High-Confidence Transcriptome Assembly Resurrects Large-scale Unstranded RNAs-seq Data

Bo-Hyun You, Sang-Ho Yoon, and Jin-Wu Nam @ Dept. of Life Science, HYU
Public data portal may include many errors due to analyzing unstranded RNA-seq data
Unstranded vs stranded RNA-seq

Stranded

Unstranded
Unstranded RNA-seq reads

Unstranded

Stranded
Problem 1. Unstranded RNA-seq causes error-prone assembly

From ENCODE RNA-seq data
Problem 1. Unstranded RNA-seq causes error-prone assembly
Problem 2. Mis-annotations causes quantification errors

- RPKM (≈ FPKM) = reads (fragment) per kilobases of exons per million mapped reads.
- 1 RPKM ~ 1 copy in a cell.

10 million mapped reads

1kb in length

0.8kb (HeLa)
Problem 3. Unstranded RNA-seq causes quantification errors
Gene annotation and personal transcriptome projects produced large-scale unstranded RNA-seq datasets

- ENCODE/GENCODE (Science 2012, Genome Res. 2012)
- modENCODE (Science 2010)
- Human Body Map project (Gene Dev. 2012)
- Zebrafish IncRNA annotation (Cell 2012)
- Worm IncRNA annotation (Genome Res. 2012)
- MiTranscriptome (Nat. Genetics 2015)
- GTEx (Nat Genet. 2013)
- TCGA (https://cancergenome.nih.gov/)
- ICGC (https://icgc.org/)
- The human protein atlas (http://www.proteinatlas.org/)
Probabilistic estimation of directions of unstranded reads

**Training k-order Markov chain**

**Step 1: Training**
- Stranded RNA-seq reads
- Random sampling
- Sampled reads
- Directions of k-nearest reads from x
- Model training

**Prediction of strand**

**Step 2: Prediction**
- Unstranded RNA-seq reads
- Directions of k-nearest reads from y
- Maximum likelihood estimation
- Assign directions
- RPDs
Antisense-overlapped reads benefited from strand prediction
Strand prediction benefits the quantification
Exon-junction and boundary updates of transfrrags
Co-assembly improves the quality of transcriptome maps
CAFE: Co-Assembly Followed by End-correction
CAFE helps us reconstruct precise full-length transcriptomes.
BIGTranscriptome from large-scale public RNA-seq data

65 Unstranded RNA-seqs (ENCODE + BodyMap) 104 Stranded RNA-seqs (ENCODE + BodyMap + GEO)

Quality control & cell-type match

62 unstranded and 60 stranded RNA-seqs (35 cell-types and tissues)

MAXIM

62 RPDs and 60 stranded RNA-seqs

TSSs (17 tissues) CPSs (four cell lines)

COCOA and BEX

338,359 transcripts (46,634 loci)

ENCODE + Body Map

(n=43)

Sensitivity (%)
High-confidence BIGTranscriptome comparable to a long-read method

<table>
<thead>
<tr>
<th>Annotation (%)</th>
<th>BIGTranscriptome</th>
<th>MiTranscriptome</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base level</td>
<td>Intron level</td>
</tr>
<tr>
<td></td>
<td>SN</td>
<td>SP</td>
</tr>
<tr>
<td>RefSeq</td>
<td>91.4</td>
<td>48.3</td>
</tr>
<tr>
<td>Gencode (manual)</td>
<td>86.6</td>
<td>66.4</td>
</tr>
<tr>
<td>Gencode (automatic)</td>
<td>90.9</td>
<td>28.5</td>
</tr>
<tr>
<td>PacBio (MCF7)</td>
<td>85.6</td>
<td>50.2</td>
</tr>
<tr>
<td>EST</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RefSeq + Gencode + PacBio + EST</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

![Graph 1](image1.png)

![Graph 2](image2.png)
BIGTranscriptome includes known and novel IncRNAs
High-confidence noncoding transcriptome map leads to better downstream analyses
BIGTranscriptome-ST from large-scale personal RNA-seq data
High-confidence coding and noncoding transcriptome maps

Bo-Hyun You, Sang-Ho Yoon and Jin-Wu Nam

*Genome Res.* 2017 27: 1050-1062 originally published online April 10, 2017
Access the most recent version at doi:10.1101/gr.214288.116

---

**Download**

<table>
<thead>
<tr>
<th>Program</th>
<th>Description</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAFE source code</td>
<td>Version 1.0.1</td>
<td>download</td>
</tr>
<tr>
<td>CAFE manual</td>
<td>User manual</td>
<td>download</td>
</tr>
</tbody>
</table>

**Annotations**

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Format</th>
<th>Download</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIGTranscriptome</td>
<td>GENCODE, Human BodyMap, and GEO (122 samples, 35 cell-types and tissues)</td>
<td>gtf</td>
<td>download</td>
</tr>
<tr>
<td>BIGTranscriptome-TS</td>
<td>GTEx and TCGA (4,821 samples, 19 tissues and tumors)</td>
<td>gtf</td>
<td>download</td>
</tr>
<tr>
<td>BIGTranscriptome IncRNA catalog</td>
<td>Known and novel IncRNAs annotated from BIGTranscriptome</td>
<td>gtf</td>
<td>download</td>
</tr>
</tbody>
</table>

**Requirement**

This program was developed on Linux environment (CentOS 6.8). Cufflinks, samtools, perl and python are required for running.

**License**

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](http://creativecommons.org/licenses/by-nc-sa/4.0/).

---

http://big.hanyang.ac.kr/CAFE
BIGLab

• http://big.hanyang.ac.kr
Strand-specific evaluation of transcriptome maps

Base level

Reference

Transfrags

FN₂, TP₂
FN₁
TP₁
TP₂
FN₁

Sensitivity = \frac{(TP₁ + TP₄ + TP₂)}{(TP₂ + TP₁ + TP₂) + (FN₂ + FN₁ + FN₁)}

Specificity = \frac{(TP₃ + TP₄ + TP₃)}{(TP₂ + TP₄ + TP₂) + (FP₁ + FP₁)}

Intron level

Reference

Transfrags

TP₀, FN₀
FN₁
FN₁

Sensitivity = \frac{(TP₀)}{(TP₀) + (FN₀ + FN₁)}

Specificity = \frac{(TP₁)}{(TP₁) + (FP₀ + FP₁)}