Published in final edited form as:

Methods Cell Biol. 2011 ; 104: 379–399. doi:10.1016/B978-0-12-374814-0.00021-5.

Sequencing-Based Expression Profiling in Zebrafish

Jin Liang, Gabriel Renaud, and Shawn M. Burgess

National Human Genome Research Institute, Bethesda MD, USA, 20892

Abstract

Gene expression profiling is a powerful technique for studying biological processes, especially tissue/organ-specific ones, at the molecular level. With the rapid development of the next-generation sequencing techniques, high throughput sequencing-based expression profiling techniques have been more and more widely adopted in molecular biology studies. In this chapter, we described a protocol for applying one of the sequencing-based expression profiling techniques, Digital Gene Expression (DGE), for zebrafish research. The protocol provides guidelines for wet-bench experimental procedures as well as for bioinformatics-focused data analyses. We also discuss potential issues/challenges with the use of DGE.

I. Introduction

Gene expression profiles of mRNA levels provide us with a comprehensive view of the molecular basis of a cell type, tissue, or organism. The most direct approach for expression profiling is to sequence all the mRNAs in the target tissue/cells to establish the expression profile. The first attempts to do so were by cDNA library or Expressed Sequence Tag (EST) library construction (Adams *et al.*, 1991; Gong, 1998; Lo *et al.*, 2003). These libraries informed the researchers of the presence/absence of specific transcripts in a sample rather than providing a true quantitative expression profile because it was impossible (or at least prohibitively expensive) to achieve the depth of sequence data needed for the quantification of gene expression level. In addition, significant biases in transcript identification were caused by the instability of some cDNA transcripts in *E. coli*. A major advance in quantitative gene expression profiling was the development of Serial Analysis of Gene Expression (SAGE) (Velculescu *et al.*, 1995). In SAGE, the idea was to generate one single short (14–27 bp) fragment (called a “tag”) from each individual mRNA transcript to represent the original transcript. The sequence of the fragment was mapped back to the original transcript based on sequence similarity. The number of times a fragment with the same sequence occurred in a sample serves as a direct measurement of the copy number of the transcript. Although the tag-based profiling strategy was greatly appreciated, it was not widely adopted by average labs mainly due to the high cost of the sequencing process.

Hybridization-based techniques (i.e. microarrays and related techniques) started to gain wide popularity in expression profiling studies. However, there are long-persisting problems with the microarray technique, e.g. cross hybridization, problematic probe design, inter-array/platform inconsistencies, even with the tremendous efforts to optimize microarray technologies over the past ten years.

The amazing progress achieved in sequencing chemistry and automation in the past few years has led to the rapid development of next-generation sequencing technologies (NGS), including Roche 454 (Margulies *et al.*, 2005), Illumina (Bennett, 2004; Bennett *et al.*, 2005),

SOLiD (Shendure *et al.*, 2005), and SMRT (Eid *et al.*, 2009). With the rapidly growing application of NGS technologies, sequencing-based strategies for gene expression profiling become more feasible due to the dramatic increase in sequencing efficiency and the accompanying sharp drop in sequencing costs. One application of the NGS technologies is for quantitative gene expression profiling by sequencing and quantifying short cDNA “tags” generated with SAGE-like protocols. For example, Solexa/Illumina (Digital Gene Expression, DGE) and Applied Biosystems (SOLiD SAGE) have developed their own SAGE-like protocols to generate short tags (21-nt for DGE and 27-nt for SOLiD SAGE) compatible with their sequencing platforms. These protocols are greatly simplified compared to the original SAGE protocol in that they do not require the concatemerization or bacterial cloning of the tags, saving significant time and labor. In this chapter, we will introduce the application of DGE (Illumina) for gene expression profiling in zebrafish.

Digital Gene Expression (also called “tag profiling”) starts with reverse transcription of mRNAs purified from biological samples into double stranded cDNAs. The cDNAs are digested with either *DpnII* or *NlaIII* (corresponding to the anchoring enzyme in the traditional SAGE protocol). The 3'-cDNA fragments are isolated and ligated with a 5'-adaptor (Adaptor I) that contains an *MmeI* recognition site. Another round of digestion with *MmeI* (corresponding to the tagging enzyme in the traditional SAGE protocol) generates cDNA tags of 20-nt long (when using *DpnII*) or 21-nt long (when using *NlaIII*). The tags are then ligated to a 3'-adaptor (Adaptor II). The tag flanked by two adaptors are immobilized onto a solid surface in a flow cell and amplified by bridge PCR. After amplification, the tags are sequenced with an Illumina Genome Analyzer sequencer (sequencing-by-synthesis) and quantified by the number of times any specific tag sequence is detected (“count”). The tag sequences are then mapped to the transcripts they represent to get the final version of a comprehensive gene expression profile with direct quantitative measurement of gene expression level. The workflow of the DGE technique is illustrated in Fig. 1.

II. Digital Gene Expression

A. Sample collection

Collect biological samples according to the questions of interest. The samples can be whole embryos/adults, dissected tissues, cell populations, or FACS-sorted dissociated cells from zebrafish.

B. Total RNA extraction

The total RNA extraction can be carried out with a traditional TRIZOL protocol or with column-based commercial kits. The total RNA yield needs to be checked by an Agilent Bioanalyzer to ensure RNA quality and integrity. The recommended starting quantity of total RNA is 1-2 µg. Protocols modified for minute amounts of starting materials have also been published (Morrissy, 2010).

C. Preparing tags for sequencing

All the reagents listed in Section II.C are contained in I-SAGE Long kit from Invitrogen unless specified otherwise. The process takes several days to complete. Optional pausing points are Steps 26, 31, 35, 51, 53, and 55.

a. mRNA isolation

Materials

- Purified total RNA (1–2 µg) from Section II.B
- Dynal Oligo(dT) magnetic beads

- DynaMag-2 magnet (Invitrogen)
  - Lysis/Binding buffer
  - Wash buffer A
  - Wash buffer B
  - First strand buffer (1X)
1. Dilute purified total RNA to 50 μ L with ultra-pure water into a sterile, nuclease-free microtube.
  2. Incubate the total RNA at 65°C for 5 min and then put it immediately on ice.
  3. Re-suspend the magnetic beads and transfer 100 μ L of bead suspension to a fresh sterile, nuclease-free, siliconized 1.5-mL microtube.
  4. Place the microtube on the magnet for 2 min and carefully remove the supernatant.
  5. Remove the microtube from the magnet and wash the beads twice with 100 μ L Lysis/Binding buffer. Collect the beads and remove the supernatant each time as described in 4.
  6. Re-suspend the beads with 50 μ L Lysis/Binding buffer and add 50 μ L total RNA. Flick the microtube to mix the beads and RNA.
  7. Incubate the mixture on a rocking platform for 5 min at room temperature.
  8. Collect the beads and remove the supernatant as described in Step 4.
  9. Wash the beads twice with 200 μ L Wash buffer A and then twice with 200 μ L Wash buffer B. Separate beads and washing buffer each time as described in Step 4.
  10. Wash the beads four times with 100 μ L First strand buffer (1X). Do not remove supernatant after the fourth wash until Step 11 is finished.

b. First-strand cDNA synthesis

Materials

- First strand buffer (5X)
- dNTPs (10 mM)
- RNaseOUT
- SuperScript II Reverse Transcriptase
- DTT (100 mM)

Equipment

- Thermomixer (Thermomixer R, Eppendorf)
- 11 Prepare the first-strand reaction mixture in a sterile, nuclease-free, siliconized 1.5-mL microtube on ice. The mixture includes
    - 10 μ L First strand buffer (5X)
    - 1 μ L RNaseOUT
    - 5 μ L DTT (100 mM)
    - 2.5 μ L dNTPs (10 mM)

29.5 μ L Ultra-pure water

- 12 Remove the supernatant from the last wash in Step 10. Resuspend the beads with the reaction mixture prepared in Step 11.
- 13 Incubate the bead/reaction mix at 42°C for 2 min, and then add 2 μ L SuperScript II Reverse Transcriptase.
- 14 Incubate the reaction at 42°C in a thermomixer that constantly mixes at 1400 rpm for 1 h.
- 15 Incubate the reaction at 70°C in a thermomixer, with mixing at 1400 rpm for 15 s every 2 min for 15 min. Place the microtube on ice afterwards.

c. Second-strand cDNA synthesis

Materials

- mRNA/cDNA hybrid from Step 15
- Second strand buffer (5X)
- dNTPs (10 mM)
- *E. coli* DNA Polymerase (10 U/ μ L)
- *E. coli* DNA ligase (10 U/ μ L)
- *E. coli* RNase H (2 U/ μ L)
- EDTA (0.5 M)
- Wash buffer C
- Wash buffer D
- Buffer 4 (1X)

Equipment

- Thermomixer (Thermomixer R, Eppendorf)
- 16 Add the following reagents to the mRNA/cDNA hybrid generated in Step 15.
    - 20 μ L ultra pure water
    - 20 μ L Second strand buffer (5X)
    - 3 μ L dNTPs (10 mM)
  - 17 Mix and incubate on ice for 5 min.
  - 18 Add the following reagents.
    - 5 μ L *E. coli* DNA Polymerase (10 U/ μ L)
    - 1 μ L *E. coli* DNA ligase (10 U/ μ L)
    - 1 μ L *E. coli* RNase H (2 U/ μ L)
  - 19 Mix and incubate the reaction at 16°C in a thermomixer, with mixing at 1400 rpm for 15 s every 2 min for 2.5 h. Pre-heat Wash buffer C to 75°C and pre-warm Wash buffer D to room temperature during the incubation.
  - 20 Place the microtube containing double-stranded (ds) cDNA products on ice and add 20 μ L EDTA (0.5 M).

- 21 Use the magnet to remove the supernatant from the beads as described in Step 4. Add 750 μ L pre-warmed Wash buffer C.
- 22 Mix thoroughly and incubate the reaction at 75°C in a thermomixer, with mixing at 1400 rpm for 15 s every minute for 12 min.
- 23 Remove the supernatant and immediately wash the beads again with 750 μ L pre-heated Wash buffer C.
- 24 Wash the beads four times with 750 μ L pre-warmed Wash buffer D.
- 25 Gently re-suspend the beads in 100 μ L Buffer 4 (1X) and transfer the beads and buffer to a fresh nuclease-free sterile 1.5-mL siliconized microtube. Rinse the old microtube with 100 μ L Buffer 4 (1X) and transfer the buffer to the new microtube to make sure all the beads are collected in the new microtube.
- 26 Wash the beads again with 200 μ L Buffer 4 (1X) as described in Step 4 and re-suspend the beads in 88 μ L ultra-pure water. (Alternatively, if needed, the beads can be temporarily stored in Buffer 4 (1X) at 4°C. Remove the buffer with the magnet and re-suspend the beads with 88 μ L ultra-pure water before proceeding to Step 27.)

d. *Nla*III digestion

Materials

- ds cDNA-containing magnetic beads in ultra-pure water (88 μ L) from Step 26
- Bovine serum albumin (BSA, 100X)
- Buffer 4 (10X)
- *Nla*III restriction endonuclease (10 U/ μ L)
- Wash buffer C
- Wash buffer D

Equipment

- Thermomixer (Thermomixer R, Eppendorf)
- 27 Keep the ds cDNA-containing magnetic beads on ice and add the following reagents.
    - 1 μ L BSA (100X)
    - 10 μ L Buffer 4 (10 $\times$)
    - 1 μ L *Nla*III (10 U/ μ L)
  - 28 Mix and incubate the reaction at 37°C in a thermomixer, with mixing at 1400 rpm for 15 s every 2 min for 1 h. Meanwhile, warm Wash buffer C to 37°C and Wash buffer D to room temperature.
  - 29 After incubation, use the magnet to remove the supernatant from the beads.
  - 30 Wash the beads twice with 750 μ L pre-warmed Wash buffer C.
  - 31 Wash the beads four times with 750 μ L pre-warmed Wash buffer D. (Alternatively, beads can be stored in Wash buffer D at 4°C if needed.) Do not remove Wash buffer D from the last wash/storage until Step 32.

e. Adaptor 1 ligation

Materials

- ds cDNA-containing beads in Wash buffer D from Step 31
- Adapter 1* (10 μ M)

$$5' \cdots_p \text{TCGGACTGTAGAACTCTGAAC} \cdots 3'$$

$$3' \text{-GTACAGCCTGACATCTTGAGACTTGGACA-} 5'$$
- Ligase buffer (1X)
- T4 DNA ligase (5 U/ μ L)
- Wash buffer D

Equipment

- Thermomixer (Thermomixer R, Eppendorf)
- 32 Remove supernatant from ds cDNA-containing beads (from Step 31) with magnet and wash the beads with 100 μ L Ligase buffer (1X).
  - 33 Re-suspend the beads with 100 μ L Ligase buffer (1X) and transfer the beads to a fresh sterile nuclease-free 1.5-mL siliconized microtube.
  - 34 Remove the supernatant with magnet and add the following reagents.
    - 41 μ L Ultra-pure water
    - 3 μ L Adapter 1 (10 μ M)
    - 5 μ L Ligase buffer (10X)
    - 1 μ L T4 DNA ligase (5 U/ μ L)
  - 35 Mix and incubate the reaction at 20°C in a thermomixer that is constantly mixing at 1400 rpm for 2 h. Meanwhile, pre-warm Wash buffer D to room temperature. (Alternatively, carry out the ligation overnight at 16°C.)
  - 36 Remove the supernatant with the magnet and wash the beads four times with 500 μ L pre-warmed Wash buffer D. Do not remove the supernatant from the last wash until ready for Step 40.

f. Mmel digestion

Materials

- ds cDNA-containing beads in Wash buffer D from Step 37
- S-adenosylmethionine (SAM, 32 mM, 800X)
- Buffer 4 (10X)
- *MmeI* restriction endonuclease (2 U/ μ L)
- Shrimp alkaline phosphatase (1 U/ μ L, Fermentas Life Sciences)
- Phase lock gel tube (2-mL, heavy; 5 PRIME)
- UltraPure phenol:chloroform:isoamyl alcohol (25:24:1; Invitrogen)

---

*Illumina customers may reproduce and create derivative works of the oligonucleotide sequences but only for use with the Illumina Genome Analyzer and associated assays. All other uses are strictly prohibited.

- Chloroform/isoamyl alcohol (24:1, Sigma-Aldrich)
- Sodium acetate (3 M, pH 5.5; Invitrogen)
- Mussel glycogen (20 mg/mL)
- 100% (pre-cooled to -20°C) and 70% ethanol

Equipment

- Thermomixer (Thermomixer R, Eppendorf)
  - SpeedVac (Thermo Scientific)
- 37** Prepare fresh SAM (10X) by 1:80 dilution of the stock (32 mM, 800X).
  - 38** Prepare fresh Buffer 4 (1X)/SAM (1X) by mixing the following reagents.
    - 800 μ L Ultra-pure water
    - 100 μ L Buffer 4 (10X)
    - 100 μ L freshly prepared SAM (10X)
  - 39** Remove supernatant from ds cDNA-containing beads (from Step 37) with magnet and wash the beads twice with 250 μ L Buffer 4 (1X)/SAM (1X).
  - 40** Add the following reagents to the beads.
    - 76 μ L Ultra-pure water
    - 10 μ L Buffer 4 (10X)
    - 10 μ L freshly prepared SAM (10X, freshly prepared)
    - 4 μ L *MmeI* (2 U/ μ L)
  - 41** Mix and incubate the reaction at 37°C in a thermomixer that constantly mixes at 1400 rpm for 1.5 h.
  - 42** After digestion, place the microtube on the magnet for 2 min and carefully transfer the supernatant to a fresh sterile, nuclease-free, siliconized 1.5-mL microtube.
  - 43** Add 2 μ L shrimp alkaline phosphatase (1 U/ μ L) to the cDNA (supernatant from Step 42) and incubate the reaction at 37°C for 1 h.
  - 44** Spin two Phase lock gel tubes (2-mL, heavy) at 12,000 - 16,000 *g* for 30 s. Add cDNA from Step 43 and 100 μ L UltraPure phenol/chloroform/isoamyl alcohol (25:24:1) to a pre-spun tube.
  - 45** Mix thoroughly and centrifuge at 12,000-16,000 *g* for 5 min to separate the phases. Carefully transfer the top aqueous phase to the other fresh pre-spun Phase lock gel tube (2-mL, heavy) and add 100 μ L chloroform/isoamyl alcohol (24:1).
  - 46** Mix thoroughly and centrifuge at 12,000-16,000 *g* for 5 min to separate the phases. Carefully collect the top aqueous phase to a fresh sterile, nuclease-free, siliconized 1.5-mL microtube.
  - 47** Precipitate the cDNA (the aqueous phase collected in Step 46) by adding the following reagents.
    - 1 μ L Mussel glycogen (20 mg/mL)

10 μ L Sodium acetate (3 M)

325 μ L 100% ethanol (pre-cooled to -20°C)

- 48 Mix and centrifuge at 16,000 g for 20 min.
- 49 Remove the supernatant and wash the pellet with 500 μ L room temperature 70% ethanol.
- 50 Air-dry the pellet or use the SpeedVac to dry it.
- 51 Dissolve the pellet in 7 μ L of ultra-pure water. (The cDNA can be stored at -20°C if needed.)

g. Adaptor 2 ligation

Materials

- cDNAs in ultra-pure water (7 μ L) from Step 51
- Adapter 2* (1.5 μ M)

$5' \text{--CAAGCAGAAGACGGCATACGANN--}3'$
 $3' \text{--GTTCTGCTCTTCTGCCGTATGCT}_p\text{---}5'$
- Ligase buffer (10X)
- T4 DNA ligase (5 U/ μ L)
- 52 Put the microtube containing 7 μ L cDNAs (from Step 51) on ice and add the following reagents.

1 μ L Adapter 2 (1.5 μ M)
 1 μ L Ligase buffer (10X)
 1 μ L T4 DNA ligase (5 U/ μ L)
- 53 Mix and incubate the reaction at 20°C for 2 h. (Alternatively, mix and incubate the reaction at 16°C overnight.)

h. Enrichment of the adapter-ligated cDNA fragments with PCR

Materials

- Adapter-ligated cDNA (10 μ L) from Step 53.
- Phusion HF buffer (5X, Finnzymes)
- Phusion hot start high-fidelity DNA polymerase (2 U/ μ L, Finnzymes)
- Primer 1* (25 μ M)

$5' \text{--CAAGCAGAAGACGGCATACGA--}3'$
- Primer 2* (25 μ M)

$5' \text{--}$
 $\text{AATGATACGGCGACCAACCGACAGGTTTCAGAGTTCTACAGTCCGA--}3'$
- dNTP mix (10 mM)

Equipment

- Thermal cycler

- 54** Set up a PCR reaction mixture in a fresh sterile PCR tube with following reagents.
- 34.5 μ L Ultra-pure water
  - 10 μ L Phusion HF buffer (5X)
  - 0.5 μ L Primer 1 (25 μ M)
  - 0.5 μ L Primer 2 (25 μ M)
  - 1.5 μ L dNTP mix (10 μ M)
  - 0.5 μ L Phusion hot start high-fidelity DNA polymerase (2 U/ μ L)
  - 2.5 μ L adaptor-ligated cDNA (from Step 53)
- 55** Amplify the adaptor-ligated cDNA in a thermal cycler using the following protocol.

---

1 cycle:	30 s at 98°C
15 cycles:	10 s at 98°C
	30 s at 60°C
	15 s at 72°C
1 cycle:	10 min at 72°C
1 cycle:	Hold at 4°C

---

i. Purification of the amplified adaptor-ligated cDNA

Materials

- Amplified adaptor-ligated cDNA (50 μ L) from Step 55.
- 10% Pre-cast Novex TBE PAGE gel (1.0 mm, 10 well, Invitrogen)
- Novex TBE running buffer (5X, Invitrogen)
- Novex Hi-Density TBE sample buffer (5X, Invitrogen)
- 25-bp DNA ladder (Invitrogen)
- UltraPure ethidium bromide (10 mg/mL, Invitrogen)
- LoTE
- Ammonium acetate (7.5 M)
- Spin-X filter spin column (Corning Life Sciences)
- Mussel glycogen (20 mg/mL)
- Sodium acetate (3 M, pH 5.5, Invitrogen)
- 100% (pre-cooled to -20°C) and 70% ethanol
- Elution buffer (Qiagen)

Equipment

- XCell SureLock Mini-Cell (Invitrogen)
- Thermomixer (Thermomixer R, Eppendorf)

- UV transilluminator
  - SpeedVac (Thermo Scientific)
- 56** Prepare TBE running buffer (1X) by 1:5 dilution of stock running buffer (5X) with de-ionized distilled water.
  - 57** Put the TBE gel in an XCell SureLock Mini-Cell and add running buffer (1X).
  - 58** Mix 1 μL of the 25-bp DNA ladder with 1 μL of Novex Hi-Density TBE sample buffer (5X) and load into a well of the TBE gel.
  - 59** Mix 50 μL of amplified adaptor-ligated cDNA (from Step 55) with 12.5 μL of sampling buffer (5X) and load into a well of the TBE gel. One lane per sample.
  - 60** Electrophorese at 200 V for 30-40 min. While waiting for the electrophoresis, prepare the gel elution buffer by mixing LoTE and ammonium acetate (7.5 M) at a 5:1 ratio. In addition, puncture the bottom of a sterile, nuclease-free, 0.5-mL microtube with a sterile 21-gauge needle 4-5 times.
  - 61** Retrieve the gel from the cassette after electrophoresis and stain with Ultrapure ethidium bromide in a clean container for 2–3 min.
  - 62** View the stained gel using a UV transilluminator.
  - 63** Use a clean scalpel to excise the 85-bp bands from each gel lane loaded with a cDNA sample. Work quickly to reduce UV exposure time.
  - 64** Put each gel slice into a 0.5-mL-microtube prepared as in Step 60.
  - 65** Place the 0.5-mL microtube (containing the gel slices) into a sterile, nuclease-free, 1.5-mL microtube. Centrifuge the stacked microtubes at full speed for 2 min at 12,000 g to collect the sheared gel pieces at the bottom of the 1.5-mL microtube.
  - 66** Discard the 0.5-mL microtube and add 200 μL of gel elution buffer to the sheared gel pieces in the 1.5-mL microtube.
  - 67** Mix well and incubate the sheared gel/buffer combination at 65°C for 1 h.
  - 68** Transfer the mixture into a Spin-X filter spin column tube and centrifuge at 12,000 g for 5 min. Save the eluate in the bottom tube.
  - 69** Add the following reagents to the eluate (~200 μL).
    - 2 μL Mussel glycogen (20 mg/mL)
    - 20 μL Sodium acetate (3 M, pH 5.5)
    - 650 μL 100% ethanol (pre-cooled to -20°C)
  - 70** Mix and centrifuge at 16,000 g for 20 minutes at 4°C.
  - 71** Carefully remove the supernatant and wash the pellet twice with 500 μL 70% ethanol.
  - 72** Air-dry the pellet or use the SpeedVac to dry it.
  - 73** Re-suspend the pellet in 10 μL Elution buffer. After this step, the tags for sequencing are generated. They can be stored at -20°C until sequencing.

D. Tag sequencing

Check the final concentration and quality of the tags (from Step 73 in Section II.C) with an Agilent Bioanalyzer. Dilute the tags with Elution buffer to desired concentrations as needed. Generate clusters and sequence on the Illumina Genome Analyzer following the recommended protocols from the company.

E. Tag mapping and annotation

Upon completion of the sequencing and subsequent base-calling, millions of short tag sequences need to be analyzed through a bioinformatics pipeline to identify the genes that the tags represent.

a. Tag mapping—A bioinformatics pipeline to implement the tag mapping can be designed with two distinct strategies in mind: 1) Map the tags to the high quality transcriptomic databases first, then map the tags that fail to match the transcriptome data to more inclusive transcriptomic databases and, finally, map the tags without any previous hit to the zebrafish genome; or 2) Map all the tags in parallel to every transcriptomic database as well as the zebrafish genome and parse the results.

While the first approach makes subsequent analysis simpler, the second approach is more easily implemented by bioinformatics programmers because they do not have to filter tags that failed to map at each step of the serial strategy. A second advantage of the parallel analysis is that it reports hits to various databases for each tag. The parallel process involves mapping all the tag sequences to databases containing: 1) well-curated mRNAs (e.g. NCBI Refseq RNAs and Vega cDNAs); 2) sequenced cDNAs and ESTs (e.g. NCBI UniGene and Ensembl cDNAs); 3) predicted transcripts (e.g. Ensembl *ab initio*); and 4) whole-genome sequences. The subsequent mapping results report the matches without any pre-filtering (Fig. 2A). The follow-up analysis can easily be done using a scripting language like Perl or Python.

The mapping step involves identifying all occurrences of a large set of short tags in a database of long contiguous sequences like mRNA transcripts or chromosomes. The sequence of a tag can be dichotomized into a short prefix, identical to all tags that result from the use of a given restriction enzyme and corresponding to its recognition sequence, and a divergent longer suffix. While alignment of numerous short sequence tags has received great attention in recent years from the computer science community, standard alignment algorithms do not normally distinguish between mismatches occurring in the first few nucleotides or in the remainder of the sequence. This can be important in tag profiling studies as mismatches occurring in the recognition site may not carry the same weight as mismatches in the remainder of the tag. An alignment program to take this into account can be readily implemented in C++. A key observation to creating an efficient aligner is that any match to the target database will be anchored to the recognition sequence, and this sequence is identical on the reverse strand when read from 5' to 3'. Hence, the database needs only to be scanned once and, upon finding the recognition site, the divergent suffixes can be found either in the sequence following it or on the reverse strand in the sequence preceding it (Fig. 3A). Many string search algorithms (Knuth *et al.*, 1977; Boyer and Moore, 1997) allow an efficient scan of the genome using a short recognition site and some even tolerate mismatches (Baeza-Yates and Gonnet, 1994).

Once the recognition site sequence (i.e. the restriction site sequence) is found in the database, the expression tags need to be searched to identify those having a divergent suffix corresponding to one of the sequences flanking the restriction site. While a linear search through the tags is easy to implement, it would be too time-consuming if the recognition site

is quite common and triggers frequent searches. A faster strategy is to use a prefix tree (Fig. 3B), referred to as “trie” (pronounced “tree”; Black, 2009) in the literature, first described by Fredkin (1960) to perform the searches. Each node in the tree represents a character at a given position in the strings and contains pointers to various child nodes containing the next characters in the sequences for the following position. However, due to the nature of the sequence tags, each internal node will have at most 5 child nodes each representing [A,C,G,T,N]. Once the end of the sequence is reached, the nodes containing the final characters are called leaf nodes and do not possess any child node. Instead, they contain a string corresponding to the identifier of the input sequence tags represented by the path leading up to that leaf node. Once the trie of all the divergent suffixes has been built, the sequences flanking both sides of the recognition site present in the database can be used to query the trie to find tags that potentially match. If no mismatches are allowed, a simple recursive search that keeps track of the flanking sequence left to search starting with the first character at the root node can be used. However, if mismatches are allowed then numerous recursive calls must be launched on child nodes. Each call must keep track of the number of mismatches found thus far. If a recursive call encounters an unresolved base pair or a character matching the current one in the flanking sequence, no mismatches are discounted and the search is launched recursively on the child nodes. If, on the other hand, a mismatch between the two aforementioned characters is found, the search must deduct a mismatch from the current number of allowed mismatches and, should they fall to -1, the search must stop. This scheme allows the trie to be efficiently searched even with mismatches while safely overlooking a large portion of the set of divergent suffixes.

By combining both the string search algorithm for the recognition site and the trie search for the divergent suffixes, the tags can be mapped to zero, one, or multiple transcript/genomic locations. Once the input tags have been mapped to various transcriptomic and genomic databases, a script can be used to produce a tab-delimited file detailing the results for each tag.

b. Tag annotation—When annotating a tag sequence, i.e. assigning a specific gene/genome location to the tag sequence, we adopted a hierarchical process (Fig. 2B). If a tag sequence can be unambiguously mapped to a known/predicted transcript, it will be annotated with the gene associated with the transcript. Genomic mapping is only considered when no match can be found to a known or predicted transcript. When mismatch is allowed during the mapping process, priority is usually given to no-mismatch mappings for tag annotation. However, it is up to the researcher whether mismatch mappings to known/predicted transcripts should have higher priorities than no-mismatch mappings to intergenic regions on the genome. For example, Fig. 2B shows an example where the mismatch transcriptomic mappings are considered before the no-mismatch genomic mappings because we want to maximize the proportion of tags annotated with genes rather than genomic locations. In addition, some tags may have been mapped to multiple transcripts/genomic locations. In that case, annotation of the tags will fail unless the tags can be unambiguously assigned with other algorithms (see Section III.B for details).

F. Data analysis and confirmation

Once gene identities and quantities are assigned, DGE data can be analyzed with commercially available software (e.g. Genesifter and MetaCore) and self-developed analysis pipelines (e.g. modules in Bioconductors). Statistical analyses are often carried out to compare different expression profiles. Most statistical analyses require normalization of the count numbers in different profiles. The most straightforward approach is to normalize the count number of each gene with respect to the total count number of the whole profile, and the normalized gene expression level is described in units of TPM (transcripts per million).

In addition, more sophisticated methods, e.g. normalizing over the count numbers of one or more house-keeping genes or over control spiked-in RNAs, can also be adopted if necessary. The choice of the statistical test usually depends on the experimental design. Although the t-test and ANOVA can be adapted for studies designed with replicates, Fisher's exact test and the χ^2 -test will be more suitable when replicates are not planned or hard to achieve. Measurement of gene expression levels also need to be further validated with correlative qRT-PCR experiments.

III. Potential issues in DGE

A. Experimental design

When designing a DGE experiment, there are several issues that need to be carefully considered depending on the biological question examined. First, how many biological and technical replications are needed to provide enough sampling events for statistical analysis? Second, how much biological tissue is needed, and how will the tissues be collected? Given the high sensitivity of DGE achieved through ultra-deep sequencing, minor sample contamination (e.g. RNA from extraneous tissues) would result in skewed expression profiles in DGE experiments (Heged *et al.*, 2009); such skewing would probably not be detected by less sensitive techniques. It is possible to detect profiles with such "contamination" by comparing biological replications (Heged *et al.*, 2009). The possibility of "contaminated" tissues should also been taken into account when calculating the number of biological replications and the amount of biological tissue collected. If only minute amounts of tissue can be acquired, the DGE protocol needs to be adjusted accordingly. Finally, inter-lane and inter-run variation should be taken into account when planning biological and technical replications on Illumina sequencers (Marioni *et al.*, 2008).

B. Challenges in tag mapping and annotation

A major challenge in using the DGE approach results from the short length of the tags. There are two main types of mapping problems: no mapping and multiple mapping. Frequently, the short sequencing reads cannot be mapped to any location on the genome/transcriptome, which can be attributed to exon-spanning reads, alternative splicing isoforms, post-transcriptional editing, polymorphisms, and sequencing errors. While it has been estimated that more than 75% of the 21-nt tags are expected to occur only once in the human genome (Saha *et al.*, 2002), the chance of mapping DGE tags to a specific gene or genomic location is further reduced by repetitive sequences, polymorphisms (e.g. SNPs), and sequencing errors. When using DGE to profile biological samples from zebrafish, mapping is even more challenging due to the high occurrence of polymorphisms, significant gene duplication, and the lower coverage depth of the zebrafish genome/transcriptome databases when compared to mouse or human resources.

Several studies have attempted to assign multi-location tags to unique genes by calculating the probabilities that those tags derived from different genes (Ge *et al.*, 2006; Malig *et al.*, 2006). Many factors were taken into account in the probability calculation, including the sequence context of the putative mapping loci (e.g. internal polyadenylation sequences) (Malig *et al.*, 2006) and gene expression data from previous SAGE and microarray studies (Ge *et al.*, 2006). In most cases, tags are mapped to the 3'-most anchoring enzyme cutting sites but exceptions have been observed as well (Asmann *et al.*, 2009). While such a method is useful for DGE of organisms with relatively well-annotated transcriptomes, it may introduce bias when the experimental organism has a less well studied transcriptome and mapping is mainly based on EST sequences rather than curated, full-length transcript sequences. Performance of multiple technical replications constrained by minimum

expression level requirements will minimize the effects of sequencing errors/artifacts on tag mapping.

Allowing mismatches will decrease the number of tags that fail in genome/transcriptome mapping. Relaxing match stringency should be considered when using DGE on organisms, such as the zebrafish, which have high rates of genetic polymorphism. However, reduction of false-negative matches is likely to increase false-positive mapping errors. In addition, it is important to take into account the tolerated mismatches and where they occur. If the number of tolerated mismatches is pre-specified for one or more given positions of the string, both restriction enzyme recognition site search and trie search only need to track the mismatch respectively rather than jointly. However, if the number of mismatches is allowed globally across the string regardless of the segment where they occurred, then naturally both procedures need to keep track of mismatches jointly. For instance, if the recognition site is found with 1 mismatch and 2 mismatches are tolerated across the whole string then the search must allow at most 1 mismatch in the trie searching procedure. We suggest that researchers carefully compare the annotation results from no-mismatch and mismatch mapping before choosing one as the mapping strategy. When annotating tags, researchers also need to decide the hierarchical priority of mismatch mapping to a transcript vs. no-mismatch mapping to a genome location (Fig. 2).

C. Tag-based profiling vs. RNA-seq

In addition to DGE, another application of NGS in gene expression profiling is full transcriptome RNA-sequencing (RNA-seq). In RNA-seq, mRNAs are converted into short cDNA fragments which are then subjected to in-depth sequencing. cDNA fragments are then assembled and annotated either using genomic sequences as reference or via *de novo* assembly. RNA-seq results are highly informative because they provide full transcriptome sequences (Sultan *et al.*, 2008; Wilhelm *et al.*, 2008), transcription start/end sites (Guttman *et al.*, 2010; Wilhelm *et al.*, 2010), exon-intron boundaries (Wilhelm *et al.*, 2008; Guttman *et al.*, 2010), splicing patterns (Sultan *et al.*, 2008; Wilhelm *et al.*, 2008; Trapnell *et al.*, 2010), post-transcriptional modifications (Picardi *et al.*, 2010), single nucleotide polymorphisms (SNPs) (Cloonan *et al.*, 2008), and transcript-level quantifications (Marioni *et al.*, 2008; Mortazavi *et al.*, 2008; Wilhelm and Landry, 2009).

Although RNA-seq clearly has advantages over DGE with respect to the accurate mapping of short sequence reads and the characterization of full-length transcripts, DGE, at the moment, is still favored over RNA-seq for generating a quantitative gene expression profile for two reasons. First, RNA-seq requires much greater sequencing depth to obtain adequate coverage of transcripts for measuring gene expression levels. For example, it has been estimated that RNA-seq required at least 40 million reads to cover 90% of a transcriptome library, whereas DGE required no more than 5 million tags (Wold and Myers, 2008). The difference in requisite sequencing depth directly translates into greater sequencing costs when RNA-seq is used for quantification. The other reason lies in the more complicated statistics needed for quantifying gene expression level from RNA-seq data. Because a single transcript gives rise to multiple cDNA fragments “randomly” in RNA-seq, the total number of cDNA fragments mapped to the transcript cannot be used to represent the copy number of the transcript until they are normalized according to the length of the transcript (Marioni *et al.*, 2008). Further normalization may also be required due to the 3' biased coverage of full-length transcripts that results from library preparation.

With further reduction in sequencing costs and optimization of the protocol, RNA-seq will become more applicable for quantitative profiling studies. What then is the future of DGE? Rather than being replaced by RNA-seq, DGE will more likely be applied in the future for routine, inexpensive gene expression profiling. Under current circumstances, a combination

of low-coverage RNA-seq and DGE would allow large-scale quantitative analysis of gene expression profiles at relatively low cost (Wang *et al.*, 2010).

Acknowledgments

This research was supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

References

- Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merrill CR, Wu A, Olde B, Moreno RF, Kerlavage AR, McCombie WR, Venter JC. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*. 1991; 252:1651–1656. [PubMed: 2047873]
- Asmann Y, Klee E, Thompson EA, Perez E, Middha S, Oberg A, Therneau T, Smith DI, Poland GA, Wieben ED, Kocher JP. 3' tag digital gene expression profiling of human brain and universal reference RNA using Illumina Genome Analyzer. *BMC Genomics*. 2009; 10:531. [PubMed: 19917133]
- Baeza-Yates RA, Gonnet GH. Fast string matching with mismatches. *Inf. Comput.* 1994; 108:187–199.
- Bennett S. Solexa Ltd. *Pharmacogenomics*. 2004; 5:433–438. [PubMed: 15165179]
- Bennett ST, Barnes C, Cox A, Davies L, Brown C. Toward the \$1000 human genome. *Pharmacogenomics*. 2005; 6:373–382. [PubMed: 16004555]
- Black, PE., editor. Dictionary of algorithms and data structures. U.S. National Institute of Standards and Technology; 2009. <http://xw2k.nist.gov/dads/>
- Boyer RS, Moore JS. A fast string searching algorithm. *Comm. ACM*. 1977; 20:762–772.
- Cloonan N, Forrest ARR, Kolle G, Gardiner BBA, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Methods*. 2008; 5:613–619. [PubMed: 18516046]
- Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009; 323:133–138. [PubMed: 19023044]
- Fredkin E. Trie Memory. *Commun. ACM*. 1960; 3:490–499.
- Gong Z. Zebrafish Expressed Sequence Tags and their applications. *Methods Cell Biol.* 1998; 60:213–233. [PubMed: 9891340]
- Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A. *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* 2010; 28:503–510. [PubMed: 20436462]
- Heged s Z, Zakrzewska A, Ágoston VC, Ordas A, Rácz P, Mink M, Spaink HP, Meijer AH. Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. *Mol. Immunol.* 2009; 46:2918–2930. [PubMed: 19631987]
- Knuth D, Morris JH Jr, Pratt V. Fast pattern matching in strings. *SIAM J. Comput.* 1977; 6:323–350.
- Lo J, Lee S, Xu M, Liu F, Ruan H, Eun A, He Y, Ma W, Wang W, Wen Z, Peng J. 15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. *Genome Res*. 2003; 13:455–466. [PubMed: 12618376]
- Malig R, Varela C, Agosin E, Melo F. Accurate and unambiguous tag-to-gene mapping in serial analysis of gene expression. *BMC Bioinformatics*. 2006; 7:487. [PubMed: 17083742]

- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005; 437:376–380. [PubMed: 16056220]
- Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*. 2008; 18:1509–1517. [PubMed: 18550803]
- Morrissey S, Zhao Y, Delaney A, Asano J, Dhalla N, Li I, McDonald H, Pandoh P, Prabhu AL, Tam A, Hirst M, Marra M. Digital gene expression by tag sequencing on the illumina genome analyzer. *Curr. Protoc. Hum. Genet*. 2010:1–36.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*. 2008; 5:621–628. [PubMed: 18516045]
- Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G. Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing. *Nucleic Acids Res*. 2010; 38:4755–4767. [PubMed: 20385587]
- Saha S, Sparks AB, Rago C, Akmaev V, Wang CJ, Vogelstein B, Kinzler KW, Velculescu VE. Using the transcriptome to annotate the genome. *Nat. Biotechnol*. 2002; 20:508–512. [PubMed: 11981567]
- Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*. 2005; 309:1728–1732. [PubMed: 16081699]
- Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keefe S, Haas S, Vingron M, Lehrach H, Yaspo ML. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*. 2008; 321:956–960. [PubMed: 18599741]
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol*. 2010; 28:511–515. [PubMed: 20436464]
- Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. *Science*. 1995; 270:484–487. [PubMed: 7570003]
- Wang X, Luan J, Li J, Bao Y, Zhang C, Liu S. *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. *BMC Genomics*. 2010; 11:400. [PubMed: 20573269]
- Wilhelm BT, Landry J. RNA-Seq--quantitative measurement of expression through massively parallel RNA-sequencing. *Methods*. 2009; 48:249–257. [PubMed: 19336255]
- Wilhelm BT, Marguerat S, Goodhead I, Bähler J. Defining transcribed regions using RNA-seq. *Nat. Protoc*. 2010; 5:255–266. [PubMed: 20134426]
- Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*. 2008; 453:1239–1243. [PubMed: 18488015]
- Wold B, Myers RM. Sequence census methods for functional genomics. *Nat. Methods*. 2008; 5:19–21. [PubMed: 18165803]

Figure 1. Steps for gene expression profiling with Digital Gene Expression (DGE) using *NlaIII* digestion.

Figure 2.

Tag mapping and annotation. (A) Tag sequences achieved from the Illumina sequencer are mapped to various transcriptomic and genomic databases by string search followed by trie search. (B) An example of hierarchical strategy to analyze the tag mapping results for tag annotation.

Figure 3.

Restriction enzyme recognition site search and trie search algorithms. (A) Illustration of mapping to the target database of a sequenced tag based on a specific restriction enzyme recognition site. Since the recognition sequence of the restriction enzyme is reverse palindromic, the same sequence can be found on the reverse strand when read from the 5' end to the 3' one. Highlighted are the potential locations where the divergent suffixes could map. (B) A trie representation using nine sequences of four nucleotides each. The common prefixes represented as common paths in the tree make exact string searches quite fast. Searches using mismatches can also be handled using multiple recursive calls where the search time increases with the number of tolerated mismatches.