Evaluation of U12-type non-canonical splicing in human ENCODE RNA-Seq datasets and analysis of biological functions for spliced sequences by Read-Split-Fly algorithm

Yongsheng Bai and Jeff Kinne

Department of Biology
Department of Mathematics & Computer Science

June 8, 2016
Human Genome Project and Next-Generation Sequencing
Cost per Genome

Moore's Law

NIH National Human Genome Research Institute

gene.gov/sequencingcosts
ENCODE Project
(https://www.encodeproject.org/)
Datasets – ENCODE
([https://www.encodeproject.org/](https://www.encodeproject.org/))

<table>
<thead>
<tr>
<th>Dataset Classification</th>
<th>Experiment</th>
<th>File Name</th>
<th>Run type</th>
<th>Forward Or Reverse Mate</th>
<th>Biological Replicate</th>
<th>Technical Replicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paired-Ended with Biological Replicate</td>
<td>ENCSR468ION ENCF002DJH</td>
<td>paired-ended</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ENCSR468ION ENCF002DJU</td>
<td>paired-ended</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ENCSR468ION ENCF002DJV</td>
<td>paired-ended</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ENCSR468ION ENCF002DJX</td>
<td>paired-ended</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Paired-Ended without Biological Replicate</td>
<td>ENCSR001UXR ENCF576OBS</td>
<td>paired-ended</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ENCSR001UXR ENCF063OPQ</td>
<td>paired-ended</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Single-Ended with Biological Replicate</td>
<td>ENCSR000BYG ENCF000MVL</td>
<td>single-ended</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>ENCSR000BYG ENCF000MVN</td>
<td>single-ended</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Transcription and Translation

- Pre-mRNA
- Splicing
- Alternatively spliced mRNA
- Translation
- Protein isoforms
The original code for the RSW algorithm worked by considering all pairs of parts of unmapped reads at once. The software was written in Perl. Results are computed much faster in RSR by considering each unmapped read separately. The algorithm was implemented in C++. 
## Comparison of number of reads for supporting *Xbp1* 26 nt spliced regions reported by RSR and other tools

<table>
<thead>
<tr>
<th>Software</th>
<th>500 nM Thapsigargin (Tg)</th>
<th>1 mM Dithiothreitol (Dtt)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Het (Ire1α +/-)</td>
<td>KO (Ire1α -/-)</td>
</tr>
<tr>
<td>Read-Split-Run (RSR)</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>TopHat</td>
<td>23</td>
<td>86</td>
</tr>
<tr>
<td>BWA</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>STAR</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Alt Event Finder</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(Bai *et al.*, 2016. To appear in BMC Genomics)
RSF Pipeline – Improved algorithm

• Rescue read halves from bowtie max file

Mouse mm9
Chr 11, position 5,424,242

GGGAGTGGAGTAAGGCTGGTGCCCGGGTCTGCTGAGTCCGCAGCAGCTCAGACTATGTGCAACCTCTGTGCAGCAGTTGCAAGGCCAGT

Max file (Read mapped onto too many locations)
RSF Pipeline – Improved algorithm

- Increased sensitivity
  - Rescue read halves from bowtie max file

<table>
<thead>
<tr>
<th></th>
<th>Xbp1 splice Supporting reads</th>
<th>Read-Split-Run</th>
<th>Read-Split-Fly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tg treated sample</td>
<td>21</td>
<td>27</td>
<td></td>
</tr>
<tr>
<td>Dtt treated sample</td>
<td>173</td>
<td>209</td>
<td></td>
</tr>
</tbody>
</table>
### RSF Pipeline – Improved algorithm

- Increased performance (Memory and Time Usage)

<table>
<thead>
<tr>
<th></th>
<th>Memory (GB)</th>
<th>Time (Hours)</th>
<th>Read-Split-Run</th>
<th>Read-Split-Fly</th>
<th>Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENCSR558LHB</td>
<td>123.2</td>
<td>15.8</td>
<td>18.9</td>
<td>11.5</td>
<td>-84.7 / -27.2</td>
</tr>
<tr>
<td>ENCSR905FLM</td>
<td>110.2</td>
<td>13.5</td>
<td>20.7</td>
<td>12.6</td>
<td>-81.2 / -6.7</td>
</tr>
<tr>
<td>ENCSR000CQY</td>
<td>55.2</td>
<td>0.55</td>
<td>46.5</td>
<td>0.42</td>
<td>-15.8 / -24</td>
</tr>
</tbody>
</table>
Extracting Spliced Sequences

- Sample data before and after bowtie-inspect-RSR file processing:

<table>
<thead>
<tr>
<th>GeneName</th>
<th>Chromosome</th>
<th># supporting reads</th>
<th>splice length</th>
<th>range of supporting reads</th>
<th>Novel or not (*)</th>
<th>Bracketed sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xbp1</td>
<td>chr11</td>
<td>22, 26</td>
<td>5424280-5424312</td>
<td>Novel GGTCCTGCTGAGTCCGCA--GCAGGTGCAGGCCAGT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ddx41</td>
<td>chr13</td>
<td>2, 74</td>
<td>55635449-55635524</td>
<td>* CCGAGCCAGCTCTGGC--GAGGGCCAGATGATGACC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nol7</td>
<td>chr13</td>
<td>4, 76</td>
<td>43494024-43494104</td>
<td>* GCGAGTGTGCTGCA--GSGATATAACCGCTTGGAGGA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gtbp4</td>
<td>chr13</td>
<td>2, 80</td>
<td>8991098--8991177</td>
<td>* GCACATATCTTTAGCAACA--TTGTCACCAAGTT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pdlim7</td>
<td>chr13</td>
<td>5, 81</td>
<td>55610170-55610255</td>
<td>* TGCTCTGAACAGCGCTGCGCT--CTGCTAGAGCCAGG</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Consequently, the spliced sequence fits snugly within the “range of supporting reads.”

- For example, the **XBP1** candidate spliced sequence begins at index 5424280 and ends at index 5424311
Consensus sequences of human U12- and U2-type intron

(Padgett, R. 2012)
# U12-type non-canonical splicing in 21 human ENCODE RNA-Seq datasets

**E-Value**: 1

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>u12</th>
<th>u12 Total</th>
<th>u2</th>
<th>u2 Total</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>known</td>
<td>novel</td>
<td>known</td>
<td>novel</td>
<td></td>
</tr>
<tr>
<td>3pFull</td>
<td>52977</td>
<td>36760</td>
<td>89737</td>
<td>13094</td>
<td>21389</td>
</tr>
<tr>
<td>5pFull</td>
<td>8840</td>
<td><strong>11618</strong></td>
<td>20458</td>
<td>3466</td>
<td>5606</td>
</tr>
<tr>
<td>branch</td>
<td>222</td>
<td>397</td>
<td>619</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>branch#extend</td>
<td>12968</td>
<td>13342</td>
<td>26310</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td><strong>Grand Total</strong></td>
<td><strong>75007</strong></td>
<td><strong>62117</strong></td>
<td><strong>137124</strong></td>
<td><strong>16560</strong></td>
<td><strong>10435</strong></td>
</tr>
</tbody>
</table>
RSF -BLAST-Workflow

Spliced Sequences Files

miRNA Query (Homo Sapiens)

BLAST Script

# Hit Sequences Results

Statistical BLAST & miRNA Results

(Input)

(Output)
Query – Sample Data

>hsa-let-7a-5p MIMAT0000062 Homo sapiens let-7a-5p
TGAGGTAGTAGGTGTATAGTT
>hsa-let-7a-3p MIMAT0004481 Homo sapiens let-7a-3p
CTATACAAATCTACTGTCTTTC
>hsa-let-7a-2-3p MIMAT0010195 Homo sapiens let-7a-2-3p
CTGTACAGCCTCCTAGCTTTCC
miRNA BLAST hits for 21 ENCODE Samples

miRNA_hits_at_evalue_0.001

miRNAs

- hsa-miR-1226-5p
- hsa-miR-1229-5p
- hsa-miR-1244
- hsa-miR-1248
- hsa-miR-1254
- hsa-miR-1266-5p
- hsa-miR-1273a
- hsa-miR-1273c
- hsa-miR-1273d
- hsa-miR-1273e
- hsa-miR-1273g-3p
- hsa-miR-1273g-5p
- hsa-miR-1273h-3p
- hsa-miR-1273h-5p
- hsa-miR-1285-3p
- hsa-miR-1285-5p
- hsa-miR-1291
- hsa-miR-1292-5p
- hsa-miR-1303
- hsa-miR-1304-3p
- hsa-miR-1972
- hsa-miR-215-3p
- hsa-miR-301a-3p
- hsa-miR-3064-3p
- hsa-miR-3132
- hsa-miR-3136-5p
- hsa-miR-3153
- hsa-miR-3159
- hsa-miR-3184-3p
- hsa-miR-3184-5p
- hsa-miR-3189-5p
- hsa-miR-3199
- hsa-miR-3529-3p
- hsa-miR-3620-5p
- hsa-miR-3651
- hsa-miR-3662
- hsa-miR-3692-3p
- hsa-miR-3692-5p
- hsa-miR-378a-3p
- hsa-miR-378c
- hsa-miR-378i
- hsa-miR-3916
- hsa-miR-3944-3p
- hsa-miR-3960
- hsa-miR-423-3p
- hsa-miR-423-5p
- hsa-miR-4488
- hsa-miR-449c-5p
- hsa-miR-454-3p
- hsa-miR-4644
- hsa-miR-466
- hsa-miR-4668-3p
- hsa-miR-4685-5p
- hsa-miR-4691-3p
- hsa-miR-4691-5p
- hsa-miR-4706
- hsa-miR-4728-3p
- hsa-miR-4763-3p
- hsa-miR-5009-5p
- hsa-miR-5088-5p
- hsa-miR-5095
- hsa-miR-5096
- hsa-miR-548aa
- hsa-miR-548ab
- hsa-miR-548ad-5p
- hsa-miR-548ae-5p
- hsa-miR-548aj-3p
- hsa-miR-548aj-5p
- hsa-miR-548am-5p
- hsa-miR-548ap-3p
- hsa-miR-548aq-3p
- hsa-miR-548as-5p
- hsa-miR-548au-5p
- hsa-miR-548ay-5p
- hsa-miR-548az-5p
- hsa-miR-548b-5p
- hsa-miR-548bb-5p
- hsa-miR-548c-3p
- hsa-miR-548c-5p
- hsa-miR-548d-3p
- hsa-miR-548d-5p
- hsa-miR-548g-5p
- hsa-miR-548h-3p
- hsa-miR-548i
- hsa-miR-548n
- hsa-miR-548o-5p
- hsa-miR-548s
- hsa-miR-548t-3p
- hsa-miR-548u
- hsa-miR-548w
- hsa-miR-548x-5p
- hsa-miR-548z
- hsa-miR-5585-3p
- hsa-miR-5585-5p
- hsa-miR-5586-3p
- hsa-miR-5684
- hsa-miR-579-3p
- hsa-miR-593-5p
- hsa-miR-603
- hsa-miR-611
- hsa-miR-619-5p
- hsa-miR-638
- hsa-miR-6514-5p
- hsa-miR-664a-5p
- hsa-miR-664b-5p
- hsa-miR-6728-5p
- hsa-miR-6735-5p
- hsa-miR-6747-5p
- hsa-miR-6758-5p
- hsa-miR-6775-5p
- hsa-miR-6782-5p
- hsa-miR-6797-5p
- hsa-miR-6801-5p
- hsa-miR-6805-3p
- hsa-miR-6807-3p
- hsa-miR-6812-5p
- hsa-miR-6817-5p
- hsa-miR-6826-5p
- hsa-miR-6831-5p
- hsa-miR-6853-5p
- hsa-miR-6855-5p
- hsa-miR-6862-3p
- hsa-miR-6873-3p
- hsa-miR-6885-5p
- hsa-miR-6894-5p
- hsa-miR-7107-3p
- hsa-miR-7113-3p
- hsa-miR-7161-3p
- hsa-miR-7-5p
- hsa-miR-7704
- hsa-miR-8075
- hsa-miR-93-5p
- hsa-miR-9-5p
Summary

• Preliminary results from 21 samples of ENCODE datasets show that there are several miRNAs are prevalent (> 50%) in studied ENCODE samples. Two of them (hsa-miR-1273d and hsa-miR-548) are associated with many diseases as suggested in the literature.

• U12-type non-canonical splicing could be noteworthy in ENCODE datasets.

• RSF for identifying U12-type splicing events using ENCODE datasets is applicable to study a range of diseases across biological systems under different experimental conditions.
## Read-Split-Run

A pipeline for detecting non-canonical spliced-regions in RNA-Seq data.

### Experiment Settings

<table>
<thead>
<tr>
<th>Mode</th>
<th>Read Type</th>
<th>Experiment Replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-comparative</td>
<td>Single</td>
<td>1</td>
</tr>
</tbody>
</table>

### Data

- **Select Genome Reference**: mm9
- **Select RNA-Seq File (FASTQ format)**: Tg_Het.txt
- **Quality Encoding**: Solexa
- **Length of Reads (in bp)**: 35

[http://bioinf1.indstate.edu/RSR](http://bioinf1.indstate.edu/RSR)
Pre-processing

Minimum Split Size (2+ bp) 11

Maximum Good Alignments Allowed Per Read 11

Candidate Selection

Minimum Distance Between Candidate Pairs (2 to 12 bp) 2

Maximum Distance Between Candidate Pairs (bp) 30k

Read Mapping Region Boundary Buffer Size (0 to 11) bp 3

Minimum Number of Supporting Reads 10

Output

Email: Yongsheng.Bai@indstate.edu

Submit Reset
Bai Lab (Spring 2016)
THANK YOU