Addressing mRNA Complexity: Utilizing the 5’ Cap Structure of mRNA for Transcriptome Analysis

Matthias Harbers

Presentation at Brown University on April 18, 2011
It is the DNAFORM Mission to contribute to human health, and prosperity of the society while respecting the environment. With our commitment to innovative *Precision Gene Technologies*, developed together with the **Oomics Science Center** of the **Institute of Physical and Chemical Research (RIKEN)** in Japan, we want to provide superior products and services for human diagnostics, food safety, and life science research.

- DNAFORM is located in Yokohama, Japan
- We facilities at Yokohama and Ami Machi
- We are offering services and clone resources to the research community
Developed functional gene annotation by designing appropriate rules and methods. The paper was followed by the draft sequence of human genome (Lander et al. 2001) a week later because they used our cDNA for gene number prediction.

Assigned functional annotations to a set of 60,770 full-length mouse cDNAs. The first project worldwide to standardize full-length mammalian cDNAs. The research was published in a special issue of Nature on the decoding of the mouse genome which was operated by Dr. Waterston and co-workers (Waterston et al. 2002)

More than 70% of the genome is transcribed. More than half of transcripts are non-protein coding RNA (RNA new continent). ncRNAs have a wide variety of functions.

Work out transcription regulation network.

...is ongoing.
Transcript information is the basis to understanding genomes, proteins, and non-coding RNAs!
The 5' cap structure is considered a specific landmark of all eukaryotic mRNA!
The Cap-Trapper Method: Modification of diol structures

7-Methylguanylate

Poly A stretch

Cap site

NaIO₄ Oxidation

Biotin Hydrazide Coupling
The Cap-Trapper Method: Selection of 5’ ends

1st Strand cDNA Synthesis

Adding Biotin group to mRNA

RNase I Digestion

RNase I Digestion will also Destroy rRNA and truncated RNA
Key limitations to gene discovery projects

Normalization and subtraction are essential tools for cloning projects!
Applying concentration dependent hybridization kinetics, the number of clones representing highly expressed genes will be reduced in the cDNAs library.
The Cap-Trapper method and normalization protocols are just 2 examples on how cDNA library and cloning methods have been improved for large scale genomic projects.

cDNA cloning and sequencing: Making clone collections

Genome → mRNA → mRNA pool

Clone resources
Great asset for research community!
Functional analysis of genes.

Databases
High throughput Sequencing
Random clone picking

Clone resources
(cDNA cloning and sequencing)

Representative clone for each gene

cDNA Library Preparation

©DNAFORM 2008
RNA

cDNA Library

Clone Picking

End Sequencing

Annotated Clone Collection

Limited by sequencing cost
Redundancy in clone collection

RNA

Shotgun sequencing

Library Screening

Clones for Targets

Much higher coverage
Focus on new targets

Approaches gene discovery
Preparation of RNA-Seq libraries from cDNA libraries

~5x10^{10} Phages
(original phage stock and high coverage)

Phage DNA Preparation
(avoid bias of using plasmid DNA)

Digestion with Homing Endonuclease
(no cleavage within cDNA inserts)

RNA Synthesis
(unbiased amplification, labeling of ends, defined strand)

DNase Treatment
(no background from phage arms)

Preparation of RNA-Seq Library
(use of standard protocols for different platforms)
Preparation of cDNA for 454 shotgun sequencing

1. **Starting material**: 75 µg total RNA
2. **GsuI Digestion**
3. **Cap Trapper**
4. **2nd strand cDNA synthesis**
5. **polyA/T**
6. **Gsul**
7. **3’-linker**
8. **Option to Normalize by DSN Digestion**
9. **PCR Amplification**
10. **Gsu I Digestion**
11. **Removal of Adapters**
Full-length cDNA

454 fragment library

454 sequencing:
Read length: 1,200,000 reads
average length 360nt length
and 260,000 reads with average
length 260 nt.

Assembly
Velvet, WGS

Visualization

Annotation 1
BLASTX against RefSeq

Annotation 2
Gene Ontology, KEGG

Total transcripts (scaffolds) assembled: 10,877
Mean contig length: 1,026
Average coverage: 19.13
N75 Contig Bases: 836

Hawkeye-compatible contigs file, Hawkeye-compatible QC file

Annotation of transcript

Annotation information for matched genes

(Data analysis by Japan Bioinformatics)
Examples for cloning projects at RIKEN and DNAFORM:

- RIKEN FANTOM mouse cDNA collection: The world’s biggest cDNA collection!
- RIKEN human cDNA collection
- RIKEN Arabidopsis cDNA collection
- Japanese rice cDNA collection (in collaboration with other institutions)
- RIKEN/DNAFORM/JBI/Hayashibara honeybee cDNA collection
- Commercial human cDNA collection for drug discovery done at DNAFORM
- Barley cDNA collection using DNAFORM cDNA library
- Soybean cDNA collection using DNAFORM cDNA library
- Cassava cDNA collection using DNAFORM cDNA library
What did we learn from cDNA cloning projects?

- End-sequencing of cDNA clones 1st approach to transcript discovery
- Great improvements in full-length cDNA cloning
- Building of large cDNA collections (FANTOM, MGC, others…)
- Limited by throughput of capillary sequencing
  (RIKEN FANTOM Pipeline: ~40,000 reads per day)
- Limited by high cost of capillary sequencing
  (Reagent cost only per read in the US$ 1 to 1.5 range)
- Hence, cDNA cloning and sequencing did not cover entire complexity of transcriptomes
- Other methods needed to uncover complexity of transcriptome
Tag-based methods for high-throughput sequencing

- Short sequences ("tags") are sufficient for transcript identification
- Short sequencing reads reduce cost
- Short sequencing reads increase throughput
- Protocol should provide 1 tag per transcript
- Digital expression profiling by counting "tags"
- Unbiased transcript discovery
- Transcript annotation using reference data
- Serial Analysis Gene Expression (SAGE)

Why Cap Analysis Gene Expression (CAGE)?

- Sequencing of 5’ends allows discovery of Transcription Start Sites
- 5’-end sequencing allows transcript identification
- 5’-end sequencing allows promoter identification
- 5’-end sequencing allows monitoring of non-polyadenylated mRNAs
- Cap-Trapper method very effective for 5’-end selection
- Cap-Trapper allows library preparation directly from total RNA
- Shift from “3’-end information” to “5’-end information”
- Cap Analysis Gene Expression (CAGE)

Flow of CAGE projects using high-speed sequencing

CAGE Samples: Starting from total RNA

Family RNA pool
- Full-length mRNA ✓
- Non-polyadenylated mRNA ✓
- Truncated mRNA No!
- rRNA No!

mRNA

Promoter
1a 1b 2 3 4 5 Genome

CAGE Clusters mapped to genome

Data processing
Illumina GAIIx (36 bp reads)

Data visualization

Illumina Sequences
Streptavidin
Biotin
EcoP15I site
Cap structure at 5’ end of mRNA

Random priming
Biotinylation of Cap
RNase treatment
Use of barcodes
27bp tag
Barcode CAGE Tag

CAGE Library Preparation

Broad Narrow

Use of barcodes

DNAFORM Precision Gene Technologies
Comparison of SAGE and CAGE data

- Directly compare SAGE (DGE) and CAGE from same samples
- Use of proliferating and differentiated C2C12 myoblasts as a model

(picture provided by Willem Hoogaars)

- Use of biological triplicates
- Use of Illumina Genome Analyzer for high-speed sequencing
- Jointly Leiden University, Genomatix, ServiceXS, DNAFORM

(Hestand, MS et al., Nucleic Acids Res. 2010 Jul 7. [Epub ahead of print])
Flow SAGE and CAGE data analysis

**CAGE**
- Prolif1-3
- Diff1-3
- Illumina sequencing (1 channel/sample, 2 technical replica) and data processing
- CAGE: Remove 1 base
- Mapping 2 mismatches allowed
  - **CAGE** 742,355 regions
- Set threshold to > 2 TPM
  - **CAGE** 41,862 regions

**SAGE**
- Prolif1-3
- Diff1-3
- SAGE: Add CATG
- Mapping 1 mismatch allowed
  - **SAGE** 361,655 regions
- Set threshold to > 2 TPM
  - **SAGE** 43,512 regions

**ElDorado mouse genome**
- 9,957 annotated exons
- 27,190 partially annotated exons
- 2,368 annotated introns
- 2,347 intergenic regions

**Annotated**
- TSS: 13,541 (32%)
- Promoter regions: 6,331 (15%)
- 3’-end of transcripts: 8,028 (19%)
- FANTOM 3 CAGE data set: 31,680 (76%)

**Assigning CAGE regions to genes (1,000 bp window)**
- **CAGE** 10,409 genes

**Assigning SAGE regions to genes (1,000 bp window)**
- **SAGE** 10,987 genes
MyoD (myogenic maker): Viewed in UCSC Genome Browser

- Narrow CAGE peak (MyoD promoter has TATA box)
- "Exon Painting"
- Transcriptional activity at 3’-end
- SAGE peak
Reproducibility of SAGE and CAGE data

(A) Sequencing replica (CAGE only)  
(B) Biological replica  
(C) Differential expression

- **A**: P=0.981  
- **B**: P=0.963  
- **C**: P=0.771  
- **D**: P=0.930  
- **E**: P=0.839
## GO terms found in SAGE, CAGE and microarray data

<table>
<thead>
<tr>
<th>CAGE GO</th>
<th>SAGE GO</th>
<th>Microarray GO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulation of striated muscle contraction</td>
<td>Regulation of muscle contraction</td>
<td>Cycline-dependent protein kinase inhibitor activity</td>
</tr>
<tr>
<td>Cardiac muscle contraction</td>
<td>Cardiac muscle contraction</td>
<td>Myogenesisis</td>
</tr>
<tr>
<td>Myogenesisis</td>
<td>Myogenesisis</td>
<td>Skeletal muscle development</td>
</tr>
<tr>
<td>Regulation of muscle contraction</td>
<td>Regulation of striated muscle contraction</td>
<td>Myoblast differentiation</td>
</tr>
<tr>
<td>Skeletal muscle development</td>
<td>Skeletal muscle development</td>
<td>6-phosphofructokinase activity</td>
</tr>
<tr>
<td>Muscle development</td>
<td>Myofibril assembly</td>
<td>Muscle development</td>
</tr>
<tr>
<td>Striated muscle contraction</td>
<td>Muscle development</td>
<td>Muscle cell differentiation</td>
</tr>
<tr>
<td>Myoblast differentiation</td>
<td>Myoblast fusion</td>
<td>Tumor suppressor activity</td>
</tr>
<tr>
<td>Muscle cell differentiation</td>
<td>Striated muscle contraction</td>
<td>Myofibril assembly</td>
</tr>
<tr>
<td>Sarcomere organization</td>
<td>Muscle cell differentiation</td>
<td>Heart development</td>
</tr>
<tr>
<td><strong>10/10 muscle related</strong></td>
<td><strong>10/10 muscle related</strong></td>
<td><strong>7/10 muscle related</strong></td>
</tr>
</tbody>
</table>
Differentially regulated TSS found in CAGE regions

- Found 196 new differentially regulated TSS in CAGE data
- Out of which 111 regions are upstream of known genes
- Out of which 85 regions are downstream of known genes
- 7 out of 8 regions tested could be confirmed by RT-PCR
Discovery of novel TSS by CAGE

- **Known TSSs**
- **Novel TSSs**

**Density of regions**

**tags per millions of reads aligned**

- 2-3
- 4-5
- 6-10
- 11-30
- 31-50
- 51-100
- 101-300
- 301-500
- 501-1k
- >1k
nanoCAGE: Preparation of CAGE libraries starting from as little as 50 ng total RNA

CAGE-Scan: Use of paired-end sequencing to link new TSS to known genes

Helicos-CAGE: Use of single-molecule sequencing to reduce bias in CAGE libraries and reduced sample requirements

Link high-speed sequencing to cDNA cloning: Creating the resources needed to study newly discovered transcripts!
Summary

- Cap-Trapper proved instrumental to big cloning projects
- Cap-Trapper is used for shotgun sequencing of cDNA
- Cap-Trapper is basis to new CAGE method to sequence 5’ ends
- CAGE established method for ENCODE and FANTOM projects
- CAGE data more complex than SAGE: 67% of gene have multiple CAGE regions
- CAGE data allowed for discovery of new TSS
- CAGE data indicate transcriptional at 3’-ends of annotated genes
- CAGE data showed some exon-painting for many transcripts
Acknowledgements

Leiden University Medical Center:
Matthew S. Hestand
Yavuz Ariyurek
Yolande Ramos
Gert-Jan B. van Ommen
Johan T. den Dunnen
Peter A.C. ‘t Hoen

Genomatix:
Andreas Klinghoff
Matthias Scherf
Thomas Werner

Service XS:
Wilbert van Workum

DNAFORM:
Makoto Suzuki
Fumiko Kobayashi

Omics Science Center RIKEN Yokohama:
Piero Carninci
Charles Plessy
Yoshihida Hayashizaki

Japan Bioinformatics:
Boris Umylny
Thank you for your interest!

Contact: Dr. Matthias Harbers at matthias.harbers@dnaform.jp