Applications on Illumina’s Genome Analyzer – From Possibilities to Publications

Shawn C. Baker, Ph.D.
Market Manager – Expression and Regulation
Illumina
THE ILLUMINA GENOME ANALYZER SYSTEM
Flow cell

8 channels

Surface of flow cell coated with a lawn of oligo pairs

Simplified workflow

- Clusters in a contained environment (no need for clean rooms)
- Sequencing performed in the flow cell on the clusters
Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation

DNA (0.1-1.0 ug)

Sample preparation

Cluster growth

Sequencing

Image acquisition

Base calling
New Analysis Algorithm Doubles Yield

- PL 1.0
- PL 1.3
- PL 1.4

x 100 for clusters per lane
New Analysis Algorithm also Improves Accuracy

<table>
<thead>
<tr>
<th></th>
<th>Clusters/tile</th>
<th>PF/tile</th>
<th>Error rates</th>
<th>Perfect reads</th>
<th>Yield* at 2x100bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL 1.3</td>
<td>239,310</td>
<td>136,455</td>
<td>R1: 1.21%</td>
<td>60%</td>
<td>26.1G</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R2: 2.06%</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16M/lane*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PL1.4</td>
<td>238,577</td>
<td>201,140</td>
<td>R1: 0.77%</td>
<td>&gt;70%</td>
<td>38.6G</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>R2: 1.34%</td>
<td>60%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>24M/lane*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Higher yield 84% PF  
Lower error

*for 120 tiles/lane
Increasing Quality, Productivity, Economy

- Increase accuracy: improved chemistry and image detection for >Q30 bases
- Increase read length (and reduce cycle time):
  - Enhance alignment and de novo assembly
- Increase density (250M paired reads / flowcell)
- Increase yield, reduce cost

<table>
<thead>
<tr>
<th>Readlength</th>
<th>Gb PF data / flowcells</th>
<th>Flowcells / 40x genome</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>3.3</td>
<td>40</td>
<td>Yoruba genomes</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>12</td>
<td>1000 genomes pilot</td>
</tr>
<tr>
<td>75</td>
<td>15</td>
<td>8</td>
<td>Melanoma pilot</td>
</tr>
<tr>
<td>100</td>
<td>20</td>
<td>6</td>
<td>Human standard</td>
</tr>
<tr>
<td>100</td>
<td>30</td>
<td>4</td>
<td>Human standard</td>
</tr>
</tbody>
</table>
New Software Improves *Yield* of Data
New Software Improves Accuracy of Data
2009 System Roadmap

On-track for 95G Output in 2009

- SBS Version 3 + Pipeline 1.3
- 2x100 PE
- 2x125 PE
- 2x150 PE
- Pipeline and GA2x
- 20G
- 35G
- 55G
- 95G

Gigabases of Quality Data per Run

Q4 '08 Jan '09 Dec '09
mRNA-Seq: Analysis and Quantitation of the Transcriptome

- Quantitate levels of RNA expression
  - Better sensitivity, dynamic range, and gene coverage than any microarray

- Transcriptome Structure and Genome Annotation
  - Provide EST-like information to annotate previously sequenced genomes
  - Data can be used to discover novel transcripts
  - Study RNA Processing and Characterize Alternative transcription

- Study transcriptome polymorphism
  - Characterize SNPs, point mutations, and insertion/deletions in mRNA
mRNA-Seq Sample Prep Workflow

<table>
<thead>
<tr>
<th>Step</th>
<th>Total Time</th>
<th>Hands-On Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Isolate Poly-A RNA</td>
<td>1.5 hr</td>
<td>1.5 hr</td>
</tr>
<tr>
<td>Fragment RNA</td>
<td>1 hr</td>
<td>10 min</td>
</tr>
<tr>
<td>Make cDNA</td>
<td>4.5 hr</td>
<td>20 min</td>
</tr>
<tr>
<td>Ligate Adaptors</td>
<td>2 hr</td>
<td>45 min</td>
</tr>
<tr>
<td>Size-Select from Gel</td>
<td>1.5 hr</td>
<td>30 min</td>
</tr>
<tr>
<td>PCR Enrich</td>
<td>1 hr</td>
<td>30 min</td>
</tr>
<tr>
<td>Sample Prep Total</td>
<td>11.5 hr</td>
<td>3.75 hr</td>
</tr>
</tbody>
</table>

Sample to Data: < 1 week
Excellent Reproducibility of Digital Counting Data from Genome Analyzer

Reproducibility of counting reads assigned to all RefSeq Genes in two UHRR Replicates

When compared across 7 lanes and 2 flow cells, all Brain/Brain and UHRR/UHRR replicates had linear r2 correlations > 0.99
Differential Expression of Genes in Brain vs. UHRR

Comparison of Digital Counts of Brain vs. UHRR

Correlations between lanes is ~0.47 for all lanes
Large Linear Dynamic Range of Quantitation

Figure taken from “Mapping and quantifying mammalian transcriptomes by RNA-Seq” by Mortazavi, Williams, McCue, Schaeffer & Wold. *Nature Methods*, 2008
Accurate Quantitation: mRNA-SEQ vs. RT-PCR

$R = 0.965$
Slope = 0.984
Overview of mRNA-Seq Assays

- Start with 1 µg (or less) of Total RNA
- Purify poly-A mRNA
- Randomly Fragment RNA

mRNA-Seq

- Random Prime mRNA → cDNA
- Make 2nd Strand cDNA
- Ligate Sequencing Adapters

- PCR Amplify 15 Cycles
- Purify, Quantify
- Grow Clusters
- Sequence on Illumina Genome Analyzer
  → Single or Paired-end reads
  → 35 to 100 base read lengths

Directional mRNA-Seq

- CIP Treatment
- PNK Treatment
- Ligate 3’- Small RNA Adapters
- Ligate 5’- Small RNA Adapters
- RT to make cDNA
mRNA-Seq: Truly Quantitative Gene Expression

FKBP8 Gene Expression

13,025 reads

Brain

8,037 reads

UHRR

RPS3 Gene Expression

3,115 reads

Brain

31,109 reads

UHRR

These screen shots were taken directly from the Illumina Genome Viewer (IGV), which is part of the GenomeStudio™ Software Suite.
A Differentially Expressed Gene – IGFBP1

“Exon” Reads

“Splice Junction” Reads
Directional RNA-Seq

- Protocol uses existing Small RNA library prep protocol
- Retains directionality of the RNA fragment
- Valuable tool for annotation of new genomes
- 1.5-day library prep protocol
- Currently in Early Access Sites – Available soon!

Directional RNA-Seq Workflow

1. Total RNA
2. Poly-A Selection
3. mRNA Fragmentation
4. RNA fragment clean-up
5. RNA Adapter Ligations
6. Perform RT-PCR Amplification
7. Purify Library
Directional RNA-Seq

Standard mRNA-Seq
This Transcript is a Product of Gene: ENSG00000007202

Gene Symbol: KIAA0100
Exons: 39
Transcript length: 7,430 bps
Translation length: 2,235 residues
24

Human transcript comparing Ensembl and Genome Analyzer data
Maher et al, Chimeric transcript discovery by paired-end transcriptome sequencing PNAS June 09
Visualization of Polymorphisms in the Transcriptome

SNP Call made in GenomeStudio
Ribo-Seq

Ribosome footprinting  Total mRNA abundance

Ribosome digestion

nuclease digestion

random fragmentation

size selection, polyadenylation

+dT primer with linkers

reverse transcription

ssDNA ligase

circularization

relinearization

deep sequencing library

Read 5' ends [A.U.]

CDS Position [nt from start]

CDS Position [nt from stop]
GRO-Seq (Global Runon Sequencing)

Fig. 1. Sample of GRO-seq data view on the University of California at Santa Cruz (UCSC) genome browser. A 2.5-Mb region on chromosome 5 showing GRO-seq reads aligned to the genome at 1-bp resolution, followed by an up-close view around the NPM1 gene. Pol II ChIP results are shown in green; mappable regions, black; GRO-seq reads on the plus strand (left to right), red; GRO-seq reads on the minus strand (right to left), light blue; RefSeq gene annotations, dark blue.
Which RNA-Seq assay and how much data are needed?

- **Counting or Profiling**
  - 5 to 10 million total reads of 35 to 50 bp length from poly-A selected RNA will give performance better than any microarray

- **Studying Alternative Splicing or quantifying cSNPs for most transcripts**
  - Deeper profiling of 50 to 100 million reads, with read lengths of 50 to 100 bps, from poly-A selected RNA using mRNA-Seq assay

- **Complete Annotation of a New Transcriptome**
  - 100 million to 1 Billion Reads of 50 to 100 bp read length
  - or Long overlapping reads
  - Combination of random primed and directional chemistry
  - Poly-A selected and ribo-minus treated total RNA
  - Combination of single-read and paired-end data
  - Also must do Small RNA analysis as well...
Wide Variety of Methods are Possible Today on the GA
Wide Variety of Methods are Published Today on the GA

- Read lengths of 35 to 100 + bps
- Random Primed or Stranded Protocols
- Poly-A Selected or Total RNA (-rRNA)
- Single-Read or Paired-End Read
- Insert Sizes of 50 to 500+ bps
Genomic Convergence Analysis of Schizophrenia: Sequencing Reveals Altered Synaptic Vesicular in Post-Mortem Cerebellum

Joann Madge
Shaun Liu*, Beavish, Yaye Langley*, John Victoria Ambrose-Blake

BMC Genomics

Mapping an
by RNA-Seq

Ali Mortazavi1,2, Brian J

We have mapped and quantitatively sequencing those genic regions represented by the provides a digital method to translate information from known patients with schizophrenia.

Methods

Methodology article

Estimating accuracy of RNA-Seq and microarrays with proteomics

Xing Fu1, Xin Fu1, Song Guo2, Zheng Yan2, Ying Xu2, Hao Hu3, Corinna Menzel2, Wei Chen,* Yixue Li1, Yicheng Wang1 and Philipp Khaitovich* 1,4

Abstract

Background: Microarrays are sensitive biotechnological experiments that are able to detect changes in gene expression at the transcriptional level. RNA-Seq and microarrays are two of the most common methods used for transcriptome analysis. RNA-Seq provides an opportunity to study the expression of genes at a higher resolution, allowing for the detection of rare or low-abundance transcripts that are not reliably detected by microarrays. However, RNA-Seq is more expensive and time-consuming than microarrays. Therefore, it is important to validate the results obtained from RNA-Seq studies using independent methods, such as qPCR or Western blotting.

Methods: In this study, we utilized the Gene Expression Omnibus database (GEO) to obtain gene expression data from published studies. We then performed RNA-Seq and microarray experiments on the same set of samples, using the same experimental conditions. We then compared the results obtained from RNA-Seq and microarrays.

Results: Our findings suggest that RNA-Seq provides superior accuracy in the detection of gene expression changes. This is particularly true for genes with low expression levels, where microarrays may fail to detect changes.

Conclusion: Our results demonstrate that RNA-Seq is a more sensitive and accurate method for transcriptome analysis than microarrays. Therefore, RNA-Seq should be considered as the method of choice for studying gene expression changes in complex biological systems.

BMC Genomics

Open Access

Published: 15 April 2009

Accepted: 1 March 2009

This article is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Microarrays are sensitive biotechnological experiments that are able to detect changes in gene expression at the transcriptional level. RNA-Seq and microarrays are two of the most common methods used for transcriptome analysis. RNA-Seq provides an opportunity to study the expression of genes at a higher resolution, allowing for the detection of rare or low-abundance transcripts that are not reliably detected by microarrays. However, RNA-Seq is more expensive and time-consuming than microarrays. Therefore, it is important to validate the results obtained from RNA-Seq studies using independent methods, such as qPCR or Western blotting.

Methods: In this study, we utilized the Gene Expression Omnibus database (GEO) to obtain gene expression data from published studies. We then performed RNA-Seq and microarray experiments on the same set of samples, using the same experimental conditions. We then compared the results obtained from RNA-Seq and microarrays.

Results: Our findings suggest that RNA-Seq provides superior accuracy in the detection of gene expression changes. This is particularly true for genes with low expression levels, where microarrays may fail to detect changes.

Conclusion: Our results demonstrate that RNA-Seq is a more sensitive and accurate method for transcriptome analysis than microarrays. Therefore, RNA-Seq should be considered as the method of choice for studying gene expression changes in complex biological systems.

BMC Genomics

Open Access

Published: 15 April 2009

Accepted: 1 March 2009

This article is distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Highly Integrated Single-Bas Maps of the Epigenome in A

Resource

High-throughput complementary DNA sequencing (RNA-Seq) is a powerful tool for whole-transcriptome analysis, supplying information about a transcript’s expression level and structure. However, it is difficult to determine the polarity of transcripts, and therefore identify which strand is transcribed. Here, we report the development of RNA-Seq protocols that provide information about transcript orientation. Using Saccharomyces cerevisiae and mouse brain transcriptomes as models, we demonstrate that knowledge of the transcript’s orientation allows for more accurate determination of the structure and expression of genes. It also helps to identify new genes and helps in studying promoter-associated and antisense transcripts. The available data sets enable the annotation of the epigenetic landscapes of the organisms. The new protocol for RNA-Seq is capable of capturing a considerable portion of transcriptome analysis in the future.

INTRODUCTION

Recent studies have demonstrated an unexpected complexity of transcript expression in eukaryotes (1–5). In addition to classical mRNAs, which cover ~1.5% of the genome, a vast number of non-coding RNAs (ncRNAs) with a wide variety of expression have been identified (6). These ncRNAs include short non-coding RNAs (sncRNAs) and long non-coding RNAs (lncRNAs), which are involved in various cellular processes such as transcription, translation, and post-translational modifications. The biological function of these ncRNAs is largely unknown, and they represent a new research area, requiring high-throughput transcriptome analysis. RNA-Seq is a new tool for whole-transcriptome analysis, providing critical information about the genome-wide expression of genes. It allows for the identification of novel transcripts and the characterization of the transcriptome, which is essential for understanding the functional interactions between genes and the regulation of biological processes. RNA-Seq technology has revolutionized the study of the transcriptome, enabling researchers to gain insights into the complexity and diversity of gene expression in eukaryotic organisms.

Transcriptome analysis by strand-specific sequencing of complementary DNA

Dmitri Parkhomchuk, Tatiana Boreidna, Yachelev Amstaldsky, Maria Banaru, Linda Haas, Sylva Kredhoch, Hane Lachmar, and Alexey Solodnokov

Max Planck Institute for Molecular Genetics, Department of Vertebrate Genomics, Hedwig, 731495, Germany

Received Apr. 7, 2009; Revised June 4, 2009; Accepted June 20, 2009

ABSTRACT

High-throughput complementary DNA sequencing (RNA-Seq) is a powerful tool for whole-transcriptome analysis, supplying information about a transcript’s expression level and structure. However, it is difficult to determine the polarity of transcripts, and therefore identify which strand is transcribed. Here, we report the development of RNA-Seq protocols that provide information about transcript orientation. Using Saccharomyces cerevisiae and mouse brain transcriptomes as models, we demonstrate that knowledge of the transcript’s orientation allows for more accurate determination of the structure and expression of genes. It also helps to identify new genes and helps in studying promoter-associated and antisense transcripts. The available data sets enable the annotation of the epigenetic landscapes of the organisms. The new protocol for RNA-Seq is capable of capturing a considerable portion of transcriptome analysis in the future.

The RNA-Seq protocol is simple, has a large dynamic range and high sensitivity, and can unequivocally identify splicing and alternative splicing events as well as alternative transcripts. RNA-Seq provides a number of advantages over conventional approaches: microarray hybridization, gene-specific amplification, and sequencing. RNA-Seq analysis does not depend on the presence of particular restriction sites within the transcript and can be used as an alternative to traditional sequencing methods. The digital character of the RNA-Seq data permits accurate quantification of transcript abundance and can potentially replace the use of traditional sequencing methods. RNA-Seq is a powerful tool for the annotation of novel transcripts and can be used to determine the expression levels of genes.
Mapping the Burkhoderia cenocepacia response via high-throughput sequenc ing


BRIEF COMMUNICATIONS

A majority (70%) of ES cell genes with two or more TSSs have both ES- and
mouse-specific binding. Individual TSSs produce both RNA subtypes (Fig. 3). Based on these
observations, we hypothesized that sense and antisense RNA subtypes may
be differentially regulated. We identified two TSSs for each gene, separated
by 500 base pairs (bp), and sequenced each of these TSSs for each
sample. The results showed that there is a higher frequency of
antisense transcription than sense transcription. This is consistent with
previous findings in other species. The data also suggest that the
regulation of antisense transcription may be influenced by the
expression of the sense transcript. Further studies are needed to
understand the role of antisense transcription in the regulation of
gene expression.
Wide Variety of Methods are Published Today on the GA

Read lengths of 35 to 100 + bps

Random Primed or Stranded Protocols

Poly-A Selected or Total RNA (-rRNA)

Single-Read or Paired-End Read

Insert Sizes of 50 to 500+ bps

Grumello-cell tumors (GCs) and cranioblastomas (CCTs) are rare head and neck neoplasms, respectively. The 10q11.21 region is commonly deleted in both GCs and CCTs, but its role in disease pathogenesis remains unclear.

METHODS

We analyzed four-adult-type chimeric trans‐sequencing. We identified putative trans‐scripts to at least three of these samples from patients with recurrent GCs and CCTs, which we validated using qRT‐PCR. The unsuspected trans‐scriptional regulation of the 10q11.21 region may provide new insights into the pathogenesis of GCs and CCTs.

RESULTS

All four GCs and two CCTs had an identical trans‐scriptome. The unexpected trans‐scriptional regulation of the 10q11.21 region may provide new insights into the pathogenesis of GCs and CCTs.

CONCLUSIONS

While trans‐scripts have not been identified as the primary driver of disease, the identification of a novel regulatory mechanism at 10q11.21 provides a potential explanation for the tumorigenic activity of this region.
Illumina’s Small RNA v1.5 library prep protocol

Just 6 hr library prep starting from total RNA

- **Enhanced workflow**
  - Shortest time from library to sequence
  - 2.5 hrs hands on time (6 hr total)
  - Single gel excision

- **NO need for small RNA purification!**
  - Starts directly from total RNA

- **Reduced RNA input requirements**
  - 1.0 µg total RNA input
    - Titrated to 100 ng total

- **Highest quality data**
  - Most accurate quantification of small RNAs
  - Superior results to RNA/DNA duplex method

- **As always, strand specific information**
  - Discover & profile microRNA
  - Sequence other non-coding RNAs

---

**Streamlined Workflow**

1. 3’ RNA Adaptor Ligation
2. 5’ RNA Adapter Ligation
3. Perform RT-PCR Amplification
4. Purify Small RNA Library

**Sample Prep**

- ~ 2.5 hours hands on time
- 6 hours Total time

**RNA to Data**

- ~ 4 days
Small RNA Discovery

- Collaboration with Tuschl
- Multiple stages/tissues of C. elegans
ChIP-Seq
*(Genome Wide Characterization of Transcription Factor Binding Sites and Epigenetic Modifications)*

- **Superior Performance**
  - High sensitivity (millions of tags per sample)
  - Very low background and false positive rate

- **Excellent Coverage**
  - Genome-wide coverage
  - High specificity/fine resolution

- **Low Sample Input**
  - 1 to 10 nanograms of DNA
ChIP-Seq:
Serum Response Factor (SRF) in Mouse C2C12 Cells

CarG Boxes in SRF-enriched Regions

Distance from peak

Site count

4.7 M reads

Wold Lab
Renaissance in the analysis of regulatory elements

- High resolution mapping of regulatory elements including promoters, enhancers, control regions
- Nucleosome positioning
- Histone modification
- Transcription Factor binding
- Chromatin structure
- RNA-binding proteins
Bisulfite Sequencing (BiS-Seq)

**BiS-Seq**

- **Simple Modification**
  - Methylated adapters
  - Bisulfite treatment

- **Genome-wide Coverage**
  - Hypothesis-free
  - Single base resolution
  - No probe design constraints

- **Discovery Applications**
  - Cancer
  - Biomarkers

**Fragment**

- gDNA

**Ligate Modified Adapters**

- 5-MeC

**Bisulfite Treatment**
“Reduced Representation” Sample Prep Method

Based on approach developed by Meissner, etc.

Genomic DNA

\[ \downarrow \]

Restriction Enzyme Digestion

\[ \downarrow \]

Size Selection 200 +/- 25 bp

\[ \times \times \]

\[ \downarrow \]

Bisulfite Treatment

\[ \downarrow \]

Sequence Ends of Selected Fragments

\[ \downarrow \]

Data Analysis
Ilumina’s Complete Expression and Regulation Solution

- DNAse1-Seq
- Nucleosome Mapping
- Hypersensitive Sites
- CH_{2}CO
- CH_{3}
- Epigenetic Modifications
- BIS-Seq
- Methyl-Seq
- Methyl Arrays
- DNA Replication Sites
- ChiP-Seq
- Small RNA Discovery
- Gene
- Long-range regulatory elements (enhancers, repressors/silencers, insulators)
- cis-regulatory elements (promoters, transcription factor binding sites)
- mRNA-Seq
- Ribo-Seq
What can you do with just one single lane of GA data?

- 2X Coverage Human
- 60X Arabidopsis
- 50X Drosophila
- 400X Yeast Genome
- 1200X E. coli