Digital Gene Expression: Tag Profiling

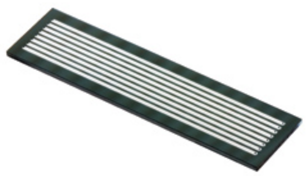
Applying Solexa® technology to gene expression analysis, Illumina introduces a universal whole-genome expression profiling platform that enables transcript discovery and analysis in any organism. Digital Gene Expression Tag Profiling produces high-quality quantitative data empowering whole-genome profiling of any polyadenylated mRNA, detection of rare transcripts, and quantification comparable to quantitative real-time PCR.

INTRODUCTION

Illumina has provided researchers with gene expression microarrays that support high-throughput biological discovery applications. Now, with Digital Gene Expression (DGE) Tag Profiling, Illumina offers a whole-genome expression platform that can generate expression profiles for any transcript from any gene in any organism. DGE Tag Profiling is a revolutionary approach to expression analysis. Driven by Solexa sequencing technology, DGE creates genome-wide expression profiles through sequencing, not hybridization. The ability to identify, quantify, and annotate expressed genes on the level of the whole genome without prior sequence knowledge enables an entirely new scale of biological experimentation, opening doors to higher-confidence target discovery, disease classification, and pathway studies. DGE Tag Profiling also offers researchers a global orthogonal hybridization array validation method and, with an almost unlimited dynamic range, a tunable depth of coverage for rare transcript discovery and quantification.

Unlike the relative expression profiles microarray hybridization technology generates, DGE records the numerical frequency of a sequence in the library population.

FIGURE 1: ILLUMINA'S GENE EXPRESSION PORTFOLIO

Platform	**BeadArray™ Direct Hybridization Assay**	**Digital Gene Expression Tag Profiling**
Application	**High-Throughput Whole-Genome Profiling**	**Novel Transcript Discovery Quantification of Rare Transcripts Orthogonal Microarray Validation**
Format	**BeadChip**	**Flow Cell** 

HIGHLIGHTS FOR DIGITAL GENE EXPRESSION TAG PROFILING

- **No sequence knowledge required:** Universal platform to study any transcript
- **Tunable coverage:** Almost unlimited dynamic range for rare transcript discovery
- **Sensitivity:** Four million tags per sample yields an average of 12 counts for transcripts at one copy per cell
- **Orthogonal validation:** Genome-wide data comparable to qPCR validation of microarrays

Digital Gene Expression is a robust platform because there is no background to subtract from a signal. Additionally, DGE data can be recorded and annotated using current genome information and easily re-annotated as genome databases evolve.

GENE EXPRESSION PROFILING THROUGH SEQUENCING

Powered by Illumina's proprietary sequencing by synthesis technology, the Illumina Genome Analyzer can be used for many different biological discovery applications such as whole-genome resequencing, DNA methylation status, protein-DNA interactions, whole-genome expression profiling, and small RNA discovery and analysis. Each application differs only in its respective sample preparation protocol and downstream data analysis.

For Tag Profiling, Illumina scientists have designed a sample protocol that enables sequencing of any mRNA. Through a simple workflow, mRNA libraries amenable to cluster generation and sequencing by synthesis are created.

The Illumina Tag Profiling template protocol builds constructs comprised of a unique, positionally known 20- or 21-base pair cDNA tag with defined adapters attached to both ends. Tag Profiling provides access to any messenger RNA through two restriction enzyme tag construction options. With a restriction site every 256 base pairs, digestion with *Nla III* captures most mRNA species. For transcripts not addressed with *Nla III* tag construction, Illumina offers an alternate method to anchor tags using *Dpn II* restriction.

The tag mRNA libraries are loaded onto the fully automated Cluster Station where they bind to complementary adapter oligos grafted onto a proprietary flow cell surface. The Cluster Station isothermally amplifies these cDNA constructs to create clonal clusters of ~1000 copies each. The resulting high-density array of template clusters on the flow cell is directly sequenced by the fully automated Illumina Genome Analyzer. Solexa sequencing uses four proprietary fluorescently labeled, reversibly terminated nucleotides to sequence the millions of clusters base by base in parallel with an accuracy rate greater than 99.6% per cycle. Because this protocol does not require any transcript-specific probes, DGE Tag Profiling enables researchers to discover and quantify transcripts in any organism, irrespective of the available annotation.

CONFIDENT NOVEL TRANSCRIPT DISCOVERY

Theoretical calculations suggest that over 99.8% of 21-base pair tags occur only once in genomes the size of the human genome[1]. Analyses based on actual sequence information from approximately 16,000 known genes suggest that >75% of 21-base pair tags are expected to occur only once in the human genome, with the remaining tags matching duplicated genes or repeat sequences[1]. Because DGE Tag Profiling creates and sequences positionally registered tags of 20 and 21 base pairs, researchers are provided with sufficient sequence information for confident identification of novel transcripts from any eukaryotic genome *(Table 1)*.

UNPARALLELED DYNAMIC RANGE

Digital Gene Expression opens up the entire mRNA spectrum for characterization. With 20 times the sensitivity of SAGE, Tag Profiling delivers an unprecedented depth of coverage. Analyzing up to four million tags per sample per flow cell channel, Tag Profiling has, in effect, unlimited dynamic range with excellent reproducibility allowing researchers to probe very low levels of expression *(Figure 2)*.

It is generally accepted that one transcript per cell can be equated to one copy of the transcript in every 350,000 transcripts. Displaying sequencing data in templates per million (TPM), the Genome Analyzer registers a single transcript copy as a signal of three TPM. Therefore, if researchers tune the sequencing depth to four million tags per flow

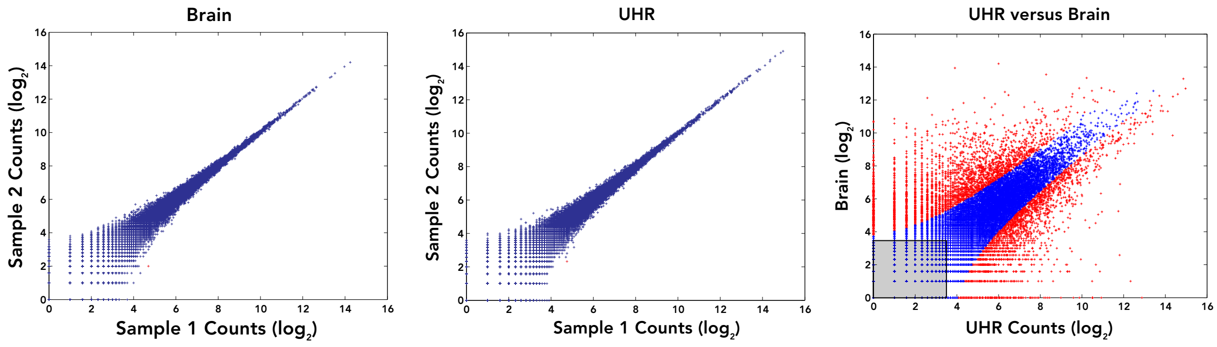
TABLE 1: EXAMPLE OF DGE TAG PROFILING DATA

GENE DESCRIPTION	SYMBOL	HUMAN BRAIN (TPM)*	UNIVERSAL HUMAN REFERENCE (TPM)	TAG SEQUENCE
Hemoglobin, Zeta	HBZ	0	4443	GATCTCCACGCAGGCCCGACA
Gamma-animobutyric acid A receptor, alpha 1	GABRA1	894	0	GATCCCAAACCCAAGTCTTGAA
PHD finger protein 20	PHF20	10	10	GATCCGGGGCTGCAGGCTTG
Glyceraldehyde-3-phospate dehydrogenase	GAPDH	10,202	21,582	GATCATGAGCAATGCCTCCT

Data were generated from an in-house experiment comparing expression profiles for universal human reference and mouse brain samples using DGE Tag Profiling. Alignment against databases allows for gene symbol, description, and other annotations to be associated with each tag.

*Templates per million

FIGURE 2: REPRODUCIBLE RESULTS AT LOW EXPRESSION LEVELS



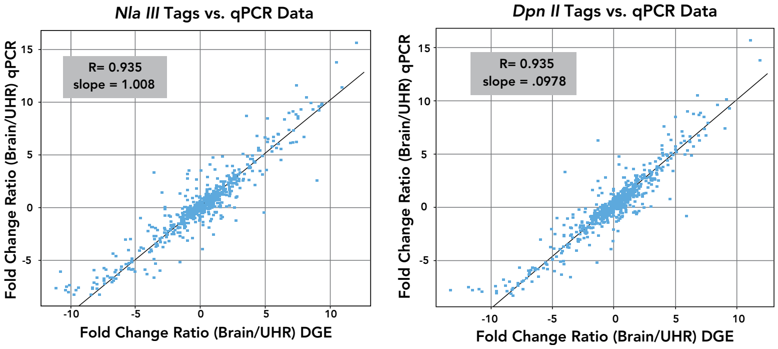
Two samples of MAQC human brain specimen were prepared using the DGE Tag Profiling protocol. Each sample was run on a single flow cell lane. The resulting data, plotted as Sample 1 versus Sample 2, clearly demonstrate the outstanding reproducibility DGE Tag Profiling delivers. Similar results were seen for the same experiment using two MAQC universal human reference (UHR) samples. Distinct differential expression is observed when the two lanes of mouse brain and UHR data are compared. Tags plotted in blue are identical in each specimen. Tags shown in red reflect a greater than 2-fold change in expression (p=0.95). On the UHR versus Brain scatter plot, the data in the shaded box show the signals observed at a level of less than one transcript per cell (three templates per million) for both samples.

cell channel, a single copy transcript will be called 12 times. If greater confidence is desired, the investigator can run multiple channels of one sample on the flow cell. This tunable depth of coverage offers unparalleled confidence calling low and zero expression levels *(Figure 2)*.

ORTHOGONAL VALIDATION

The most commonly employed microarray validation methods are (1) qPCR and (2) replicating the experiment with another microarray system. While qPCR offers robust orthogonal validation, it is an expensive and time-consuming protocol that limits whole-genome or global analysis. While more cost effective, validating microarray data with a different microarray platform does not exclude potential bias associated with hybridization assays. Digital Gene Expression combines the strengths of both methodologies by providing whole-genome orthogonal validation comparable to qPCR *(Figure 3)* at a radically lower cost.

FIGURE 3: COMPARISON OF DGE TAG PROFILING GENE EXPRESSION DATA WITH QUANTITATIVE PCR



Experiments using MAQC samples show the data correlation between qPCR and the Tag Profiling protocol are greater than 0.93. After being assayed by qPCR, 629 and 625 RefSeq genes were quantified using the *Nla III* and *Dpn II* protocols, respectively. Unlike microarray data, there is no observed "ratio compression" in the data as evidenced by the slopes equal to ~1.

DATA ANALYSIS

Tag Profiling generates data using open architecture software, allowing researchers to tailor Illumina Genome Analyzer data analysis software to address their specific needs. Investigators are able to run image analysis, base calling, and standard filtering to generate a list of sequence tags and counts. It is also possible to annotate tags with genomic information and analyze differential gene expression. For the Tag Profiling application, the Genome Analyzer software provides canonical sequences for human and mouse

ORDERING INFORMATION

CATALOG NO.	PRODUCT	DESCRIPTION
FC-102-1005	DGE-Tag Profiling for *Nla III* Sample Prep Kit (1)	Contains reagents for preparing *Nla III* tags from eight total RNA samples (eight samples can be loaded on one flow cell)
FC-102-1006	DGE-Tag Profiling for *Nla III* Sample Prep Kit (5)	Contain reagents for preparing *Nla III* tags from 40 total RNA samples (40 samples can be loaded on five flow cells)
FC-102-1007	DGE-Tag Profiling for *Dpn II* Sample Prep Kit (1)	Contains reagents for preparing *Dpn II* tags from eight total RNA samples (eight samples can be loaded on one flow cell)
FC-102-1008	DGE-Tag Profiling for *Dpn II* Sample Prep Kit (5)	Contains reagents for preparing *Dpn II* tags from 40 total RNA samples (40 samples can be loaded on five flow cells)
FC-103-1004	DGE-Tag Profiling *Nla III* Cluster Generation Kit (1)	Contains reagents, one flow cell, one amplification and one hybridization manifold for processing up to eight samples.
FC-103-1005	DGE-Tag Profiling *Nla III* Cluster Generation Kit (10)	Contains reagents, ten flow cells, ten amplification manifolds, and ten hybridization manifolds for processing up to 80 samples.
FC-103-1006	DGE-Tag Profiling *Dpn II* Cluster Generation Kit (1)	Contains reagents, one flow cell, one amplification manifold, and one hybridization manifold for processing up to eight samples.
FC-103-1007	DGE-Tag Profiling *Dpn II* Cluster Generation Kit (10)	Contains reagents, ten flow cells, ten amplification manifolds, and ten hybridization manifolds for processing up to 80 samples.
FC-104-1001	18-cycle Solexa Sequencing Kit (1)	Contains reagents for generating 17 base pair sequences for eight DGE Tag Profiling samples (one flow cell)
SY-301-2001	Illumina Cluster Station	
SY-301-1001	Illumina Genome Analyzer	

genomes. Expression profiles for other species can be compared easily against public databases like the NCBI RefSeq database (www.ncbi.nlm.nih.gov) and the University of California Santa Cruz (UCSC) genome browser.

SUMMARY

Digital Gene Expression Tag Profiling offers unparalleled depth, specificity, and sensitivity for confident novel and rare transcript discovery in any organism. With a simple workflow that does not require previous sequence knowledge, and transcript quantification comparable to qPCR, Illumina Tag Profiling delivers a truly universal whole-genome expression analysis platform. Like all Illumina gene expression products, Tag Profiling delivers industry-leading levels of accuracy, flexibility, and affordability.

REFERENCES

(1) Saha S, Sparks AB, Rago C, Viatcheslav A, Wang CJ, et al. (2002) Using the transcriptome to annotate the genome. Nat Biotech 20: 508-512.

ADDITIONAL INFORMATION

To learn more about Illumina's RNA Analysis Solutions and other Solexa sequencing applications, visit www.illumina.com or contact us at the address below.

Illumina, Inc.
Customer Solutions
9885 Towne Centre Drive
San Diego, CA 92121-1975
1.800.809.4566 (toll free)
1.858.202.4566 (outside the U.S.)
techsupport@illumina.com
www.illumina.com

FOR RESEARCH USE ONLY

Pub. No. 370-2007-011 23Jul07

illumina®