Published in final edited form as:
Curr Protoc Mol Biol. 2011 January ; 0 25: Unit25B.9. doi:10.1002/0471142727.mb25b09s93.

Quantification of gene transcripts with Deep Sequencing Analysis of Gene Expression (DSAGE) from 1-2μg total RNA

DC Christodoulou[1], **JM Gorham**, **M Kawana**, **SR DePalma**, **DS Herman**, and **H Wakimoto**
[1]Harvard Medical School, Boston, Massachusetts, USA.

Abstract

Deep sequencing analysis of gene expression (DSAGE) measures global gene transcript levels by massively parallel sequencing of cDNA tags, using 1-2μg total RNA. 21-bp cDNA tags are generated by NlaIII digestion of the cDNA, followed by MmeI cleavage offset from the NlaIII site. cDNA tags are then queried by massively parallel sequencing and aligned to a reference genome and transcriptome, or any available gene sequences, using Bowtie, an ultra high-throughput short-read aligner, and Tophat, a fast splice-junction mapper. Analysis of 10-20-million tags, acquired using one lane of an Illumina Genome Analyzer II, provides sufficient depth to quantify gene expression and detect rare transcripts. Typically, we observe the expression of 15,000 genes in the cardiac left ventricle, including gene transcripts expressed as low as one copy per cell. These expression profiles are highly reproducible (r>0.99 between technical replicates), enabling sensitive detection of differences between experimental conditions as well as assessment of relative transcript abundance between different genes. The significance of these differences is assessed, while accounting for multiple comparisons, using a false discovery rate approach (Audic and Claverie, 1997). Thus, DSAGE can be used to quantify gene expression profiles and assess differential expression with high sensitivity requiring small amounts of biological material.

INTRODUCTION

This protocol describes the construction of 21bp cDNA tag libraries appropriate for massively-parallel sequencing (see Basic Protocol 1) and the analysis of the resulting sequence data. The adapter oligonucleotides used in the protocol are optimized for sequencing with current Illumina massively-parallel sequencers. A step-by-step implementation of the analysis protocol is described (see Basic Protocol 2), which is compiled into 3 steps (Alternate Analysis Protocol 1).

BASIC PROTOCOL 1

DSAGE Library Construction

DSAGE library construction begins with the extraction of 1-2 μg high quality RNA. If it is desirable to minimize biological noise, RNA can be pooled from 3-5 sample replicates. RNA can be extracted using the RNAeasy kit (Qiagen) or the standard protocol using Trizol (Invitrogen). The quality of the RNA can be assessed by running an aliquot on a Bioanalyzer (Agilent) (Schroeder et al., 2006) or an agarose gel. Good quality RNA has a ~2:1 ratio of the 28S and 18S RNA.

An overview of the subsequent steps is presented in Figure 1. Poly(A) RNA is captured on Dynal(dT) beads. Reverse transcription, second strand synthesis, NlaIII digestion, and adapter 1 ligation are performed on the beads. NlaIII cleaves cDNA at defined sites and subsequent washes of the beads remove all but the 3′ fragment of the cDNA. Adapter 1 is then ligated to the overhang resulting from the NlaIII digestion, creating an MmeI site at the overlap of the NlaIII site (cDNA) and the adapter sequence. MmeI cuts 18/20-bp downstream of its recognition site, releasing a 21bp tag with a 2-bp overhang attached to adapter 1. Adapter 2 is then ligated to the released fragment, and the resulting molecules are PCR amplified. To avoid amplification bias, a mock reaction is monitored by real-time PCR and the subsequent library amplification is performed within the log phase of the reaction. The PCR product is then purified, quantified and submitted for sequencing at a concentration of 10 nM of library molecules. The sequencing is typically performed by a core facility and is described in detail in Bentley et al. (2008).

The adapter oligonucleotides are designed to work with standard Illumina Genome Analyzer flow cells and amplification primers. We describe the use of a sequencing primer (Reagents and Solutions) that overlaps the common NlaIII site and reduces the required cycles needed for sequencing to 17.

Materials—Trizol (Invitrogen, cat#15596-018) or RNeasy kit (Qiagen, cat#74104)

Dynabeads mRNA DIRECT Kit (Invitrogen, cat#610-12)

DynaMag-2 magnet (Invitrogen, cat#123-21D)

SuperScript II Reverse Transcriptase (Invitrogen, cat#18064-022)

Second-Strand Buffer (Invitrogen, cat#10812-014)

E. coli DNA Polymerase I (Invitrogen, cat#18010025)

E. coli DNA Ligase (Invitrogen, cat#18052019)

E. coli Rnase H (Invitrogen, cat#18021071)

NlaIII (New England Biolabs, cat#R0125L)

MmeI (New England Biolabs, cat#R0637L)

100X BSA (New England Biolabs, cat#B9001S)

T4 DNA Ligase (high concentration) (Invitrogen, cat#15224-041)

Novex 20% TBE gels(Invitrogen, cat#EC6315BOX)

Low Molecular Weight DNA ladder (New England Biolabs, cat#N3233S)

SYBR Green I (Invitrogen, cat#S7563)

Phenol:Chloroform (adjust to pH7.5-8) (Ambion, cat#AM9732)

Spin-X Centrifuge Tube Filters (Corning, cat#8161)

Glycogen (Roche, cat#10901393001)

GlycoBlue (Ambion, cat#AM9515)

Non-Stick RNase-free Microfuge Tubes, 1.5ml (Ambion, cat#12450

Non-Stick RNase-free Microfuge Tubes, 0.5ml (Ambion, cat#12350)

Platinum Taq DNA polymerase (Invitrogen, cat#10966034)

Qubit fluorometer (Invitrogen, cat#Q32857)

Quant-iT dsDNA HS Assay Kit (Invitrogen, cat#Q32851)

2X Bind and Wash (BW) Buffer (Reagents and solutions)

Buffer C (Reagents and Solutions)

Buffer D (Reagents and Solutions)

Buffer 4 (Reagents and Solutions)

LoTE (Reagents and Solutions)

MGB Buffer (Reagents and Solutions)

Oligonucleotides (Reagents and Solutions) (Integrated DNA Technologies)

Stage I: RNA capture, cDNA and Second Strand synthesis on the beads

1. Transfer 100µl of the oligo(dT) beads (Dynabeads) to a 1.5ml tube (non-stick tubes are used throughout the experiment). Capture the beads by placing the tube on the DynaMag-2 magnet, pipette out the buffer and quickly resuspend the beads in 500µl Lysis buffer (Dynabeads kit). Add the RNA and incubate for 10 minutes with gentle agitation.

2. Capture the beads with the bound RNA on the magnet and remove the supernatant. Perform the following washes:

   Wash 2x with 500µl Buffer A (Dynabeads kit)

   Wash 1x with 500µl Buffer B (Dynabeads kit)

   Wash 2x with 200µl 1X First Strand Buffer (SuperScript II kit)

   Resuspend the beads in the following mix:

   18µl 5X First Strand buffer

   9µl 0.1M DTT

   4.5µl 10mM dNTP

55.5µl ddH_2O

Vortex, spin down and keep on ice.

3. Place the tubes at 37°C for 2 minutes and add 3µl of Superscript Reverse Transcriptase. Vortex gently and incubate for 1 hour. Mix beads every 20 minutes by gentle vortexing.

4. Place the tubes on ice and add the following components in the order shown:

465µl pre-chilled ddH_2O

150µl of 5X Second Strand Buffer

15µl 10mM dNTP

5µl E. coli DNA ligase

20µl E. coli DNA pol I

5µl E. coli RNAse H

Incubate at 16°C in a water bath for 2 hours. Vortex gently every 20 minutes.

In the meantime: Pre-heat Buffer C (Reagents and Solutions) to 75°C and prepare the NlaIII mix for each sample:

172µl LoTE

2µl 100X BSA

20µl 10X NEB buffer 4 (New England Biolabs)

Keep above mix on ice.

5. Following second strand synthesis, add 45µl of 0.5M EDTA to stop the reaction. Capture the beads on the magnet and remove the reaction liquid.

6. Resuspend the beads in the pre-heated buffer C (450µl) and incubate at 75°C for 12 minutes (with intermittent gentle vortexing).

7. Capture the beads and wash once more with preheated Buffer C. In these conditions, beads tend to clump on the tube. If this occurs, scrape off the clumped beads from the tube using a pipette tip and proceed quickly since SDS tends to precipitate.

8. Capture beads and resuspend in 500µl of Buffer D. Transfer resuspended beads to a new tube and wash twice with 200µl Buffer D, followed by two additional washes of 200µl Buffer 4.

Stage II- Cleavage of cDNA with NlaIII, Adapter 1 ligation on the beads

9. Add 6µl of NlaIII to the prepared NlaIII mix from step 5. Resuspend the beads in the mix and incubate at 37°C for 1 hour.

10. Wash beads 2 times with 450µl Buffer C (preheated to 37°C). Wash quickly since SDS tends to precipitate. Then, wash twice with 200µl Buffer D and transfer the bead

slurry to a new tube. **At this stage the beads can be kept overnight at 4°C rotating (if needed).

11. Wash once more with Buffer D, 2x with 200μl of 1X Ligase Buffer, and 1x with 50μl of 1X Ligase Buffer. At this stage, keep the beads in ligase buffer on ice if multiple samples are being handled.

12. Capture beads and resuspend in the following mix:

11.5μl LoTE

4μl 5X DNA Ligase Buffer

2μl Adapter 1 (50μM)

Incubate for 2 minutes at a 50°C heating block followed by a 10 minute incubation at room temperature, then place the tubes on ice.

13. Add 2.5μl of T4 DNA Ligase (high conc.) and incubate in a 16°C water bath for 2 hours. Mix beads every 20 minutes by gentle vortexing.

Stage III – Digestion with MmeI, Adapter 2 ligation, and gel purification—In the meantime: Dilute the 32mM SAM (provided with MmeI) in NEB buffer 4 and prepare the MmeI mix (make both fresh):

10X SAM:

3μl 32mM SAM

24μl 10X NEB buffer 4

213μl ddH$_2$O

Prepare MmeI mix (per 1 sample):

20μl 10X NEB buffer 4

20μl 10X SAM (from above)

150μl ddH$_2$O

14. After ligation, wash the beads 3x with 200μl of Buffer D, resuspend in 200μl Buffer 4 and transfer the bead slurry to a new tube. Wash once more with Buffer 4.

15. Preheat the MmeI mix for 2 minutes at 37°C. Capture the beads on the magnet and discard Buffer 4. Quickly add 10μl of MmeI enzyme to the preheated MmeI mix and use the mix (160μl/sample) to resuspend the beads. Vortex gently and incubate for 2 hours at 37°C. Mix beads every 20 minutes by gentle vortexing.

16. Centrifuge at 14,000 rpm at room temperature for 2 minutes. Transfer the supernatant to a new tube. In order to collect the residual material, add 100μl of LoTE to the beads, mix, and recentrifuge. Place on the magnet and add the 100μl supernatant to the previous tube to a total of 300μl. Discard the beads.

17. To assess contamination, remove a 40µl aliquot from one sample to a new tube (label “ligase minus control”) and adjust the volume to 300µl with LoTE. Process this sample identically until Step 21.

18. Add 300µl of Phenol-Chloroform to each sample. Vortex and centrifuge at 14,000 for 2 minutes. Transfer aqueous phase to new tubes.

19. To precipitate add 2µl of Glycoblue, 133µl of 7.5M ammonium acetate and mix. Add 1000µl of 100% EtOH, vortex immediately and precipitate at −80°C for at least 30 minutes. **At this stage the sample can be kept at −80°C overnight or longer as needed.

20. Centrifuge for 30 minutes at 4°C. Wash 2x with 500µl 70% ethanol (cold). Allow the pellet to air dry and resuspend in 2µl of LoTE.

21. Add the following to the library and mix by pipetting:

2µl 5X Ligase Buffer

2µl Adapter 2 (50µM)

2µl ddH_2O

2µl T4 DNA Ligase (high conc.)

Incubate overnight at 16°C in a thermocycler.

Add water instead of ligase to the “ligase minus control”. After ligation, amplify for 40 cycles as in Step 30. Run on a 4% agarose gel to assess contamination. If an 88bp band is present, follow steps in the Troubleshooting table.

22. Add the following to the ligation mix:

2µl LoTE

3µl 5x Gel Loading Buffer

Load to 1 lane of 20% TBE gel and run until the blue dye front reaches the bottom of the box (approximately 2 hours). Use the Low molecular weight DNA ladder in a separate lane.

23. Stain with SYBR green for 15 minutes. Visualize with UV and cut out the gel band corresponding to 66-bp (Figure 2A).

24. Crush the extracted gel by spinning at 14,000 rpm for 6 minutes through a 0.5ml tube with a hole in the bottom punctured by an 18G needle. Collect crushed gel in a 1.5ml tube.

25. Add 250µl LoTE and 50µl 7.5M ammonium acetate. Vortex each tube and place on a 60°C heating block for 15 minutes. Pre-wet a SpinX column with 5µl of LoTE and add the sample. Centrifuge for 5 minutes at 14,000 rpm at room temperature.

26. Add 3µl of glycogen (Ambion) and 133µl of 7.5M ammonium acetate. Mix and precipitate with 1ml 100% ethanol at −80°C.

**The sample can be stored at −80°C overnight or longer as needed.

27. Centrifuge for 15 minutes at 14,000 rpm at 4°C. Wash 2x with 500μl of 70% ethanol, air dry pellet and resuspend in 14μl LoTE.

**Library can be stored at −20°C.

Stage IV – Library amplification

28. To determine the number of cycles, prepare a mock reaction for qPCR. For each sample assemble:

5μl 10X MGB PCR Buffer

2.5μl DMSO

3μl 10mM dNTP mixture

1μl GexPCR A (100μM)

1μl GexPCR B (100μM)

35μl ddH$_2$O

1μl Platinum Taq (5U/μl)

0.5μl 10x SYBR Green

1μl Library

Cycle conditions:

1st Stage: 94°C for 1′.

2nd Stage: 94°C for 30″; 57°C for 30″; 72°C for 1′. Repeat 24 times.

3rd Stage: 72°C for 5′.

29. Determine cycle number for final amplification. An optimal cycle can be chosen at the mid-upper part of the linear phase of the curve.

30. Perform final amplification:

5μl 10X MGB PCR Buffer

2.5μl DMSO

3μl 10mM dNTP mixture

1μl GexPCR A (100μM)

1μl GexPCR B (100μM)

35.5μl ddH$_2$O

1μl Platinum Taq (5U/μl)

1μl Library

31. Add 10μl of 6X loading dye to the sample and load into 3 lanes of a 20% TBE gel.

Repeat steps 22 to 27 to extract the 88bp band (Figure 2B) and elute in 14μl LoTE. (The added 22bp sequence is used during sequencing.) As before, use the Low molecular weight ladder in a separate lane.

32. Use 1μl with Qubit fluorometer to assess the concentration and submit ~0.6ng/μl for sequencing. Only 17 cycles are needed when the sequencing primer provided in this unit is used.

BASIC PROTOCOL 2

Data Analysis

For the purpose of this analysis, we assume the DNA sequences are available in FASTQ format. To recapitulate the full-length cDNA tag sequence, the NlaIII site is appended to the beginning of DNA sequences. The resulting 21-bp tag is first aligned to the genome and then the transcriptome using Bowtie and Tophat (Langmead et al., 2009; Trapnell et al. 2009). We provide Tophat with a list of known splice-junctions to enable alignment to the transcriptome. These junctions can be generated from a UCSC annotation table (see Support Protocol 1). The number of sequence tags aligned to each gene transcript is then tallied to generate gene expression profiles. Sample profiles are normalized to 1 million gene-aligned tags. Between each pair of samples the fold-difference and significance of the difference are calculated (Audic and Claverie, 1997). An overview of the analysis pipeline is shown in Figure 3.

Materials—The software package requires a Unix computer, and a computing cluster is strongly recommended

(runtime is approximately 1 hour using a 2.4GHz computing node)

Software pre-requisites*:

- Bowtie 0.12.3
- Tophat 1.0.10
- Samtools 0.1.7a
- Bio-SamTools 1.16
- Perl v5.8.8
- BioPerl module for SAGE comparison
- Integrative Genomics Viewer

*Recommended that later or earlier versions be tested

Analysis package includes:

- Analysis programs written in Perl
- Reference files

1. Software pre-requisites installation. Download and install the following programs:

a. Bowtie: http://bowtie-bio.sourceforge.net/index.shtml

   Follow installation instructions and add the program to the PATH

b. Tophat: http://tophat.cbcb.umd.edu/

   Version 10 (recommended) can be downloaded here:

   http://tophat.cbcb.umd.edu/downloads/tophat-1.0.10.tar.gz

   Add the installed program to the PATH

c. Samtools: http://samtools.sourceforge.net/

   Add the installed program to the PATH

d. Bio-Samtools module:

   http://search.cpan.org/~lds/Bio-SamTools/

   Add the module's path to the PERL library environment variable. (PERL5LIB for Perl 5 if the module cannot be installed in the standard BioPerl location.)

e. Perl: http://www.perl.org/

f. BioPerl module for SAGE comparison by algorithm of Audic and Claverie

   http://search.cpan.org/~scottzed/Bio-SAGE-Comparison-1.00/

   Add the module's path to the PERL library environment variable (as above)

g. Integrative Genomics Viewer (IGV): http://www.broadinstitute.org/igv

h. Download the analysis package from this location:

   http://seidman.med.harvard.edu/gs/DSAGE

   Unpack the files in a new directory (substitute '/…/' below with the directory location) 'tar xvfz <downloaded_package>'

   Add the following as environment variables to ~/.bash_profile (assumes the bash shell is used)

   ```
   'export DSAGE_TABLES=/…/DSAGE/tables'
   'export PATH=$PATH:/…/DSAGE/bin'
   source ~/.bash_profile
   ```

2. Append NlaIII sequence

   Run 'appendNlaIII.pl <Fastq_file>'

   The output will be <Fastq_file>.NlaIII

3. For steps 4-7:

   For <SPECIES> use:

If species used is mouse: mm9

If species used is human: hg19

If species used is chicken: galGal3

For <Sample_name> use a one-word description of the sample

4. Align the tags from the appended Fastq file with Tophat allowing 0 (zero) mismatches and by supplying known junctions:

Run Tophat in the same folder (Best if the job is submitted by bsub if a computer cluster is used)

'tophat --solexa1.3-quals -o <Sample_name> --segment-mismatches 0 -j $DSAGE_TABLES/<SPECIES>.juncs <REF_GENOME> <Fastq_file>.NlaIII'

For <REF_GENOME> follow the instructions on the bowtie website to generate or download the genome

The output is a read alignment file (.sam) and a coverage file (.wig) saved in a new folder.

5. Create a sorted binary file (BAM file) and an associated index file for the aligned tags:

First create a temporary unsorted BAM file:

Run 'samtools import $DSAGE_TABLES/<SPECIES>.faidx <input_sam> samfile.tmpbam'

(For Tophat v.10 the <input_sam> is 'accepted_hits.sam')

Sort and name the BAM:

'samtools sort samfile.tmpbam <Sample_name>'

Create an index file:

'samtools index <Sample_name>.bam'

6. At this stage the tags and the tag-depth can be visualized on the UCSC Browser as illustrated in Figure 4.

Normalize, name and compress the wiggle/bedgraph file:

'normwig.pl <input_wig> <Sample_name>'

(For Tophat v.10 the <input_wig> is 'coverage.wig')

The output file is <Sample_name>_norm.wig.gz and can be uploaded to the UCSC

Browser as a custom track. The BAM file can also be uploaded.

For more on how to generate custom tracks, see these instructions: http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html#CustomTracks

For genes that are highly expressed, load SAM or BAM file on IGV.

7. Count sequence tags aligned to each gene transcript to construct the expression profile

   Inside the Tophat folder, run 'countreads.pl <Sorted_BAM_prefix>

   $DSAGE_TABLES/<SPECIES>.ann'

   The output file is "genes.counts" which is the expression profile of the sample

8. Assemble, normalize, and compare expression profiles. Samples sequence tag counts are normalized to a total of one million.

   Run 'expression_analyze.pl <Sample_name_1> dir:<Corresponding_folder_1> <Sample_name_2> dir:<Corresponding_folder_2> […] <Sample_name_N> dir:<Corresponding_folder_N>'

   As shown above, the "dir:" must precede the folder name (no spaces).

   Supplying the directory name is not necessary if the Sample name matches the Tophat folder name and if the command is executed in the directory one level above the Tophat output folders.

   E.g. 'expression_analyze.pl <Sample_name_1> <Sample_name_2>..<Sample_name_N>'

   (Also, the program will work when directories are specified for some of the samples: E.g. 'expression_analyze.pl <Sample_name_1> <Sample_name_2> dir:<Sample_2_corresponding_folder>')

   The output file is called <Sample1_Sample2.._SampleN>.expr.

ALTERNATE ANALYSIS PROTOCOL 1

While a step-by-step approach is needed to test individual components of the process (as when running the analysis pipeline for the first time), to troubleshoot or to make modifications, automation is essential for the analysis of multiple samples as it can save time and minimize errors. This alternate analysis protocol offers tools to automate the generation of expression profiles for multiple samples into a single step.

Materials

Same as in Basic Protocol 2

1. Create a tab-delimited text input file, e.g. 'Samples.txt'

   First column: FASTQ file (full path if not running in the same directory)

   Second column: Sample name

   Third column: Species (e.g. mm9, hg19, galGal3). Name must match the Bowtie index genome.

2. To execute steps 2-7 from Basic Protocol 2:

   ‘DSAGEanalyze_nobsub.pl <Text_file from Step 1>’.

   Example perl script for batching process to linux cluster:

   ‘DSAGEanalyze_bsub.pl <Text_file from Step 1>’

3. For sample comparisons and normalization of the data follow the directions in Step 8 of the Basic Protocol 2.

SUPPORT PROTOCOL 1: MAKING REFERENCE FILES

Materials

Analysis package from Basic Protocol 2

Follow the steps below to make reference files for other other species (or to remake the annotation tables to utilize updated gene annotations and/or genome assemblies)

1. Download a gene annotation table from UCSC (http://genome.ucsc.edu/)

   Click “Tables” link to proceed to the Table Browser.

   - Define genome and assembly

   - Set group to ‘Genes and Gene Prediction Tracks’

   - Set track to ‘UCSC genes’

   - Set table to “knownGene”

   - Set output to “selected fields from primary and related tables”

   - Click “get output” button

   - Check the boxes for the first 10 fields from knownGene (abbreviated: UCSC name,chr,strand,txstart,txend,cdsstart,cdsend,noexons,exon starts, exon ends)

   - Check the box for “geneSymbol” from kgXref

   - Click “get output” button again.

2. To make a junctions file:

   ‘makejunctions.pl <INPUT_UCSC_TABLE> <SPECIES>’

3. To make a gene annotation file for gene expression using Refseq names:

   ‘makeannexpression.pl <INPUT_UCSC_TABLE> <SPECIES>’

4. Move the two generated files to the $DSAGE_TABLES folder.

The <SPECIES> as named here should be used in the Basic Protocol 2 and Alternate Analysis Protocol.

REAGENTS AND SOLUTIONS

Bind and Wash (BW) Buffer (2X) recipe:

200ml 5M NaCl

2.5ml 2M Tris-HCl pH7.5

1ml 0.5M EDTA

296.5ml ddH_2O

Buffer D recipe:

25ml 2x BW Buffer (see above)

1ml 100X BSA (New England Biolabs)

24 ml ddH_2O

Buffer 4 recipe:

5ml NEB 4(New England Biolabs)

1ml 100X BSA (New England Biolabs)

44ml ddH_2O

(stable in 4°C)

Buffer C recipe:

5ml 10% SDS

25ml 2X BW

50μl glycogen

20ml ddH_2O

(stable in 4°C; heat prior use to dissolve SDS)

LoTE recipe:

150μl 1M Tris-HCl (pH 7.5)

20μl 0.5M EDTA

49.83ml ddH_2O

MGB Buffer (10X) recipe:

8.3ml 1M $(NH_4)_2SO_4$

16.75ml 2M Tris-HCl (pH8.8)

3.35ml 1M $MgCl_2$

0.351ml β-mercaptoethanol

21.25ml ddH_2O

*Adapter 1

5′ Phosphate-TCGGACTGTAGAACTCTGAAC-NH_2

5′ ACAGGTTCAGAGTTCTACAGTCCGACATG

*Adapter 2

5′ CAAGCAGAAGACGGCATACGANN

5′ Phosphate-TCGTATGCCGTCTTCTGCTTG- NH_2

PCR Primers

5′ CAAGCAGAAGACGGCATACGA

5′AATGATACGGCGACCACCGACAGGTTCAGAGTTCTACAGTCCGA

Sequencing Primer

5′ CCGACAGGTTCAGAGTTCTACAGTCCGACATG

*For each adapter set, anneal in a thermocycler (in 50μM with LoTE) using the program: 2′ at 95°C, −0.1°C/second to 65°C, 10′ at 65°C, −0.1°C/second to 37°C, 10′ at 37°C, −0.1°C/second to 25°C, 20′ at 25°C, hold at 4°C. The annealed adapters can be stored at −20°C until use.

COMMENTARY

Background

This unit describes the protocol for performing deep sequencing analysis of gene expression (DSAGE), a sequencing-based approach to quantifying gene transcript expression. Analog methods that involve hybridization of cDNA to microarrays of oligonucleotides have been used extensively to assess gene expression. However, such approaches only provide relative measurements with a blunted dynamic range, are limited to a set of known oligonucleotide probe sequences, and are hindered by hybridization biases (Draghici et al., 2006). In contrast, digital sequencing approaches to studies of gene expression involve unbiased counting of observations of gene transcripts, offering quantification and more comprehensive analyses. Early sequencing-based methods, such as Serial Analysis of Gene Expression (SAGE) (Velculescu et al., 1995), were limited in their depth to ~100,000 tags because of high Sanger dideoxy sequencing costs. We previously described Polony Multiplex Analysis of Gene Expression (PMAGE), which utilized a novel massively-parallel sequencing method that facilitated a dramatic increase in the sequencing throughput to 5 million tags and a substantial reduction in cost (Kim et al., 2007). DSAGE was adapted from PMAGE to use the Illumina Genome Analyzer II (Illumina Inc., CA). Currently 15-25 million tags can be sequenced with one lane of Illumina, and this number is solely dependent on the state of the sequencing technology used.

We found that analysis tools developed for RNA-seq, Tophat (Trapnell et al. 2009) and Bowtie (Langmead et al., 2009), efficiently assign DSAGE tags to their corresponding genes. This approach facilitates the integration of DSAGE data with other data types, including RNA-seq and genomic variation, and is able to map tags across exon splice junctions. This approach provides much flexibility in the reference transcriptome to which tags are mapped, permitting study of organisms with incomplete genome or gene sequences. In this protocol we use transcript annotations from the University of California Santa Cruz

(UCSC) genome browser (Kent et al., 2002), downloaded with the UCSC browser retrieval tool (Karolchik et al., 2004).

Critical parameters and Troubleshooting

The key element in constructing a DSAGE library is the starting quality of the RNA. RNA with RNA Integrity Number (Agilent) greater than 8 may be used for optimal results. The yield does not just depend on the quality and quantity of the RNA used, but also on the state of the materials. An indicator of the yield is the presence of the 66bp pre-amplification library which is expected to be visible on the polyacrylamide gel after SYBR green staining. It should be noted that inability to visualize the band should not prompt discontinuation of the library preparation. In that case, the marker may be used to excise the band, carefully avoiding contamination with adapter dimmer.

Contamination of the pre-amplified library with an amplified library can have magnified effects. Since DNA is stable at room temperature, avoiding such contamination requires maintaining a separate lab bench area to be used when processing the amplified library. Similarly the instruments used (including the refrigerator and freezer areas) for the amplified library should be marked as “Post-PCR” and avoided when handling pre-amplified libraries. Any materials can be tested for contamination by running the PCR amplification reaction similarly as performed for the no-ligase control.

The amount of the library following amplification depends on the final amplification step and extraction from the gel. Cycles in the mid-upper part of the log-phase of the reaction should be sufficient to generate enough material to be used for sequencing (the cycles should never exceed the log part of the reaction). Also, special care should be taken to crush the extracted polyacrylamide gel into fine pieces (if crushing was incomplete, the process can be repeated).

While lower yield libraries will require more cycles to amplify to sufficient levels for sequencing, the number of cycles needed for the amplification should not be expected to exceed 15. Libraries requiring more cycles to amplify would display low complexity when sequenced, indicating that the dynamic capture of the low-expressing genes would be compromised. Test the efficiency of the enzymes and purification steps. Note that NlaIII must be kept in −80°C for long-term storage (> 3 months).

The software package pre-requisites must be installed and tested according to the developers’ instructions. Although we recommend using Tophat version 10 as it appears to function best on our cluster with 21bp reads, later versions can also be tested and used.

If the libraries are constructed properly, technical replicates are expected to strongly replicate (with $r > 0.99$). Similar biological samples are also expected to strongly correlate as biological noise can be minimized by pooling RNA from biological replicates. If the expression profiles of such samples do not strongly correlate, the possibilities of poor RNA quality, amplification or complexity of the amplified library, as well as the presence of contamination should be addressed.

As PMAGE and SAGE, DSAGE requires the presence of an NlaIII site on the transcript. An NlaIII site is present on the majority of transcripts, enabling almost complete profiles to be generated. To address that possibility, the mRNA sequence can be retrieved and queried for the NlaIII site.

In this analysis protocol, we strongly recommend that alignment does not allow any mismatches between the tag and the reference genome used; as a consequence, perfect matches are used to construct the transcriptional profiles. A complication of this, however, is in the case of a polymorphism being present on the tag, in which case the tag will not be able to align to the genome. This may become an issue when comparing samples of different biological origin. However unlikely, this can be seen by identifying the NlaIII site on the transcript and identifying polymorphisms using the UCSC genome browser. Alternatively, Bowtie and Tophat could align the reads with allowing 1 or 2 mismatches.

Anticipated Results

In regard to library preparation, the 66bp library is expected to be visible on the polyacrylamide gel after SYBR green staining when the yield of the previous reactions is optimal. For a library with good complexity, less than 15 cycles are required for amplification. The 88bp amplified library should always be visible. Beyond primer dimers, non-specific amplification products should not be present.

Current Illumina sequencers are expected to sequence more than 10 million tags. We observe the expression of greater than 10,000 genes in the mouse heart.

Time considerations

Library preparation is expected to take 4 days. Sequencing is expected to take about 3 days (not taking into account scheduling queues at the sequencing facility). Data analysis should be completed within a day given that runtime is 1 hour per sample and that parallel processing can be used with a computer cluster.

Acknowledgments

The authors are grateful to the principal investigators Christine & Jonathan Seidman. This work was funded by SysCODE Grant #NIH PL1 HL092552, C & J Seidman.

LITERATURE CITED

Audic S, Claverie JM. The significance of digital gene expression profiles. Genome Res. 1997; 7(10): 986–95. [PubMed: 9331369]

Bentley DR, Balasubramanian S, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456(7218):53–9. [PubMed: 18987734]

Blackshaw S, Croix BS, et al. Serial analysis of gene expression (SAGE): experimental method and data analysis. Curr Protoc Mol Biol Chapter. 2007; 25 Unit 25B 6.

Draghici S, Khatri P, et al. Reliability and reproducibility issues in DNA microarray measurements. Trends Genet. 2006; 22(2):101–9. [PubMed: 16380191]

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. Jan 1; 2004 32(Database issue):D493–6. [PubMed: 14681465]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. Genome Res. Jun; 2002 12(6):996–1006. [PubMed: 12045153]

Kim JB, Porreca GJ, et al. Polony multiplex analysis of gene expression (PMAGE) in mouse hypertrophic cardiomyopathy. Science. 2007; 316(5830):1481–4. [PubMed: 17556586]

Langmead B, Trapnell C, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10(3):R25. [PubMed: 19261174]

Li H, Handsaker B, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. [PubMed: 19505943]

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita P, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer L, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer T, Clawson H, Barber GP, Haussler D, Kent WJ. The UCSC Genome Browser database: update 2010. Nucleic Acids Res. Jan; 2010 38(Database issue):D613–9. Epub 2009 Nov 11. [PubMed: 19906737]

Schroeder A, Mueller O, et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol. 2006; 7:3. [PubMed: 16448564]

Trapnell C, Pachter L, et al. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25(9):1105–11. [PubMed: 19289445]

Velculescu VE, Zhang L, et al. Serial analysis of gene expression. Science. 1995; 270(5235):484–7. [PubMed: 7570003]

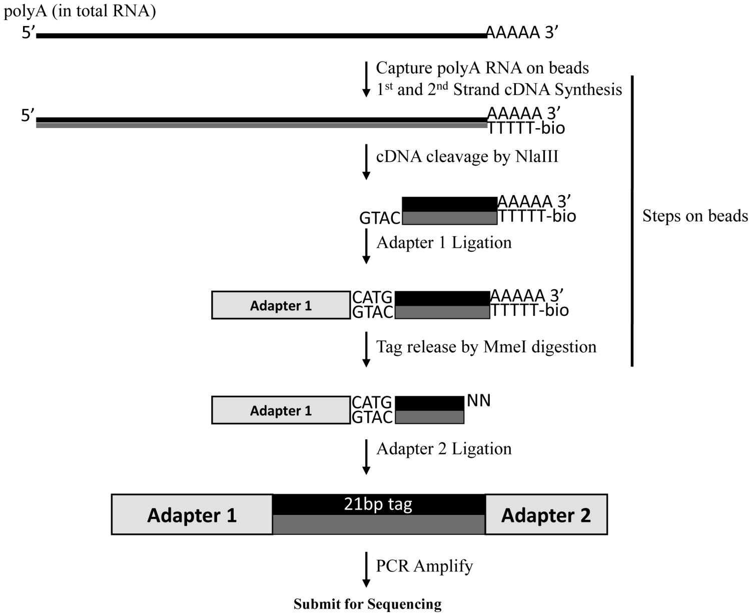

Figure 1.

DSAGE Library construction

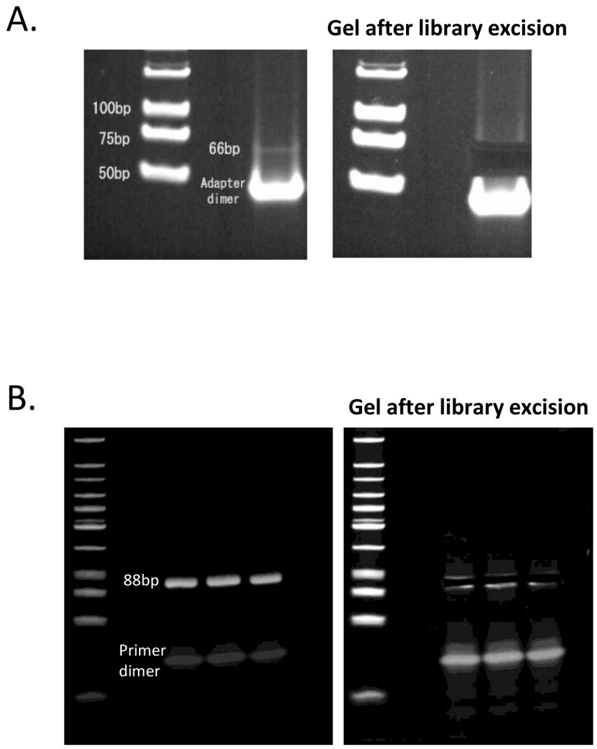

Figure 2.

Library excision from a polyacrylamide gel. The 66bp pre-PCR library (A) and the final amplified 88bp library (B) are shown.

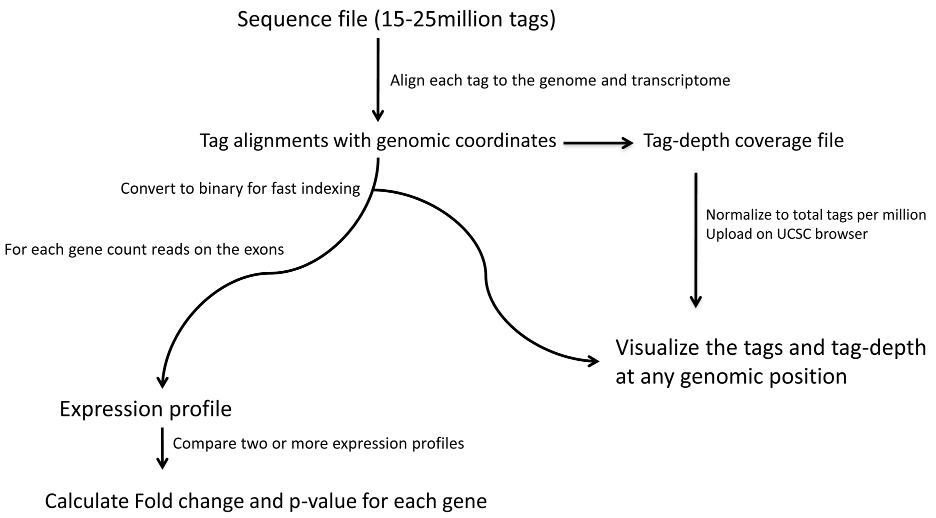
Sequence file (15-25million tags)
Align each tag to the genome and transcriptome
Tag alignments with genomic coordinates
Tag-depth coverage file
Convert to binary for fast indexing
Normalize to total tags per million
Upload on UCSC browser
For each gene count reads on the exons
Visualize the tags and tag-depth at any genomic position
Expression profile
Compare two or more expression profiles
Calculate Fold change and p-value for each gene

Figure 3.

Data analysis pipeline

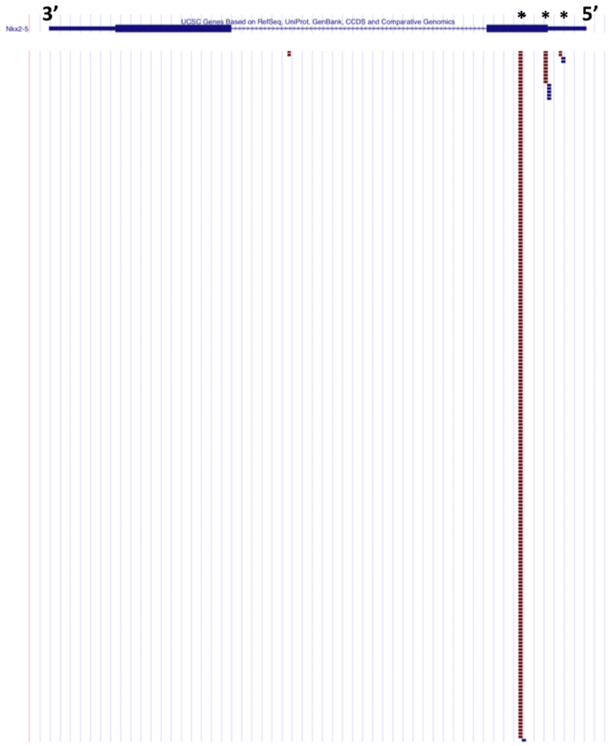

Figure 4.

Tags aligned to the Nkx2-5 locus from mouse left ventricle. Nkx2-5 has three NlaIII sites found in the first exon (denoted by an asterisk). The majority of the sequence tags were aligned to the most 3′ NlaIII site indicating that digestion with NlaIII was nearly complete.

Table 1

Troubleshooting for DSAGE library construction and data analysis

Problem	Possible Cause	Solution
PCR for no-ligase negative control produced an 88bp band	Contamination from another library	Replace affected reagents. Keep a separate bench area (and materials) for pre- and post-amplification reactions
The 66bp pre-amplification library is not visible	Low yield	Use the marker to excise the 66bp.
No signal for low-expressing genes	Low complexity library or insufficient sequence depth	Check activity of enzymes and purification steps. The amount of material before amplification correlates with the complexity of the library. A total of 10-15 cycles should be sufficient for amplification.
The genes in the gene profiles display mostly null values	Genome assembly used for alignment may not match the annotation tables	Make sure that the genome assembly used for alignment matches the assembly of the reference tables (The provided reference tables with the package are for assemblies hg19, mm9, galGal3)
No signal for a gene that is expressed in the tissue	A small number of mRNAs does not have an NlaIII site	Verify that the gene has an NlaIII site by retrieving the mRNA sequence
The gene profiles of similar biological samples do not strongly correlate	Poor RNA quality	Use RNA with RNA Integrity Number > 8
	Library may not be uniformly amplified.	Keep amplification cycles within the exponential range
	Low complexity library Contamination from another library	Increase yield (see above) See above
Some program commands fail to return an output	Software may be incorrectly set-up. Also, insufficient memory, or other problems with the computing node.	Test-run software using a small dataset.