP.

4. Mix gently by vortexing and microcentrifuge briefly at room temperature. Incubate 2 min at 37°C, then add 5 μl of 200 U/μl SuperScript II reverse transcriptase and mix well. Incubate an additional 1 hr at 37°C.

5. After incubation, place tube on ice to terminate the reaction. Add the components of the second-strand-synthesis mixture to the first-strand reaction on ice in the order shown:

   93 μl DEPC ddH_2O, 4°C
   30 μl 5× second-strand buffer
   3 μl 10 mM dNTP
   1 μl 10 U/μl *E. coli* DNA ligase
   4 μl 10 U/μl *E. coli* DNA polymerase I
   1 μl 2 U/μl *E. coli* RNase H.

   Vortex gently to mix.

6. Incubate 2 hr at 16°C. Intermittently mix by gentle flicking. Add 2 μl 5 U/μl T4 DNA polymerase and incubate 5 min at 16°C. Place tubes on ice and terminate reaction by adding 10 μl of 0.5 M EDTA, pH 7.5.

   T4 DNA polymerase is used in the reverse-SAGE protocol to fill in 5′ overhangs generated after second-strand synthesis.

7. Add 150 μl PC8 and vortex thoroughly. Microcentrifuge 5 min at maximum speed, room temperature. Remove and save aqueous layer (∼150 μl).

   Unlike microSAGE, the reverse-SAGE protocol synthesizes DNA onto unbound biotinylated oligonucleotides, making purification (i.e., phenol-chloroform extraction followed by ethanol precipitation) easier. As a result, the heat denaturation and multiple wash steps in the SAGE protocol are unnecessary.

8. Ethanol precipitate aqueous layer in a fresh standard 1.5-ml microcentrifuge tube by adding the following reagents:

   2 μl SeeDNA
   70 μl 7.5 M ammonium acetate
   500 μl 100% ethanol.

   Vortex thoroughly, then microcentrifuge 20 min at maximum speed, 4°C. Wash pellet in 70% ethanol.

9. Resuspend in 20 μl LoTE buffer.

   Samples may be stored at 4°C up to a week or frozen at −20°C for months. However, it is best to leave at 4°C overnight and resume the protocol the following day.

Cleave cDNA with anchoring enzyme (NlaIII) and ligate linkers

10. Cleave cDNA with the anchoring enzyme (*Nla*III) using the following mixture:

    20 μl cDNA (step 9)
    148 μl H_2O
    2 μl 100× BSA
    20 μl 10× NEBuffer 4
    10 μl 10 U/μl *Nla*III.

    Mix and incubate 1 hr at 37°C.

    It is best to proceed with prewashing the streptavidin-Dynabeads (step 11) during this incubation such that the beads will be ready for use in the subsequent steps.

11. Thoroughly resuspend Streptavidin Dynabeads, exercising care to avoid excessive vortexing as streptavidin may be sheared off the magnetic beads. Transfer 200 μl beads to a No-stick siliconized 1.5-ml microcentrifuge tube and place in a magnetic rack. After ∼1 min remove supernatant. Wash beads twice in 200 μl of 1× BW then let stand in 200 μl of 1× BW until ready for use (up to several hours).

    All manipulations with Dynabeads are done using siliconized microcentrifuge tubes to avoid loss of yield due to products sticking to tube walls. All other manipulations, especially ethanol precipitations, should be done in standard microcentrifuge tubes.

    Dynabead washes are executed in the same fashion as done in the primary method (see Basic Protocol 1, step 2). Briefly, the beads are placed in the magnet 1 to 2 min. While the siliconized tube is still in the magnet, the buffer is gently pipetted off. The tube is then taken off the magnet and fresh buffer/wash is added to the tube and the beads are resuspended by agitation by hand or gentle vortexing. It is critical that the Dynabeads are not allowed to dry between the wash steps.

12. Ethanol precipitate cDNA from step 10 as described in step 8. Resuspend cDNA pellet in 200 μl of 1× BW.

13. Remove 1× BW from Dynabeads (step 11) and replace with 200 μl cDNA in BW. Mix gently by pipetting the mixture up and down. Incubate 15 min at room temperature with intermittent agitation by hand. Wash three times with 200 μl of 1× BW. Add 200 μl of 1× T4 DNA ligase buffer.

14. Prepare the following mix:

    2 μl 200 ng/μl linkers 2A and 2B (annealed)
    28 μl LoTE
    8 μl 5× T4 DNA ligase buffer.

    Remove 1× ligase buffer from the Dynabeads by pipetting and add the above mixture.

15. Mix bead slurry bound with cDNA gently, but well. Heat the tube 2 min at 50°C then incubate 15 min at room temperature.

16. Add 2 μl of 5 U/μl (high-concentration) T4 DNA ligase and incubate 2 hr at 16°C. Mix beads intermittently during ligation.

    It is best to use annealed linkers 2A/2B that are <1 month old.

Release cDNA with AscI

17. After ligation, wash beads four times with 1× BW/1× BSA. Wash in 1× NEBuffer 4/1× BSA and proceed immediately to the next step.

18. Resuspend the beads by adding the following components:

    85 μl LoTE buffer
    10 μl 10× NEBuffer 4
    2 μl 100× BSA
    2 μl 10 U/μl *Asc*I.

    Mix contents gently, but well using a pipette.

19. Incubate 1 hr at 37°C, agitating intermittently by hand every 15 min.

20. After digestion, collect supernatant carefully with a magnet. Place supernatant into a fresh nonsiliconized microcentrifuge tube. Add 50 μl LoTE to sample. Extract with PC8 as described in step 7.

21. High-concentration ethanol precipitate by combining the following:

    150 μl sample
    2 μl SeeDNA
    70 μl 7.5 M ammonium acetate
    500 μl 100% ethanol.

    Microcentrifuge 20 min at full speed, room temperature. Wash with 70% ethanol and resuspend in 25 μl LoTE.

    This is the concentrated rSAGE product, which may be stored indefinitely at −20°C. Avoid repeated freeze-thaw.

Amplify rSAGE-library dilutions by PCR

22. Make several dilutions of rSAGE product in LoTE.

Usually 1 µl of 1:25, 1:50, and 1:100 dilutions are recommended for PCR. Due to frequent variations in yield, this can vary widely. These dilutions are good starting point, however.

23. Prepare the following PCR reaction:

1 µl rSAGE dilution
5 µl 10× SAGE PCR buffer
3 µl DMSO
3 µl 10 mM dNTPs
1 µl 350 ug/µl M13 forward primer
1 µl 350 ug/µl primer 2
36 µl ddH$_2$O
1 µl 5 U/µl Platinum *Taq* DNA polymerase.

Repeat for all dilutions.

24. Use the following PCR cycling conditions:

Initial step:	2 min	94°C	(denaturation)
25 cycles:	45 sec	94°C	(denaturation)
	1 min	57°C	(annealing)
	1 min	70°C	(extension)
1 cycle:	5 min	70°C	(fill-in)
Final step:	indefinite	4°C	(hold).

25. Analyze 10 µl of each PCR product on a 4% to 20% Novex TBE acrylamide gel along with 1-kb ladder. Stain with 1× SYBR Green I in TBE buffer for 30 min and visualize under UV light.

A smear predominantly in the 200 to 500 bp range should be observed. Choose the highest rSAGE dilution that gives reliable results. The authors usually use the amplified 1:50 dilution of the rSAGE product. Amplified rSAGE libraries may be stored at −20°C for months.

PCR amplify using SAGE tag–specific primer and M13F primer

26. Prepare the following PCR mixture per reaction:

1 µl amplified rSAGE library (step 24)
5 µl 10× SAGE PCR buffer
2.5 µl DMSO
3 µl 10 mM dNTPs
10 µl 5 M betaine
1 µl 350 µg/µl M13 forward primer
1 µl 350 µg/µl SAGE tag–specific primer
25.5 µl H$_2$O
1 µl 5 U/µl Platinum *Taq*.

See Critical Parameters and Troubleshooting for a discussion of SAGE tag–specific primers.

27. Amplify under the following PCR cycling conditions:

Initial step:	2 min	93°C	(denaturation)
1 cycle:	30 sec	93°C	(begin touchdown)
	1 min	60°C	
	1 min	70°C	
15 cycles:	30 sec	93°C	(touchdown cycles)
	1 min	60° − 1°C/cycle	
	1 min	70°C	
30 cycles:	30 sec	93°C	(amplification cycles)
	1 min	44°C	
	1 min	70°C	
1 cycle:	5 min	70°C	(fill-in)
Final step:	indefinite	4°C	(hold).

These PCR cycling conditions are only guidelines that happen to work well for most SAGE tag–specific primers. A prolonged touchdown is pivotal for the specificity of priming. Optimal annealing temperatures may vary depending upon the nucleotide makeup of the SAGE tag–specific primer. Therefore, the touchdown annealing temperature should begin at least 10°C above the predicted oligonucleotide melting point (T_m). Over the 15 touchdown cycles, the annealing temperature should, by −1°C increments, settle upon the predicted SAGE-tag-specific primer's annealing temperature, where the rest of the 30 amplification cycles will proceed. It is not advisable to go below an annealing temperature of 40°C, regardless of how low the oligonucleotide T_m might be. Despite the apparent numerous amplification cycles used in this prolonged touchdown approach, the Taq polymerase remains very much active, mostly attributable to the protective effects of high-concentration betaine. See Critical Parameters and Troubleshooting for further discussion.

28. Visualize 5 µl of the PCR products on a 1.5% TBE agarose gel (*UNIT 2.7*).

The expected amplicons are usually between 100 to 400 bp, sometimes larger or smaller. Sometimes multiple bands may be amplified. If a control rSAGE amplified library was constructed, the band that is more intense in the experimental rSAGE library should be selected for further characterization. Often, multiple closely sized bands are amplification products of the same cDNA, attributable to variable oligo-dT priming along the poly-A tract during reverse transcription.

29. Load 25 µl of PCR products into a 1.5% TBE agarose gel and electrophorese until individual bands can be resolved (*UNIT 2.7*). Carefully excise the amplicon in the smallest agarose piece possible without sacrificing yield and place into a preweighed microcentrifuge tube.

30. Purify PCR product using the Qiaquick gel-extraction kit according to manufacturer's instructions (Qiagen). Elute Qiaquick columns with 30 µl EB Buffer. Proceed immediately to cloning using 4 µl eluant and the TOPO TA Cloning Kit or Cloning Kit for Sequencing per manufacturer's instructions.

If the only goal for the rSAGE procedure is to sequence the cDNA fragment, then the standard TA cloning vector pCR2.1 (Invitrogen) should suffice. However, if there are future plans for in vitro transcription of the cloned cDNAs, then it is advisable to use the TA cloning vector pCR4-TOPO (Invitrogen), which has both T7 and T3 RNA polymerase recognition sequences flanking the multiple cloning site.

IMPORTANT NOTE: *After TOPO TA cloning, do not use the M13 forward primer for subsequent colony PCR or cycle sequencing, as the M13 forward site will be embedded in the cloned cDNA. The M13 forward primer will not discriminate between M13 forward sites in the cDNA clone and the vector.*

31. Sequence TA cloning products using conventional methods (e.g., *CPMB UNIT 7.4A*).

PHOSPHORYLATING AND ANNEALING LINKERS

SUPPORT PROTOCOL 3

It is critical that the linkers be both annealed into double-stranded products and efficiently phosphorylated prior to ligation onto *Nla*III-digested cDNAs during SAGE-library construction. Even if linkers are ordered prephosphorylated, it is critical to test the efficiency of linker phosphorylation by self-ligation prior to SAGE library construction so as not to lose precious time and material. The following protocol details linker phosphorylation, annealing, and self ligation.

Additional Materials *(also see Basic Protocol 1)*

Linkers 1A, 1B, 2A, and 2B (see recipe)
10× kinase buffer (New England Biolabs)
10 mM ATP
10 U/μl T4 polynucleotide kinase (New England Biolabs)

Phosphorylate linkers

1. If linkers 1B and 2B are not already phosphorylated on their 5′ ends, prepare the following mixture:

   9 μl 350 ng/μl linker 1B or 2B
   6 μl LoTE buffer
   2 μl 10× kinase buffer
   2 μl 10 mM ATP
   1 μl 10 U/μl T4 polynucleotide kinase.

   Incubate 30 min at 37°C, then heat inactivate 15 min at 65°C.

Anneal linkers

2. Add 9 μl of 350 ng/μl linker 1A to 20 μl phosphorylated linker 1B.

3. Add 9 μl of 350 ng/μl linker 2A to 20 μl phosphorylated linker 2B.

4. Perform the following incubations on each linker pair:

   2 min at 95°C
   10 min at 65°C
   10 min at 37°C
   20 min at room temperature.

5. Dilute to 2 ng/μl with LoTE prior to use in SAGE-library construction.

Perform and check ligation

6. Prepare the following ligation reaction:

   0.5 μl annealed undiluted 350 ng/ml linker 1A + phosphorylated linker 1B (step 4)
   0.5 μl annealed undiluted 350 ng/ml linker 2A + phosphorylated linker 2B (step 4)
   7 μl H_2O
   1 μl 10× T4 DNA ligase buffer
   1 μl 5 U/μl (high-concentration) T4 DNA ligase buffer.

   Incubate 4 hr at 16°C.

   All linkers, whether ordered prephosphorylated or phosphorylated in house, should be checked for self-ligation.

7. Analyze product on a prepoured 20% polyacrylamide/TBE gel. Visualize as described (see Basic Protocol 1, step 37).

   Phosphorylated linkers should allow linker-linker dimers (80 to 100 bp) to form after ligation, while nonphosphorylated linkers will prevent self-ligation. Only linker pairs that self-ligate >70% should be used in further steps.

BASIC PROTOCOL 3

USING THE SAGE DATA ANALYSIS APPLICATION

The SAGE Data Analysis Application is a statistical computational program implementing a Poisson-based algorithm for analysis of SAGE data (Cai et al., 2004). The application allows users to compare two or multiple SAGE libraries, and to perform cluster analysis. The purpose of cluster analysis is to group tags (i.e., genes) with significant changes in expression levels that behave similarly under different conditions. It has been applied in a number of genomics studies in mouse retinal development (Blackshaw et al., 2004), fetal gut development (Lepourcelet et al., 2005), and diseases, such as cancer (Lepourcelet et al., 2005; Allinen et al., 2004).

There are two user platforms for the SAGE Data Analysis Application: one is an online web-based application and the other is a Microsoft Windows desktop-based application (stand-alone version). Both platforms perform the same set of analyses, the difference being that the web-based application does not require users to download and install the application onto a local computer. All data analyses are performed interactively. The potential drawback of the web-based application is that users need to submit their SAGE data onto the online application web server, which may risk the exposure of data to the public. If data security is a concern, the authors recommend that users use the Windows desktop-based application. The instructions in this protocol describe use of the online version.

Materials

Hardware

Computer with Internet access

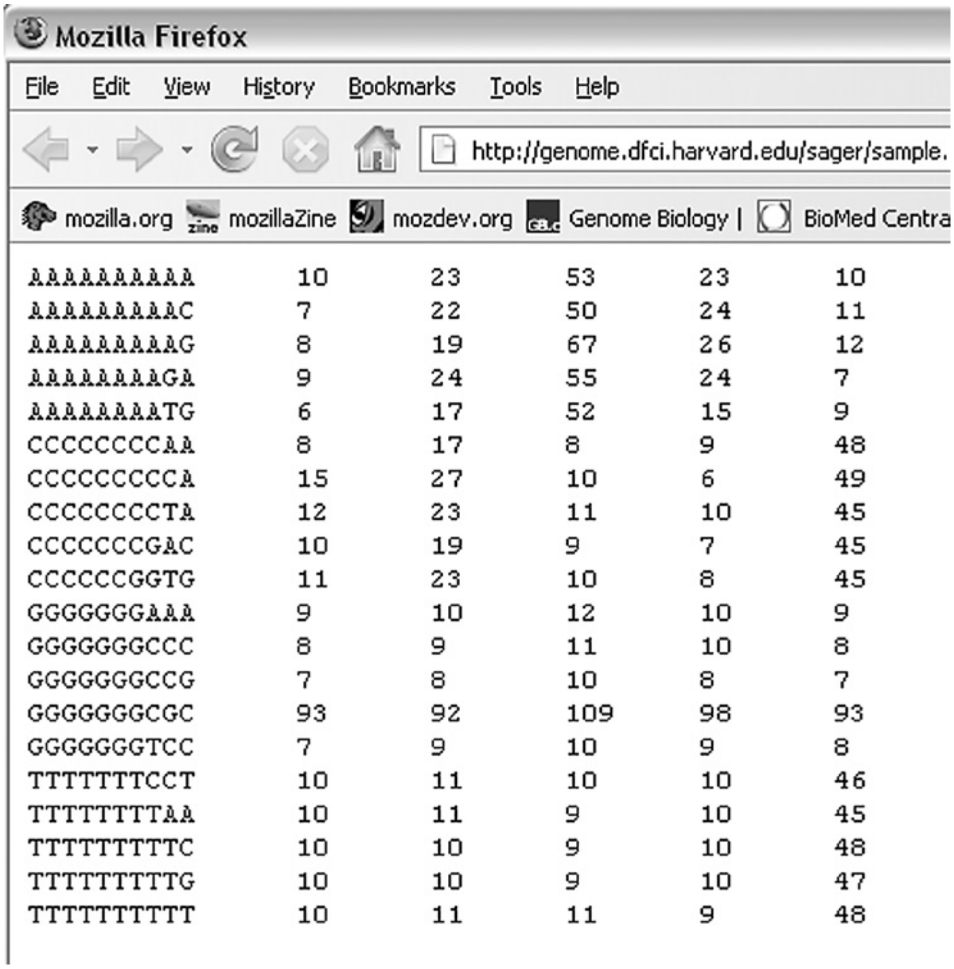
Software

An up-to-date Internet browser, such as Internet Explorer (*http://www.microsoft.com/ie*); Netscape (*http://browser.netscape.com*); Firefox (*http://www.mozilla.org/firefox*); or Safari (*http://www.apple.com/safari*).

Files

Raw SAGE data should be in tab-delimited text format with tags in rows and SAGE libraries in columns. The data file should contain a header line specifying what the data in each column is (see Fig. 11.7.3). The SAGE Data Analysis Application requires that the data file be sorted by the tag sequence column (see the first column in Fig. 11.7.3). In a Unix system this can be done with the "sort" command, and in Microsoft Windows system this can be done by choosing from the menu "Records" -> "Sort" in Microsoft Access or "Data" -> "Sort" in Excel After sorting, export or save data as a tab-delimited text file. If the Cluster Analysis module is used, the columns that contain tag counts for all libraries in the data file must be next to each other, i.e., if the libraries start from the second column and there are 5 libraries, the 2nd through 6th columns should be the columns for tag counts from each individual SAGE library (see Fig. 11.7.3).

NOTE: The data file can have as many extra columns as desired. As long as the correct column numbers are specified for the tag counts and first library the program should work.


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://genome.dfci.harvard.edu/sager/sample.
mozilla.org mozillaZine mozdev.org Genome Biology | BioMed Centra

```
AAAAAAAAAA     10     23     53     23     10
AAAAAAAAAC     7      22     50     24     11
AAAAAAAAAG     8      19     67     26     12
AAAAAAAAGA     9      24     55     24     7
AAAAAAAATG     6      17     52     15     9
CCCCCCCCAA     8      17     8      9      48
CCCCCCCCCA     15     27     10     6      49
CCCCCCCCTA     12     23     11     10     45
CCCCCCCGAC     10     19     9      7      45
CCCCCCGGTG     11     23     10     8      45
GGGGGGGAAA     9      10     12     10     9
GGGGGGGCCC     8      9      11     10     8
GGGGGGGCCG     7      8      10     8      7
GGGGGGGCGC     93     92     109    98     93
GGGGGGGTCC     7      9      10     9      8
TTTTTTTCCT     10     11     10     10     46
TTTTTTTTAA     10     11     9      10     45
TTTTTTTTTC     10     10     9      10     48
TTTTTTTTTG     10     10     9      10     47
TTTTTTTTTT     10     11     11     9      48
```

Figure 11.7.3 Screen shot of a sample SAGE data file. SAGE data file needs to be in tab-delimited format. All columns of SAGE libraries (tag counts) need to be arranged next to each other. Column 1 is SAGE tag, columns 2 to 6 are tag counts for 5 different SAGE libraries. For online version, the column headers are removed to keep data unidentifiable by other users.

Uploading a data file

1. Navigate to the home page for the SAGE Data Analysis Application (shown in Fig. 11.7.4) at *http://genome.dfci.harvard.edu/sager/*. Upload a tab-delimited data file by clicking “Browse” under “Step 1.” Navigate to the data file, select it and then click “Send” (Fig. 11.7.4).

   A new screen appears showing your data alongside some other samples.

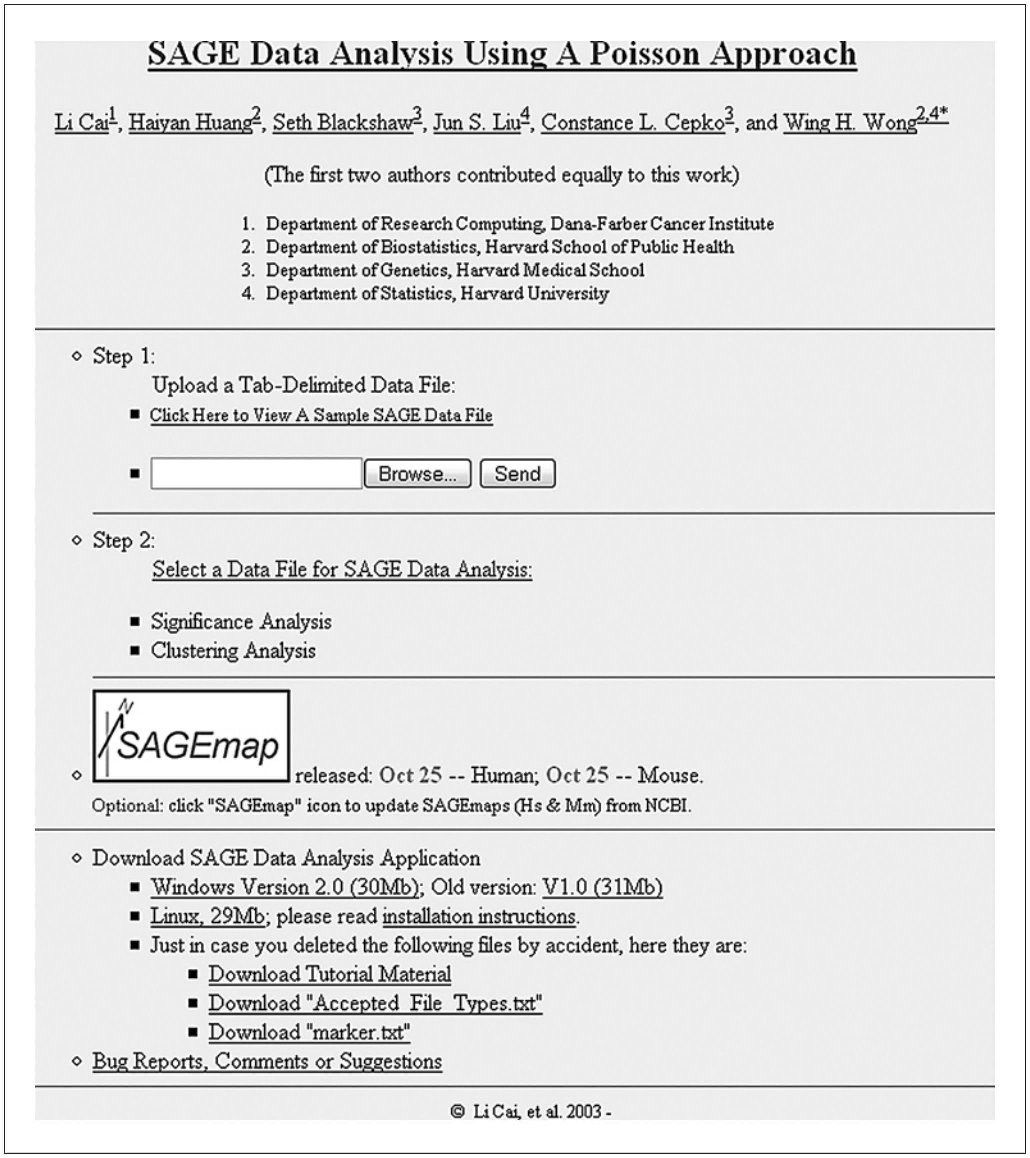
2. Select the data file of interest under “Step 2” on the screen.

   The choice of data file is confirmed and two calculation options are given (Fig. 11.7.5)–for significance analysis (Step 3a) and for cluster analysis (Step 3b).

Performing significance analysis

Significance analysis allows users to compare two or more different libraries and calculate *P*-values. The description of the algorithm used for Poisson-based significance analysis is in Appendix A at the end of this unit.

3. Click the numbered boxes to select SAGE libraries under “Step 3a.” As shown in Figure 11.7.5, boxes “1”, “3”, and “5” are selected for significance analysis.



SAGE Data Analysis Using A Poisson Approach

Li Cai[1], Haiyan Huang[2], Seth Blackshaw[3], Jun S. Liu[4], Constance L. Cepko[3], and Wing H. Wong[2,4*]

(The first two authors contributed equally to this work)

1. Department of Research Computing, Dana-Farber Cancer Institute
2. Department of Biostatistics, Harvard School of Public Health
3. Department of Genetics, Harvard Medical School
4. Department of Statistics, Harvard University

- Step 1:
  Upload a Tab-Delimited Data File:
  - Click Here to View A Sample SAGE Data File
  - ▭ Browse... Send
- Step 2:
  Select a Data File for SAGE Data Analysis:
  - Significance Analysis
  - Clustering Analysis
- SAGEmap released: Oct 25 -- Human; Oct 25 -- Mouse.
  Optional: click "SAGEmap" icon to update SAGEmaps (Hs & Mm) from NCBI.
- Download SAGE Data Analysis Application
  - Windows Version 2.0 (30Mb); Old version: V1.0 (31Mb)
  - Linux, 29Mb; please read installation instructions.
  - Just in case you deleted the following files by accident, here they are:
    - Download Tutorial Material
    - Download "Accepted_File_Types.txt"
    - Download "marker.txt"
- Bug Reports, Comments or Suggestions

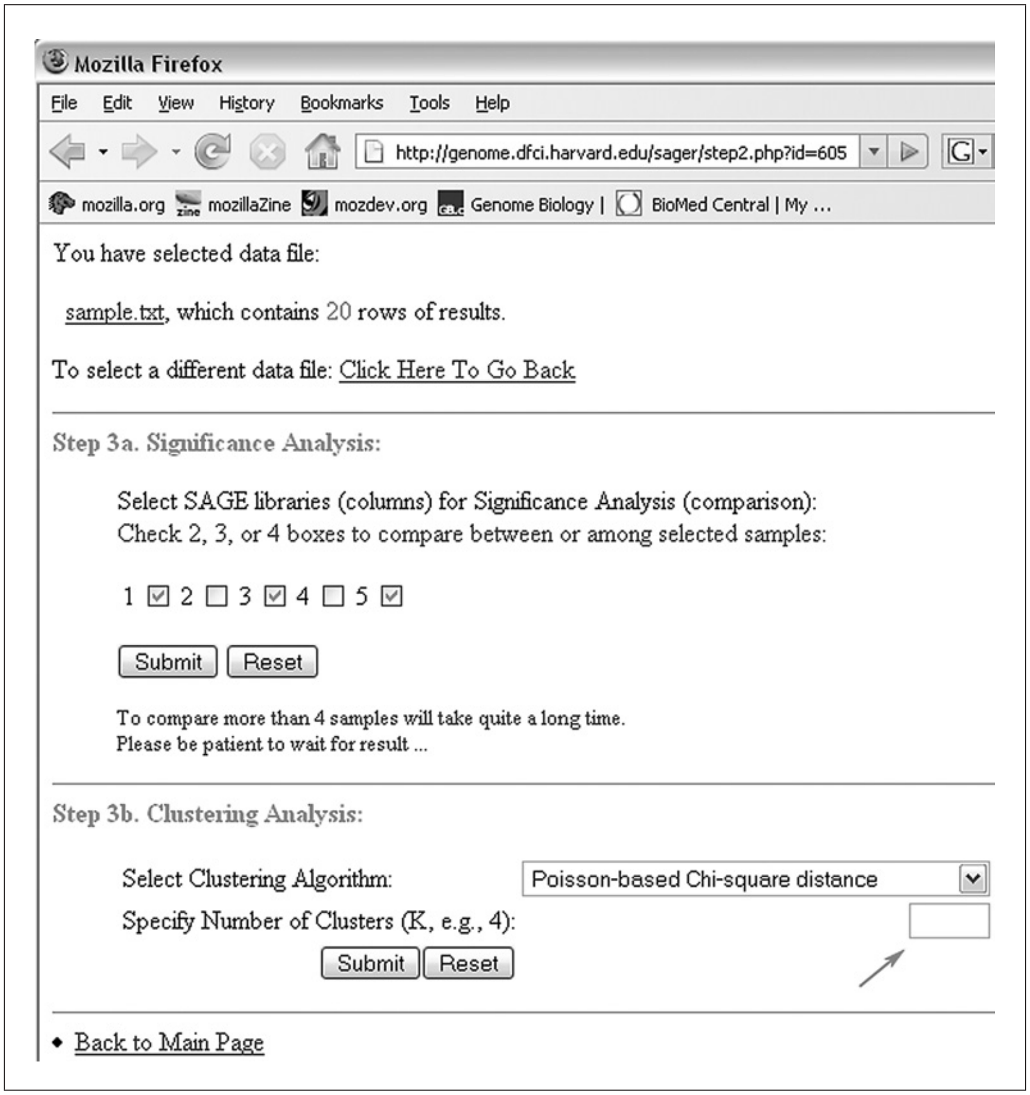
© Li Cai, et al. 2003 -

Figure 11.7.4 Screen shot of the main page of the online version of the SAGE Data Analysis Application.

4. Click "Submit." A new screen appears (Fig. 11.7.6).
5. Click the link to the result file ("sample.txt.1370.txt" in this example) to view or download results with calculated *P*-values. The result is shown in Figure 11.7.7. The last column contains the calculated *P*-values. The result file is a tab-delimited text file. Users can open the result file in Microsoft Excel or Access. Result data can be sorted by *P*-value to allow the selection of tags that are most significantly differentially expressed. The smaller the *P*-value, the more significantly differentially expressed the tag.
6. To annotate the data, select an organism for SAGE tag gene mapping (Fig. 11.7.6). Click "Submit." The annotated results appears on screen (Fig. 11.7.8).

Performing cluster analysis

Cluster analysis is more appropriate for multiple SAGE library data sets rather than simple pair-wise comparisons between libraries. Cluster analysis allows users to select several different algorithms (distances), including Poisson-based (PoissonC), Pearson


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://genome.dfci.harvard.edu/sager/step2.php?id=605
mozilla.org mozillaZine mozdev.org Genome Biology | BioMed Central | My ...
Submit Reset
Submit Reset
Back to Main Page

You have selected data file:

sample.txt, which contains 20 rows of results.

To select a different data file: Click Here To Go Back

Step 3a. Significance Analysis:

Select SAGE libraries (columns) for Significance Analysis (comparison):
Check 2, 3, or 4 boxes to compare between or among selected samples:

1 ☑ 2 □ 3 ☑ 4 □ 5 ☑

To compare more than 4 samples will take quite a long time.
Please be patient to wait for result ...

Step 3b. Clustering Analysis:

Select Clustering Algorithm: Poisson-based Chi-square distance

Specify Number of Clusters (K, e.g., 4):

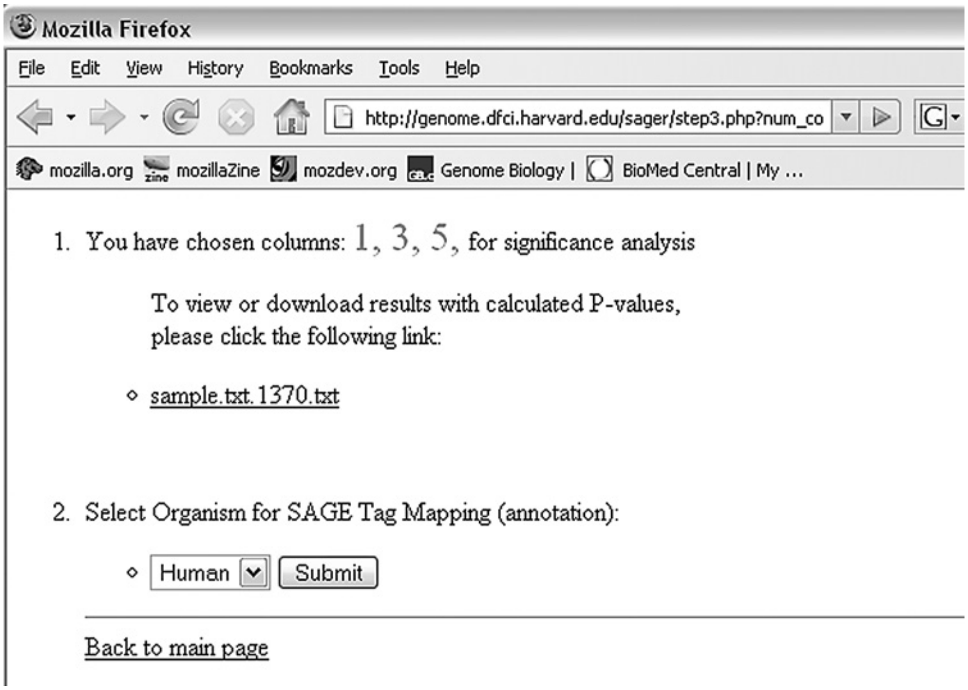
Figure 11.7.5 Screen shot for SAGE data significance analysis.

correlation (PearsonC), and Euclidean, etc., to group SAGE data into a user-defined number of clusters (k). We have found that Poisson-based clustering is generally most robust algorithm for analyzing SAGE data (Cai et al., 2004). The number of clusters cannot be more than the number of tags (genes) contained in the data file. It is recommended that users test a range of values for k. A more detailed discussion of how to set the value for k is found in (Hartigan, 1975).

To start cluster analysis, users begin with “Step 3b” as shown in Figure 11.7.5.

7. Select the clustering algorithm from the pull-down menu, as indicated by the arrow in Figure 11.7.5.
8. Enter the desired value in the “Specify Number of Clusters” box.
9. Click “Submit.”

   The run time is usually <1 min, but will vary depending on how large the dataset is. For example, for a large dataset containing >4000 unique tags, the run time could be as long as half of an hour. When clustering is finished a new screen appears, similar to that shown in Figure 11.7.6.


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://genome.dfci.harvard.edu/sager/step3.php?num_co
mozilla.org mozillaZine mozdev.org Genome Biology | BioMed Central | My ...

1. You have chosen columns: 1, 3, 5, for significance analysis

   To view or download results with calculated P-values,
   please click the following link:

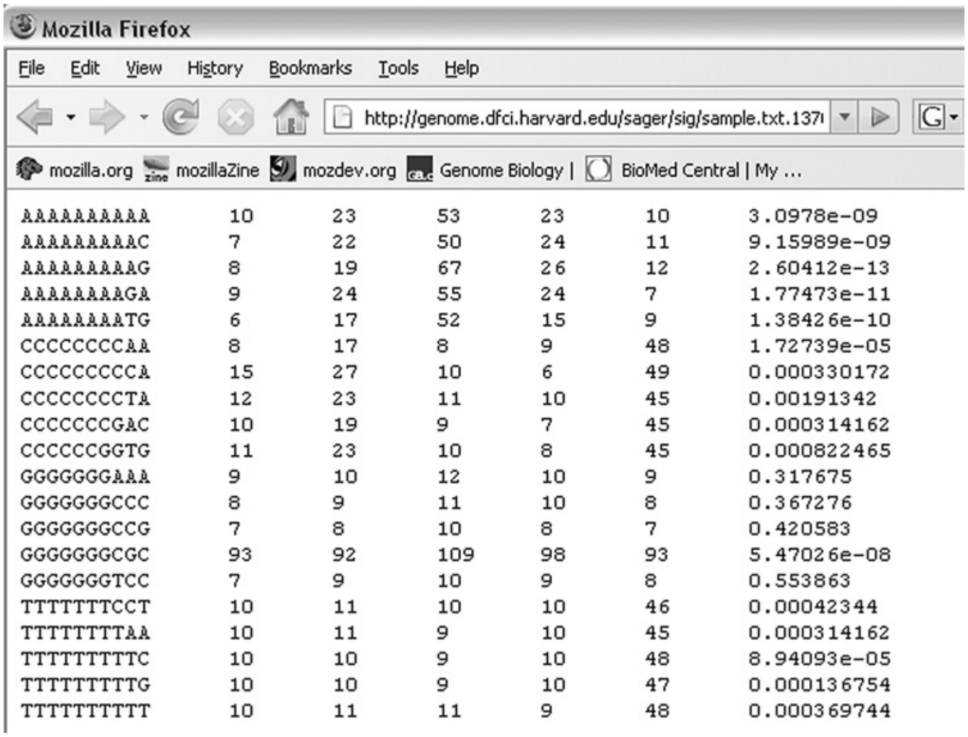
   - sample.txt.1370.txt

2. Select Organism for SAGE Tag Mapping (annotation):

   - Human Submit

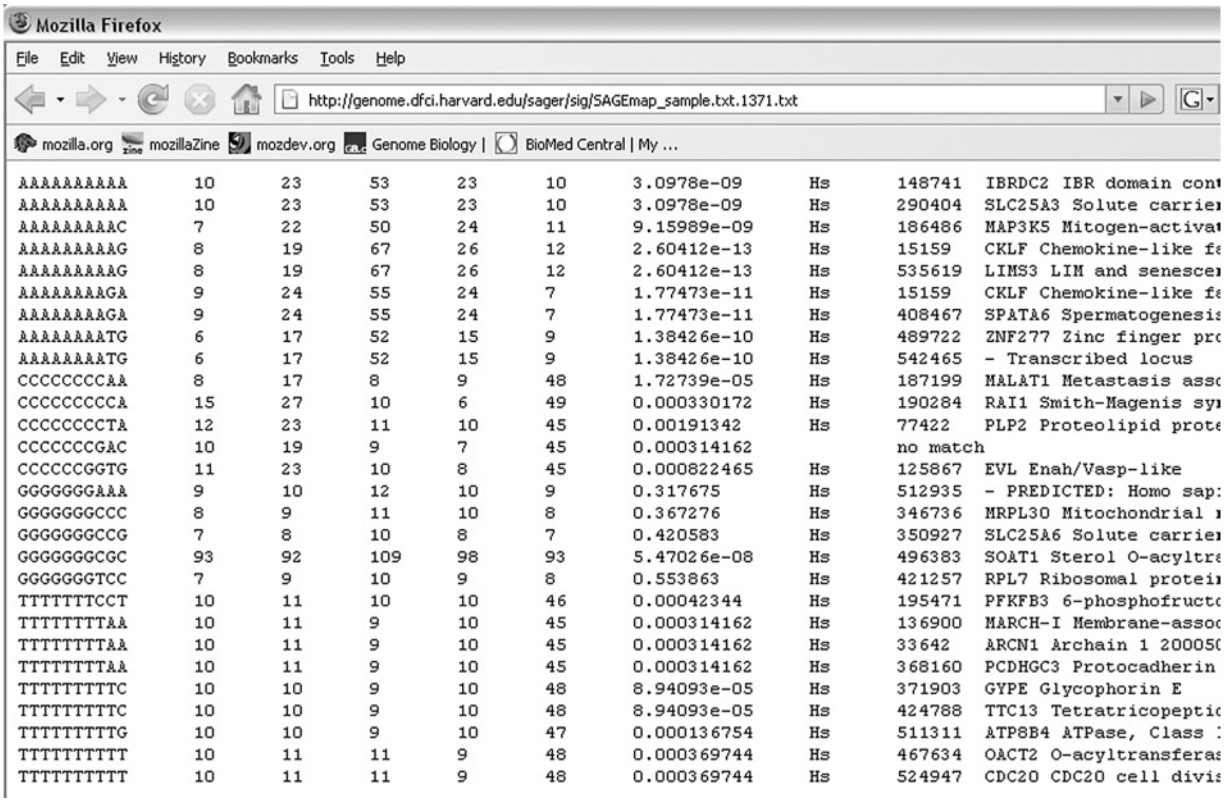
Back to main page

Figure 11.7.6 Selection of libraries 1, 3, and 5 for significance analysis.


Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://genome.dfci.harvard.edu/sager/sig/sample.txt.137
mozilla.org mozillaZine mozdev.org Genome Biology | BioMed Central | My ...

```
AAAAAAAAAA      10      23      53      23      10      3.0978e-09
AAAAAAAAAC      7       22      50      24      11      9.15989e-09
AAAAAAAAAG      8       19      67      26      12      2.60412e-13
AAAAAAAAGA      9       24      55      24      7       1.77473e-11
AAAAAAAATG      6       17      52      15      9       1.38426e-10
CCCCCCCCAA      8       17      8       9       48      1.72739e-05
CCCCCCCCCA      15      27      10      6       49      0.000330172
CCCCCCCCTA      12      23      11      10      45      0.00191342
CCCCCCCGAC      10      19      9       7       45      0.000314162
CCCCCCGGTG      11      23      10      8       45      0.000822465
GGGGGGGAAA      9       10      12      10      9       0.317675
GGGGGGGCCC      8       9       11      10      8       0.367276
GGGGGGGCCG      7       8       10      8       7       0.420583
GGGGGGGCGC      93      92      109     98      93      5.47026e-08
GGGGGGGTCC      7       9       10      9       8       0.553863
TTTTTTTCCT      10      11      10      10      46      0.00042344
TTTTTTTTAA      10      11      9       10      45      0.000314162
TTTTTTTTTC      10      10      9       10      48      8.94093e-05
TTTTTTTTTG      10      10      9       10      47      0.000136754
TTTTTTTTTT      10      11      11      9       48      0.000369744
```

Figure 11.7.7 Results from the significance analysis. Column 1 is the SAGE tag, columns 2 to 6 are 5 different SAGE libraries, column 7 is calculated *P*-value.



```
AAAAAAAAAA      10      23      53      23      10      3.0978e-09      Hs      148741  IBRDC2 IBR domain con
AAAAAAAAAA      10      23      53      23      10      3.0978e-09      Hs      290404  SLC25A3 Solute carrie
AAAAAAAAAC      7       22      50      24      11      9.15989e-09     Hs      186486  MAP3K5 Mitogen-activa
AAAAAAAAAG      8       19      67      26      12      2.60412e-13     Hs      15159   CKLF Chemokine-like fa
AAAAAAAAAG      8       19      67      26      12      2.60412e-13     Hs      535619  LIMS3 LIM and senesce
AAAAAAAAGA      9       24      55      24      7       1.77473e-11     Hs      15159   CKLF Chemokine-like fa
AAAAAAAAGA      9       24      55      24      7       1.77473e-11     Hs      408467  SPATA6 Spermatogenesis
AAAAAAAATG      6       17      52      15      9       1.38426e-10     Hs      489722  ZNF277 Zinc finger pro
AAAAAAAATG      6       17      52      15      9       1.38426e-10     Hs      542465  - Transcribed locus
CCCCCCCCAA      8       17      8       9       48      1.72739e-05     Hs      187199  MALAT1 Metastasis asso
CCCCCCCCCA      15      27      10      6       49      0.000330172     Hs      190284  RAI1 Smith-Magenis sy
CCCCCCCCTA      12      23      11      10      45      0.00191342      Hs      77422   PLP2 Proteolipid prote
CCCCCCCGAC      10      19      9       7       45      0.000314162             no match
CCCCCCGGTG      11      23      10      8       45      0.000822465     Hs      125867  EVL Enah/Vasp-like
GGGGGGGAAA      9       10      12      10      9       0.317675        Hs      512935  - PREDICTED: Homo sap:
GGGGGGGCCC      8       9       11      10      8       0.367276        Hs      346736  MRPL30 Mitochondrial
GGGGGGGCCG      7       8       10      8       7       0.420583        Hs      350927  SLC25A6 Solute carrie
GGGGGGGCGC      93      92      109     98      93      5.47026e-08     Hs      496383  SOAT1 Sterol O-acyltra
GGGGGGGTCC      7       9       10      9       8       0.553863        Hs      421257  RPL7 Ribosomal protei
TTTTTTTCCT      10      11      10      10      46      0.00042344      Hs      195471  PFKFB3 6-phosphofructo
TTTTTTTTAA      10      11      9       10      45      0.000314162     Hs      136900  MARCH-I Membrane-asso
TTTTTTTTAA      10      11      9       10      45      0.000314162     Hs      33642   ARCN1 Archain 1 20005
TTTTTTTTAA      10      11      9       10      45      0.000314162     Hs      368160  PCDHGC3 Protocadherin
TTTTTTTTTC      10      10      9       10      48      8.94093e-05     Hs      371903  GYPE Glycophorin E
TTTTTTTTTC      10      10      9       10      48      8.94093e-05     Hs      424788  TTC13 Tetratricopepti
TTTTTTTTTG      10      10      9       10      47      0.000136754     Hs      511311  ATP8B4 ATPase, Class
TTTTTTTTTT      10      11      11      9       48      0.000369744     Hs      467634  OACT2 O-acyltransfera
TTTTTTTTTT      10      11      11      9       48      0.000369744     Hs      524947  CDC20 CDC20 cell divis
```

Figure 11.7.8 Annotated results after tag matching with SAGEmap. Column 1 is the SAGE tag, columns 2 to 6 are 5 different SAGE libraries, column 7 is calculated *P*-value, column 8 is organism (Hs = *homo sapiens*), column 9 is unigene ID, column 10 is gene symbol and gene description.

10. Select an organism for annotation, then click "Submit." A screen with graphs appears. To view members of a cluster, click on the individual graphs. A new window appears with all members in the clicked cluster (Fig. 11.7.9).

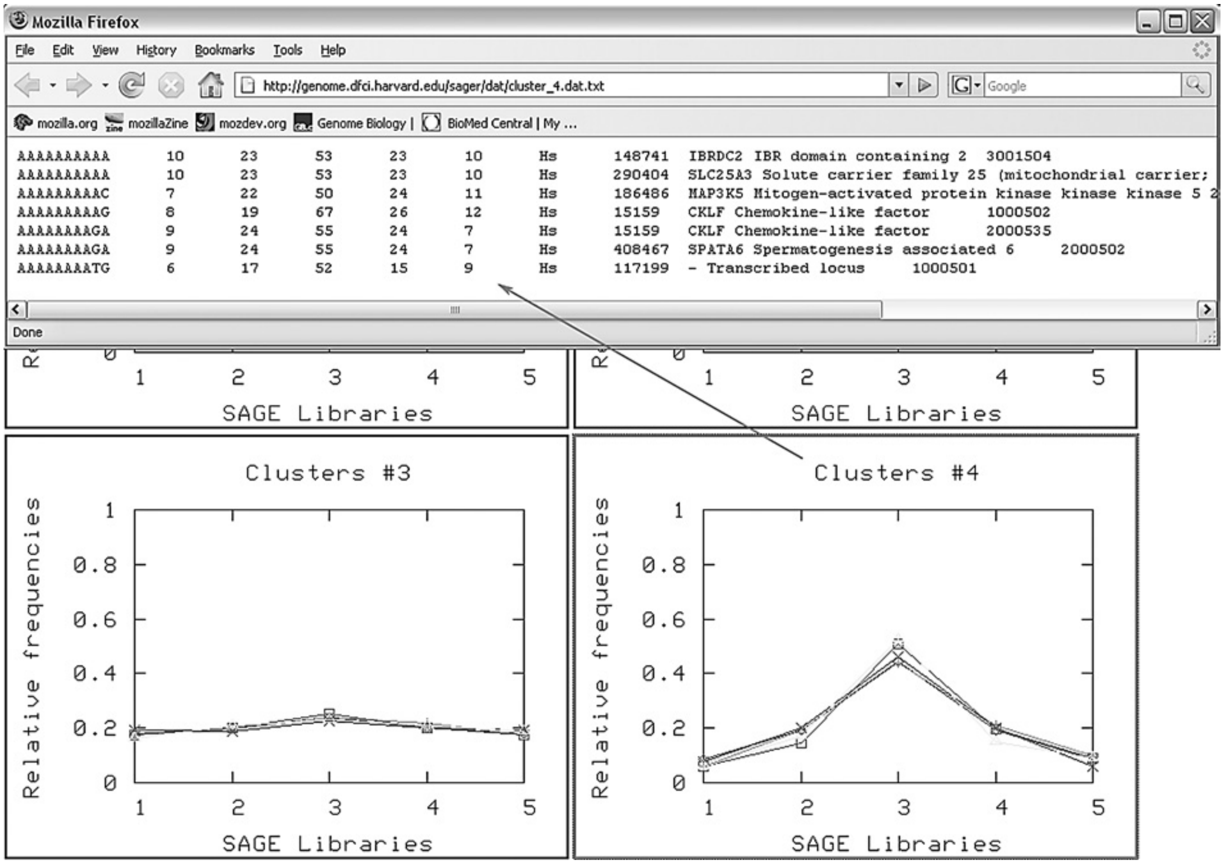
11. To save graphs, right click on individual graphs. Select "Save Image As . . ." from the menu. The user then selects a directory where the graph is to be saved.

Prioritizing data for further analysis

Once particular clusters of interest have been identified, genes can be prioritized for further study based on a variety of criteria. Genes that match specific SAGE tags can be rapidly functionally annotated with Gene Ontology criteria using Web-based programs such as EASE (Hosack et al., 2003), and genes that match a particular function of interest can then be selected. Genes can also be prioritized based on abundance levels or by relative tissue-specificity. It can be very useful to include additional SAGE libraries from public repositories in the analysis to help generate more robust clusters. Some sources of this data include the SAGEmap (*http://www.ncbi.nlm.nih.gov/SAGE/*) and SAGE Genie (*http://cgap.nci.nih.gov/SAGE*) sites found at NCBI and at *http://www.mouseatlas.org*.

Suggestions for improved results

Given the fact that observed SAGE tag levels are actually found in a Poisson distribution about their actual abundance level (Audic and Claverie, 1997), an abundance threshold can be usefully applied to the data prior to submission for cluster analysis. The exact value to use should be determined empirically, and largely depends on how many false positives one is willing to tolerate in each cluster. Tag counts ≥ 5 in at least one of the SAGE libraries is a good value to start with. Significance analysis indicates that when



Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://genome.dfci.harvard.edu/sager/dat/cluster_4.dat.txt
Google
mozilla.org mozillaZine mozdev.org Genome Biology | BioMed Central | My ...
Done
Clusters #3
Clusters #4
Relative frequencies
0
0.2
0.4
0.6
0.8
1
1
2
3
4
5
SAGE Libraries

```
AAAAAAAAAA      10      23      53      23      10      Hs      148741  IBRDC2 IBR domain containing 2  3001504
AAAAAAAAAA      10      23      53      23      10      Hs      290404  SLC25A3 Solute carrier family 25 (mitochondrial carrier;
AAAAAAAAAC      7       22      50      24      11      Hs      186486  MAP3K5 Mitogen-activated protein kinase kinase kinase 5 2
AAAAAAAAAG      8       19      67      26      12      Hs      15159   CKLF Chemokine-like factor      1000502
AAAAAAAAGA      9       24      55      24      7       Hs      15159   CKLF Chemokine-like factor      2000535
AAAAAAAAGA      9       24      55      24      7       Hs      408467  SPATA6 Spermatogenesis associated 6     2000502
AAAAAAAATG      6       17      52      15      9       Hs      117199  - Transcribed locus     1000501
```

Figure 11.7.9 Screen shot shows cluster #2 and all of its members after clicking on the graph of cluster #2.

comparing 2 or more libraries, with tag count 5 in one library versus tag count 0 or 1 in the other library, $p \leq 0.05$. This means SAGE tags that are included in clustering analysis are significantly differentially expressed tags.

Using the stand-alone version of the software

To perform the analysis of SAGE data on a desktop computer, obtain a copy of the application from the SAGE Data Analysis Application website (*http://genome.dfci.harvard.edu/sager/*) and store it onto the desktop computer, and double click the downloaded file to start installation. Follow instructions to finish the installation process. There will be an application icon called "SAGE Data Analysis" on the desktop. Double click the icon to start the program. The instructions and tutorial of the stand-alone version are included in the software download package. The program is free for public use. This program is distributed in the hope that it will be useful for research purpose, but WITHOUT ANY WARRANTY.

REAGENTS AND SOLUTIONS

Use double-distilled water in all recipes and protocol steps. For common stock solutions, see **APPENDIX 2D***; for suppliers, see* **SUPPLIERS APPENDIX***.*

BRS1 primer

5′-Biotin-CCGGGCGCGCCGTAAAACGACGGCCAG$(T)_{19}$-3′

Order HPLC purified from a trusted supplier. The authors recommend using Integrated DNA Technologies (IDT).

BW buffer, 1×

For 2 *stock:*
10 mM Tris·Cl, pH 7.5 (*APPENDIX 2D*)
1 mM EDTA
2.0 M NaCl
Store up to 1 year at room temperature
Dilute to 1× with H_2O just before use

Linkers

Linker 1A: 5′TTTGGATTTGCTGGTGCAGTACAACTAGGCTTAATAGGGACATG 3′
Linker 1B: 5′TCCCTATTAAGCCTAGTTGTACTGCACCAGCAAATCC[amino mod C7] 3′
Linker 2A: 5′TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGGGGACATG 3′
Linker 2B: 5′TCCCCGTACATCGTTAGAAGCTTGAATTCGAGCAG[amino mod C7] 3′

The authors recommend using Integrated DNA Technologies for ordering oligonucleotides.

LoTE buffer

3 mM Tris·Cl, pH 7.5 (*APPENDIX 2D*)
0.2 mM EDTA, pH 7.5 (*APPENDIX 2D*)
Store up to 1 year at room temperature

PC8

480 ml phenol, warmed to 65°C
320 ml 0.5 M Tris·Cl, pH 8.0 (*APPENDIX 2D*)
640 ml chloroform

Add in sequence and place at 4°C. After 2 to 3 hr, shake again. After an additional 2 to 3 hr, aspirate aqueous layer. Store up to 1 year in aliquots at −20°C or 6 months at 4°C.

Commercially available 1:1 (v/v) phenol:chloroform mix can also be substituted, as long as the pH is preset to 8.0.

PCR primers

Primer 1:	5′ GGATTTGCTGGTGCAGTACA 3′
Primer 2:	5′ CTGCTCGAATTCAAGCTTCT 3′
M13 forward:	5′ GTAAAACGACGGCCAGT 3′
M13 reverse:	5′ GGAAACAGCTATGACCATG 3′

The authors recommend using Integrated DNA Technologies for ordering oligonucleotides.

SAGE PCR buffer, 10×

166 mM ammonium sulfate
670 mM Tris·Cl, pH 8.8 (*APPENDIX 2D*)
67 mM $MgCl_2$
100 mM 2-mercaptoethanol
Dispense into aliquots and store up to 1 year at −20°C.

SAGE tag–specific primer

5′-GACATGXXXXXXXXXX-(10-bp SAGE tag)-3′

If the SAGE-tag-specific primer has a calculated annealing temperature below 40°C, incorporate additional bases further 5′ on linker 2A (see recipe for linkers) to increase the oligonucleotide melting temperature. The full linker 2A-SAGE tag sequence is as follows:

5′-TTTCTGCTCGAATTCAAGCTTCTAACGATGTACGGGGACATGXXXXXX XXXX-(10-bp SAGE tag)-3′

The SAGE 2000 software has the ability to extract an additional base for an 11-base tag. This may be helpful, as any additional sequence-specific bases may yield a more specific product.

Zeocin-containing low-salt LB plates

For 1 liter:
10 g tryptone
5 g yeast extract
5 g NaCl

Adjust the pH to 7.5 and add 15 g bactoagar. Autoclave solution and allow to cool before adding zeocin to 100 mg/ml.

COMMENTARY

Background Information

Serial analysis of gene expression (SAGE) was first developed in 1995 (Velculescu et al., 1995), and has since been used to generate a large variety of data from normal and cancerous human tissue (Zhang et al., 1997; Boon, et al. 2002), yeast (Velculescu et al., 1997), *C. elegans* (Halaschek-Wiener et al., 2005), *D. melanogaster* (Gorski et al., 2004), mouse (Virlon et al., 1999; Blackshaw et al., 2004), rat (Klimaschewski et al., 2000), and even (with modifications) human oocytes (Neilson et al., 2000).

SAGE is a powerful method for providing genome-wide gene-expression data. In much the same fashion as EST libraries, SAGE utilizes cDNA "tags" which are sequenced and quantified. The 14-bp SAGE tags differ from ESTs essentially by size, allowing subsequent concatenation and high-throughput sequencing in much greater volumes. The location of the anchoring enzyme site is essentially sufficient to uniquely identify the cognate cDNA or gene. The original protocol required relatively large amounts of starting material (2 to 5 μg of polyA mRNA) and was technically quite challenging, frequently giving variable results even in experienced hands. Major improvements were made to the protocol by a number of groups (Virlon et al., 1999; Datson et al., 1999; St. Croix et al., 2000), which collectively gave rise to a version of the protocol known as microSAGE (see Basic Protocol 1), owing to the fact that over 1000-fold less starting material could be readily used for library construction. The critical modifications appear to have been anchoring the mRNA to magnetic beads prior to cDNA synthesis (rather than after cDNA sythesis via incorporation of a biotinylated oligo(dT) primer as in the original protocol) and optimization of the quantities of reagents used, in particular, the quantities of linkers. Additional improvements, such as heating the ditag concatemers prior to gel purification (Angelastro et al., 1999), have resulted in SAGE libraries with substantially higher insert frequency and larger insert size than in the original protocol. These technical improvements, coupled with the drop in the cost of DNA sequencing, have combined to allow the generation of over 3.5 million human SAGE tags alone, many of which are publicly available for analysis (*http://www.ncbi.nlm.nih.gov/SAGE*).

SAGE analysis has a number of unique advantages over hybridization-based measures of global gene expression, such as microarray analysis (*UNITS 11.1, 11.3*), or approaches such as subtractive hybridization (*CPMB UNITS 25B.1 & 25B.2*) and differential display methodologies (*UNIT 11.5*). Since very few mRNAs lack *Nla*III sites, SAGE generates a tag for virtually every cellular mRNA, providing a level of coverage unequaled by any microarray yet available for humans or mice. For these same reasons, SAGE can also serve as a tool for gene discovery and transcript annotation even in species with fully sequenced genomes. The sensitivity of SAGE is limited only by the number of tags that one has the desire or resource to sequence and, with larger numbers of tags sequenced, it becomes possible to determine relatively small (<2-fold) changes in gene expression between samples. Since individual SAGE tag levels are expressed as a percentage of total tags, it is straightforward to compare tag levels among libraries generated by other labs. As more SAGE libraries are generated and made public, these data sets can be used to generate a large-scale atlas of gene expression that is of great use to the whole scientific community. Such a resource is already available for human normal and malignant tissues at NCBI (*http://www.ncbi.nlm.nih.gov/SAGE*), and libraries from other species are available from various sources (see Internet Resources for a partial list).

The main drawbacks of SAGE analysis are the time and expense required to generate sufficient numbers of tags to examine expression of low and moderate-abundance mRNAs. The price of sequencing has dropped considerably in the past few years, but real costs still remain around $0.25/tag. For researchers simply hoping to identify a handful of differentially expressed genes in their sample of interest, subtractive hybridization (*CPMB UNIT 25B.1*), differential display methodologies (*UNIT 11.7*), or even the use of commercially available microarray technology may prove more cost-effective. An additional drawback of SAGE is the requirement that a large body of cDNA/EST sequence must be available from the organism being studied in order to match SAGE tags to the specific mRNAs. This effectively limits the use of SAGE to model organisms. Another drawback of the method is the occasional failure of a SAGE tag to match a predicted gene or to be long enough to easily isolate a full-length cDNA clone. While this happens at relatively low frequency for high abundance transcripts in model organisms, it can limit the interpretation of the data in some cases.

As a result, several approaches, most of which are variations of conventional RT-PCR, have been developed to identify these unknown or anonymous SAGE tags. There has been marked improvements in strategies used to identify unknown SAGE tags by reverse-cloning cDNA fragments, collectively called reverse SAGE (rSAGE; see Basic Protocol 2). First, the cloning process is similar to the original SAGE protocol; therefore, only cDNA pieces which are 3′ to the most 3′ anchoring enzyme site are used as templates for subsequent PCR amplification and subcloning (Polyak et al., 1997; also see Internet Resources, SAGEnet). Second, the use of betaine allows for a prolonged PCR touchdown that results in more specific priming.

Critical Parameters and Troubleshooting

MicroSAGE

The two key determinants of a successful SAGE library are quantity and purity of ditags. To ensure obtaining many ditags, carefully optimize the starting reaction and scale up the number of PCR reactions as desired. For certain low-yield preparations, the authors have gone as high as 700 PCR reactions of 50 µl to generate the starting material. For purity, ensure that the 102-bp and the 80-bp bands are well separated, and be very careful not to extract any of the 80-bp band. Run the gel as long as possible and do not overload the wells (no more than 10 µl per well, despite the large number of gels this will require). Do the same for the 26-bp cut ditag band (avoiding the 40-bp linker band).

One other problem that has been encountered occasionally is contamination of reagents following construction of libraries, which will result in 102-bp bands in the no-ligase control in the initial optimization PCR reactions. To avoid this, be very careful to avoid splashes and not reuse tips during the scale-up or initial purification of the 102-bp band. Make separate aliquots of LoTE buffer, PC8, ammonium acetate, and ethanol for each library during these steps to reduce the likelihood of contamination. Use aerosol-barrier tips wherever possible.

A final common cause of experimental failure is low-quality reagents. Wherever possible, order supplies from the sources specified in the protocol. The authors have most frequently observed problems with the *Nla*III enzyme and the linkers. Always store *Nla*III in aliquots at −80°C, do not reuse aliquots, and try to have the enzyme shipped on dry ice if possible. The authors order linkers prekinased, but always check via self-ligation to ensure that a sufficiently large fraction of the linkers is properly phosphorylated.

rSAGE

For the rSAGE procedure, much depends on the quality of RNA used in the sample. It would be best to use the same batch of RNA

Table 11.7.1 Troubleshooting for SAGE Reactions

Problem	Possible Cause	Solution
MicroSAGE		
No PCR product with control primers following cDNA synthesis	Dynabeads inactive	Store Dynabeads at 4°C only; do not freeze
	Reverse transcriptase inactive	Replace reverse transcriptase
	RNA degraded prior to homogenization	Minimize delay between tissue harvesting and homogenization
	Cells insufficiently lysed	Homogenize tissue thoroughly. Use homogenization by Polytron only.
Ditag PCR product is slightly shorter (running at ∼90 bp) and will not redigest with *Nla*III	Failure to completely remove *E. coli* DNA polymerase I following second strand cDNA synthesis	Do not omit or shorten SDS washes or 75°C heat inactivation step
PCR product in no-ligase control at 100 bp	Contamination of reagent by ditags from a previously constructed SAGE library	Use separate aliquots of LoTE buffer, ammonium acetate, and PC8 for each large-scale ditag purification. Use aerosol pipet tips.
Ditag yield low (100-bp band <80-bp band)	Ratio of linkers to cDNA too high	Reduce amount of linkers in ligation
	cDNA synthesis inefficient	See advice in steps 1-11
	PCR conditions not optimized	Titrate dNTP concentration and number of amplification cycles
	Quantity of starting material too low	Increase amount of starting material
Ditags do not cut with *Nla*III following purification	*Nla*III inactive	Store enzyme in aliquots at −80°C. Do not reuse thawed aliquots.
	Ditags insufficiently pure	Run Qiaquick gel extraction on eluate (see step 44)
Ditag concatemers not generated efficiently	Insufficient quantity of purified ditags used in ligation	Increase quantity of cDNA used for large-scale ditag prep and/or increase number of cycles of amplification
	Ditags used in ligation are insufficiently pure	Run preparative gel longer to more efficiently separate 100- and 80-bp bands
Ditag concatemers most concentrated at high molecular weight (3kb) and do not clone efficiently	Concatemer ligation not heated properly	Heat at 65°C and chill on ice immediately

continued

Table 11.7.1 Troubleshooting for SAGE Reactions

Problem	Possible Cause	Solution
Concatemers have a high (5%) frequency of duplicate ditags	Too many cycles of PCR used to reamplify ditags	Reduce number of cycles used. Increase amount of starting material when making cDNA
rSAGE procedure		
No SAGE tag-specific PCR product	Degraded RNA	Use fersh RNA
	Error in generation of rSAGE amplified library	Reconstruct rSAGE library
	Poor tag specific primer design	Redesign primer (see Critical Parameters and Troubleshooting)
	Low abundance transcript	Increase the number of amplification cycles
Multiple SAGE tag–specific PCR products	Multiple splice variants or multiple gene identities for a given SAGE tag	Clone all PCR amplicons
		Construct a control rSAGE library and select amplicons not present in control library
	Nonspecific priming	Start the PCR touchdown cycles at a higher temperature
Poor sequence quality	Use of M13 forward primer for cycle sequencing	Use another universal primer on the pCR vector for sequencing—e.g., M13 reverse, T3, T7

that was originally used to construct the SAGE library. As most interesting SAGE tags are those that are expressed in abundance in one RNA sample and not in another, it is advisable to make a reverse-SAGE library of such a control tissue. It is not uncommon to generate multiple PCR bands from a tag-specific rSAGE amplification. Identifying a PCR product that is specific to the experimental rSAGE library and not present (or less apparent) in the control would help in the cloning and identification process.

The most technically challenging aspect of reverse-cloning SAGE tags is the PCR of a specific cDNA with the tag-specific primer. The rSAGE-amplified library used as a template for this PCR reaction consists solely of 3′-cDNA ends which have the linker2-SAGE tag on the 5′ end and a oligo dT-M13 forward sequence on the 3′ end. The PCR of a specific product is difficult when the reverse primer (M13 Forward) anneals to all templates, and the forward primer (SAGE-tag specific) shares the same sequences on the 5′ end. Specificity is conferred only by the last 10 bases on the forward primer, representing the unique 10-base SAGE tag. One may also choose to incorporate an additional SAGE-tag base, information that the SAGE 2000 software can extract from the raw data. The SAGE tag–specific PCR is executed with a prolonged touchdown using an automatic hot-start *Taq* polymerase (i.e., Platinum *Taq*; Invitrogen). As a 15-cycle touchdown requires 46 denaturing cycles, betaine is used as a *Taq* polymerase protectant. The authors strongly advise against switching to a proofreading DNA polymerase, such as *Pfu* or Vent, in the PCR reactions. Proofreading enzymes have significant 3′-5′ exonuclease activity which may digest the 3′ end of the SAGE tag–specific primer. Even one-base differences may reduce the specificity of the PCR product.

Designing of SAGE tag–specific primers is a matter of much debate. Only the 3′-most ten bases of the oligonucleotide contains tag-specific sequences, and the rest of the primer at the 5′ end consists of linker sequences which are shared by all the cDNAs in the amplified rSAGE library. As a result, the authors empirically use CACATG-XXXXXXXXXX as a guideline for primer design where the Xs refer to the specific sequence in the SAGE tag of interest. Only six bases are nonspecific, and the relatively low annealing temperatures allow for an extended touchdown starting at a temperature that is well above the oligonucleotide melting point. However, if the rSAGE-specific primer has an annealing temperature which is too low, there is a risk of the primers melting off the template before the extension cycle. Therefore, if the calculated T_{m} of the SAGE tag specific primer is below 40°C, it is advisable to incorporate more of the linker sequence to raise the melting temperature of the oligo.

In the rare case that the SAGE tag in question lies immediately 5′ to the polyA tail, reverse-SAGE may yield no additional information, and the PCR product may be too small to adequately visualize on a 1.5% agarose gel.

Additional troubleshooting guidelines are presented in Table 11.7.1.

Anticipated Results

If Basic Protocol 1 is followed closely, libraries containing >85% inserts with an average size of 30 to 50 tags (450 to 750 bp) should be routinely generated. This should enable one to obtain a SAGE data set of 50,000 tags after ∼2000 individual sequencing reactions.

If the above guidelines for rSAGE (see Basic Protocol 2) are followed, one should be able to clone the cDNA, usually 75 to 400 bp, from which a given SAGE tag is generated. This cDNA fragment would stretch from the 3′-most anchoring-enzyme site to the poly-A tail. The additional sequence data can be used to BLAST genome databases (*UNIT 6.8*) or be used to generate primers for 5′ RACE (*CPMB UNIT 15.6*). The cloned fragment may also be used for northern analyses (*APPENDIX 3K*) or in situ hybridizations (Chapter 4).

Time Considerations

MicroSAGE

The time typically taken for RNA preparation through *Bsm*FI digestion is 10 to 14 hr. Blunt-ending and ditag-ditag ligation take 2 to 3 hr. Ditag amplification and PCR optimization take 2 to 3 hr and large-scale ditag amplification and purification take 6 to 8 hr/day for 2 days. Ditag digestion and purification take 6 to 8 hr. Concatemer formation, purification, and subcloning take 6 to 8 hr. Template cleanup and transformation take 4 to 6 hr. PCR of library clones and gel analysis take 4 to 5 hr.

If a high-quality SAGE library is produced, it will require ∼2000 sequencing reactions to obtain 50,000 tags. This will take anywhere from an additional 1 week to 3 months, depending on the resources and sequencing capacity.

rSAGE

Generating purified double-stranded cDNA typically takes 4.5 hr. Cleaving the cDNA with the anchoring enzyme (*Nla*III), magnetic bead purification, ligating linkers to cDNA, and release of 3′ cDNA fragments from magnetic beads with *Asc*I typically takes 6 to 8 hr. PCR generation of amplified rSAGE libraries takes 2.5 to 3.5 hr. SAGE tag-specific PCR takes 3.5 to 4.5 hr. TOPO-TA cloning and subsequent sequencing is user-dependent.

Acknowledgement

The authors are grateful to the collaborators who kindly provided the data, and to the many users who provided valuable feedback, suggestions, and help. The authors wish to thank Feng X. Zhao and members of Research Computing at Dana-Farber Cancer Institute.

Literature Cited

Allinen, M., Beroukhim, R., Cai, L., Brennan, C., Lahti-Domenici, J., Huang, H., Porter, D., Hu, M., Chin, L., Richardson, A., Schnitt, S., Sellers, W.R., and Polyak, K. 2004. Molecular characterization of the tumor microenvironment in breast cancer. *Cancer Cell* 6:17-32.

Angelastro, J.M., Kenzelmann, M., and Muhlemann, K. 1999. Substantially enhanced cloning efficiency of SAGE (serial analysis of gene expression) by adding a heating step to the original protocol. *Nucl. Acids Res.* 27:917-918.

Audic, S. and Claverie, J. M. 1997. The significance of digital gene expression profiles. *Genome Res.* 7:986-995.

Blackshaw, S., Harpavat, S., Trimarchi, J., Cai, L., Huang, H., Kuo, W.P., Weber, G., Lee, K., Fraioli, R.E., Cho, S.H., Yung, R., Asch, E., Wong, W.H., and Cepko, C.L. 2004. Genomic analysis of mouse retinal development. *PLoS Biol.* 2:E247.

Boon, K., Osorio, E.C., Greenhut, S.F., Schaefer, C.F., Shoemaker, J., Polyak, K., Morin, P.J., Beutow, K.H., Strausberg, R.L., De Souza, S.J., Riggins, G.J. 2002. An anatomy of normal and malignant gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 99:11287-11292.

Cai, L., Huang, H., Blackshaw, S., Liu, J.S., Cepko, C., and Wong, W.H. 2004. Clustering analysis of SAGE data using a Poisson approach *Genome Biol.* 5(7) R51.

Datson, N.A., van der Perk-de Jong, J., van den Berg, M.P., de Kloet, E.R., and Vreugdenhil, E. 1999. MicroSAGE: A modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucl. Acids Res.* 27:1300-1307.

Gorski, S.M., Chittaranjan, S., Pleasance, E.D., Freedman, J.D., Anderson, C.L., Varhol, R.J., Coughlin, S.M., Zuyderduyn, S.D., Jones, S.J., and Marra, M.A. 2003. A SAGE approach to discovery of genes involved in autophagic cell death. *Curr. Biol.* 13:358-363.

Halascheck-Wiener, J., Khattra, J.S., McKay, S., Pouzyrev, A., Stott, J.M., Yang, G.S., Holt, R.A., Jones, S.J., Marra, M.A., Brooks-Wilson, A.R., and Riddle, D.L. 2005. Analysis of long-lived C. elegans *daf-2* mutants using serial analysis of gene expression. *Genome Res.* 15:603-615.

Hartigan, J. 1975. Clustering Algorithms. Wiley, New York and London.

Hosack, D.A., Dennis, G., Jr., Sherman, B.T., Lane, H.C., and Lempicki, R.A. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol.* 4(10) R70.

Klimaschewski, L., Tang, S., Vitolo, O.V., Weissman, T.A., Donlin, L.T., Shelanski, M.L., and Greene, L.A. 2000. Identification of diverse nerve growth factor-regulated genes by serial analysis of gene expression (SAGE) profiling. *Proc. Natl. Acad. Sci. U.S.A.* 97:10424-10429.

Lepourcelet, M., Tou, L., Cai, L., Sawada, J., Lazar, A.J., Glickman, J.N., Williamson, J.A., Everett, A.D., Redston, M., Fox, E.A., Nakatani, Y., and Shivdasani, R.A. 2005. Insights into developmental mechanisms and cancers in the mammalian intestine derived from serial analysis of gene expression and study of the hepatoma-derived growth factor (HDGF). *Development* 132:415-427.

Neilson, L., Andalibi, A., Kang, D., Coutifaris, C., Strauss, J.F. 3rd, Stanton, J.A., and Green, D.P. 2000. Molecular phenotype of the human oocyte by PCR-SAGE. *Genomics* 63:13-24.

Polyak, K., Xia, Y., Zweier, J.L., Kinzler, K., and Vogelstein, B. 1997. A model for p53 induced apoptosis. *Nature* 389:300-305.

St. Croix, B., Rago, C., Velculescu, V., Traverso, G., Romans, K.E., Montgomery, E., Lal, A., Riggins, G.J., Lengauer, C., Vogelstein, B., and Kinzler, K.W. 2000. Genes expressed in human tumor endothelium. *Science* 289:1197-1202.

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

Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Hieter, P., Vogelstein, B., and Kinzler, K.W. 1997. Characterization of the yeast transcriptome. *Cell* 88:243-251.

Virlon, B., Cheval, L., Buhler, J.M., Billon, E., Doucet, A., and Elalouf, J.M. 1999. Serial microanalysis of renal transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* 96:15286-15291.

Zhang, L., Zhou, W., Velculescu, V.E., Kern, S.E., Hruban, R.H., Hamilton, S.R., Vogelstein, B., and Kinzler, K.W. 1997. Gene expression profiles in normal and cancer cells. *Science* 276:1268-1272.

Internet Resources

http://www.sagenet.org

SAGEnet. Contains instructions for obtaining SAGE analysis software, downloadable SAGE libraries from human, mouse and yeast, and a comprehensive bibliography of SAGE papers.

http://www.ncbi.nlm.nih.gov/SAGE

Serial analysis of gene expression at NCBI.

http://www.ncbi.nlm.nih.gov/CGAP

Cancer Genome Anatomy project. Contains full downloadable predicted tag data for human, mouse, rat, zebrafish, and cow. Also contains a large number of downloadable human SAGE libraries (containing >3.5 million total tags), as well as tools for submitting SAGE data for public access and tools for searching tag abundance levels in the publicly available human SAGE data.

http://www.umich.edu/˜ehm/eSAGE

eSAGE at University of Michigan. Helpful software for SAGE data analysis.

http://www.invitrogen.com

iSAGE at Invitrogen. Integrated kit and software package for conducting microSAGE. The protocol used is very similar to the one described here.

http://arep.med.harvard.edu/labgc/adnan/projects/Utilities/mergesagetags.html

Merge SAGE tags at Harvard Medical School. Helpful tool for merging SAGE data files and downloaded predicted tag identify files (from NCBI).

Contributed by Seth Blackshaw
Johns Hopkins University School of Medicine, Baltimore, Maryland

Brad St. Croix
National Cancer Institute
Frederick, Maryland

Kornelia Polyak
Dana-Farber Cancer Institute
Boston, Massachusetts

Jae Bum Kim
Brigham and Women’s Hospital
Boston, Massachusetts

Li Cai
Rutgers University
Piscataway, New Jersey

APPENDIX: ALGORITHM FOR POISSON-BASED SIGNIFICANCE ANALYSIS

In a SAGE experiment, a set of transcripts from a cell or tissue is sampled for tag extraction. Considering the numerous types of transcripts present in a cell or tissue and the small probability of sampling a particular type of transcript, the authors assume that the number of sampled transcripts of each type is approximately Poisson distributed. Statistically, when this actual sampling process is random enough, Poisson would be the most practical and reasonable assumption compared to other probability models. This assumption leads to the following probability models used for significance analysis and clustering analysis of SAGE data.

Based on Poisson assumption, the authors developed a significance analysis algorithm ("SA algorithm") to detect differentially expressed tags in SAGE data. The input to the SA algorithm is a tab-delimited file containing multiple sage libraries. The SA algorithm can simultaneously compare two or more SAGE libraries. The output of SA algorithm is a set of *P*-values of tests for the significance of the difference in gene expression. Genes with significantly small *P*-values are identified as differentially expressed across different libraries. The *P*-values are calculated in the following way:

Letting X_{ij} be the number of copies of tag *i* in library *j*, three sums are defined:

$$M_{i.} = \sum_{j} X_{ij}; \text{ The sum of counts of tag } i \text{ over all libraries.}$$

$$M_{.j} = \sum_{i} X_{ij}; \text{ The sum of tag counts over all tags in library } j.$$

$$M = \sum_{i,j} X_{ij}; \text{ The sum of tag counts over all tags and libraries.}$$

Equation 11.7.1

Under the null hypothesis that there is no expression difference across libraries, M_iM_j/M copies are then expected to be observed for tag *i* in library *j*. Further, considering that the tags are extracted from a random sample of transcripts in cell, it is reasonable to assume X_{ij} is Poisson distributed with means $\lambda_{ij} = M_iM_j/M$.

The χ^2 statistic is used to test the deviation of observed counts from expected counts:

$$TS_i = \sum_{j=1}^{k} \frac{(X_{ij} - \lambda_{ij})^2}{\lambda_{ij}}$$

Equation 11.7.2

Equation 7.11.2 where *k* is the number of libraries compared.

When *k* is large or λ_{ij} is not small (<5), TS_i is approximately χ^2 distributed with degree of freedom of $k-1$ ($\chi^2{}_{k-1}$), the SA algorithm calculates the *P*-values using the approximated $\chi^2{}_{k-1}$. However, when *k* and λ_{ij} are small, there is a large departure of TS_i from $\chi^2{}_{k-1}$, the SA algorithm calculates exact *P*-value of observed TS_i based on the Poisson distribution of X_{ij}.