SQANTI: Classification, Curation and Quantification of a PacBio transcriptome

Manuel Tardaguila, PhD

Ana Conesa Lab
PacBio in Oligodendrogenesis

Neural Stem Cells (neurospheres) → Oligodendrocyte Progenitor Cells → Immature Oligodendrocytes → Premyelinating mature Oligodendrocytes → Myelinated mature Oligodendrocytes

- Illumina Short reads → PacBio RIs → ToFU transcripts
- RefSeq → Curated Transcriptome → Corrected Transcriptome → SQANTI

Tardaguila et al. SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification Pre-print BioRxiv (2017)
1. CLASSIFICATION OF PACBIO TRANSCRIPTS
Splice-based classification (I): Known Transcripts

**Full Splice Match (FSM)**

- Reference
- PacBio
- Donor
- Acceptor
- Splice Junction

Absolute number of transcripts: \( n = 16,104 \)

**Incomplete Splice Match (ISM)**

- Reference
- PacBio

**Isoform distribution across structural categories**

- Coding prediction
  - Coding
  - Non Coding

- FSM: 55.8%
- ISM: 8.9%
Splice-based classification (II): Novel Transcripts from known genes

Novel In Catalog (NIC)
- Novel SJs with known Donors and acceptors

Novel Not in Catalog (NNC)
- Novel SJs from novel Donors and/or Acceptors

Reference
- PacBio

New combination of Known SJs

Retained Intron

Isoform distribution across structural categories

Coding prediction
- Coding
- Non Coding

% Transcripts

n=16,104

FSM  55.8%
ISM  8.9%
NIC  19.9%
NnO  14.1%
Splice-based classification (III): Novel Transcripts from novel genes

Intergenic

PacBio

Gene A

Gene B

Genic Intron

Exon A

Intron A-B

Exon B

PacBio

Fusion Transcripts

Gene A

Gene B

PacBio

Isoform distribution across structural categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Coding</th>
<th>Non Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSM</td>
<td>55.8%</td>
<td></td>
</tr>
<tr>
<td>ISM</td>
<td>8.9%</td>
<td></td>
</tr>
<tr>
<td>NIC</td>
<td>19.9%</td>
<td></td>
</tr>
<tr>
<td>NIO</td>
<td>14.1%</td>
<td></td>
</tr>
</tbody>
</table>

0.4% 0.2% 0.2% 0.2%
Splice-based classification (IV): Antisense and Genic Genomic transcripts

**Genic Genomic**

Exon A ➔ Intron A-B ➔ Exon B

*PacBio*

**Antisense**

Gene A

Ref 1 ➔ Ref 2

*PacBio*

*Antisense*

**Isoform distribution across structural categories**

<table>
<thead>
<tr>
<th>Coding prediction</th>
<th>% Transcripts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coding</td>
<td>55.8%</td>
</tr>
<tr>
<td>Non-Coding</td>
<td>44.2%</td>
</tr>
</tbody>
</table>

FSM: 8.9%, ISM: 19.9%, NLC: 14.1%, NCO: 0.2%, Genic Genomic: 0.3%, Genic Antisense: 0.4%, Fusion: 0.2%, Intergenic: 0.2%, Genic Intron: 0.2%
Splice based classification allows to separate known transcripts (FSM and ISM) from novel transcripts arising from novel splice junctions or new combinations of already known splice junctions (NIC and NNC).

The predominant categories in our transcriptome are FSM, ISM, NIC and NNC and they are mostly coding (oligodT library preparation).
2. CURATION OF PACBIO TRANSCRIPTS
40% novel isoforms in mouse... Are all of them real?
PacBio output needs curation

- NNC transcripts characterized by novel Donors and/or Acceptors of splicing concentrate traits of low quality transcripts
Using Short reads and Long reads to mine QC features

SQANTI

- SQANTI classification
  - Gene model annotation
  - ORF annotation
  - FL number
  - Min coverage SJ and location
  - Non-canonical splicing
  - ...

- SQANTI SJ level annotation file
  - Known/novel
  - Canonical/Non-canonical
  - SJ coverage
  - Position relative to TSS
  - Presence of RT-switching
  - ...

Corrected Transcriptome

SQANTI transcript level annotation file

Expression Matrix

Short-read/Full-Length Coverage

Optional
Features selected for Random Forest Classification

- **bite**
  - **NEG**
  - **POS**
  - [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]

- **test set**
  - **NEG**
  - **POS**
  - [2, 5, 10, 20, 50, 100, 200, 500, 2000, 5000]

- **geneExp**
  - **NEG**
  - **POS**
  - [1, 100, 10000]

- **isoExp**
  - **NEG**
  - **POS**
  - [1, 2, 5, 10, 20]

- **Min_sample_cov**
  - **NEG**
  - **POS**
  - [1, 100, 10000]

- **MinCov**
  - **NEG**
  - **POS**
  - [20, 50, 100, 200, 500, 1000, 2000, 5000]

- **MinCovPos**
  - **NEG**
  - **POS**
  - [1, 2, 5, 10, 20]

- **nIndels**
  - **NEG**
  - **POS**
  - [1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0]

- **nIndelsJunc**
  - **NEG**
  - **POS**
  - [1.0, 1.2, 1.4, 1.6, 1.8, 2.0]

- **ratioExp**
  - **NEG**
  - **POS**
  - [1, 10, 100, 1000, 10000]

- **sdCov**
  - **NEG**
  - **POS**
  - [1.0, 1.2, 1.4, 1.6, 1.8, 2.0]

- **FSM_class**
  - **NEG**
  - **POS**
  - [0.0, 0.2, 0.4, 0.6, 0.8, 1.0]

- **exons**
  - **NEG**
  - **POS**
  - [1000, 2000, 3000, 4000, 5000, 7000]
PCR validation in an independent set of transcripts

<table>
<thead>
<tr>
<th>R.H. (dT) PCR</th>
<th>FSM</th>
<th>NIC</th>
<th>NNC</th>
<th>Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>23</td>
<td>7</td>
<td>8 (3 nc)</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>2</td>
<td>12 (8 nc)</td>
<td>2</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>9</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>Positive</td>
<td>5 (3 nc)</td>
<td>7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>0</td>
<td>6 (3 nc)</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

1.2) Results on the PCR set:

**Figure 3:** ROC curve obtained on the PCR set. The red curve was obtained on the complete PCR set, the dark red curve on the PCR set reduced to the novel transcripts.

**Figure 4:** Boxplot of the values of the random forest probabilities to be positive obtained on the PCR sets (red for the PCR negative transcripts, green for the PCR positive transcripts).

The area under roc obtained on the PCR sets is comprise between 96.66 (if all the type of sequences are evaluated) and 95.56 if the classifier is only applied to the novels.
Splice Junction independent curation: Intrapriming features

- oligoDT can prime outside polyA regions in A rich regions inside transcripts
- We looked for transcripts showing >= 80% Adenines in the 20 nts downstream (DNA) the end of the transcript
- If this transcripts lack a consensus poly Adenilation site we filter them out
- 605 transcripts filtered out
SQANTI filter eliminates transcript with bad QC features

Isoform distribution across structural categories

Quality control attributes across structural categories
SQANTI filter works across different PacBio datasets: Maize

- **Before SQANTI**

- **AFTER SQANTI**
SQANTI filter works across different PacBio datasets: MCF-7

- Before SQANTI

- AFTER SQANTI
• PacBio output needs curation, specially in NNC novel acceptor/donor transcripts
• SQANTI mines features based in splice junctions, expression and structure to feed a RF classifier that succeeds in filtering out transcripts that fail to validate in an independent PCR set
• SQANTI works across different datasets
3. QUANTIFICATION OF PACBIO TRANSCRIPTOME
Quantification: PacBio curated vs RefSeq: Transcripts

- PacBio finds a set of highly expressed transcripts and reproducible transcripts in common with RefSeq.
- RefSeq detects lots of lowly expressed, difficult to reproduce transcripts. However it also captures exclusively 30% of highly expressed ones (PacBio 18%).
Most of the PacBio exclusive transcripts are new combinations of already known Splice Junction

![Bar chart: % of novel transcripts](image)

- **ENSEMBL**
- **NIC: new combination of known SJ**
- **NIC: novel SJ of known D/A**
- **NNC**
- **antisense**
- **fusion**
- **genic genomic**
- **genic intron**
- **intergenic**

3,674 PacBio exclusive transcripts

- Found by both PacBio Exclusive OLP replicate 1
  - High Exp: n=2,729
  - Medium Exp: n=4,181
  - Low Exp: n=174

- Found by both PacBio Exclusive OLP replicate 2
  - High Exp: n=3,330
  - Medium Exp: n=6,589
  - Low Exp: n=3,315

- Found by both RefSeq Exclusive OLP replicate 1
  - High Exp: n=7,440
  - Medium Exp: n=7,440
  - Low Exp: n=8,379

- Found by both RefSeq Exclusive OLP replicate 2
  - High Exp: n=3,330
  - Medium Exp: n=6,589
  - Low Exp: n=3,315

- % of total Genes
  - 0
  - 20
  - 40
  - 60
  - 80
  - 100

- Isoforms per gene
  - 0
  - 20
  - 40
  - 60
  - 80
  - 100

- % of novel transcripts iPacBio Exclusive
  - 0
  - 10
  - 20
  - 30
  - 40
  - 50
  - 60

- 239
- 8,782
- 21,934
- 3,674

Transcript expression (log(TPM))
RefSeq detects lots of lowly expressed genes. However it captures an important fraction of highly expressed ones 19% (PacBio 1.5%)
Imposing a higher EXP threshold highlights advantages and disadvantages of PacBio quantification.

- **PacBio**
  - High Exp: n=6,350
  - Medium Exp: n=3,167
  - Low Exp: n=19

- **RefSeq**
  - High Exp: n=2,698
  - Medium Exp: n=4,488
  - Low Exp: n=186

**Ploidy Experiment (log(TPM))**
- 0.0
- 2.5
- 5.0
- 7.5
- 10.0

**OLP replicate 1**
- R² = 0.99
- R² = 0.87

**OLP replicate 2**
- R² = 0.99
- R² = 0.98

**Found by both**
**PacBio Exclusive**
**RefSeq Exclusive**

**% of novel transcript pairs**
- PacBio: 2,111 PacBio exclusive transcripts
- RefSeq: 7,425 PacBio exclusive transcripts
Analysis of Most Expressed Transcript (MET) reveal unaccounted 3’ end variability that is captured by PacBio.

**Dhrs7b - Dehydrogenase / reductase (SDR family) member 7B**

- **ExR** vs. **GIR**
- **Mean Transcript expression (TPM)**
- **Lowest SJ coverage by short reads**
- **Lowest mean coverage of an exon**
- **Distance to reference TTS (nts)**

**ExR vs. GIR**

<table>
<thead>
<tr>
<th>Same MET</th>
<th>Different MET</th>
</tr>
</thead>
</table>

**Lowest SJ coverage by short reads**

<table>
<thead>
<tr>
<th>Same MET</th>
<th>Different MET</th>
</tr>
</thead>
</table>

**Lowest mean coverage of an exon**

<table>
<thead>
<tr>
<th>Same MET</th>
<th>Different MET</th>
</tr>
</thead>
</table>

**Distance to reference TTS (nts)**

<table>
<thead>
<tr>
<th>Same MET</th>
<th>Different MET</th>
</tr>
</thead>
</table>
• PacBio captures a robust fraction of transcripts and genes being expressed and at the same time allow for novel discovery.
• RefSeq captures more transcripts and genes, many lowly expressed and hardly reproducible ones
• However, PacBio TOFU fails to capture many highly expressed genes detected by RefSeq.
• PacBio transcriptome reveals unaccounted for 3’ end variability in known transcripts that hamper RefSeq quantification.

Tardaguila et al  SQANTI: extensive characterization of long read transcript sequences for quality control in full-length transcriptome identification and quantification Pre-print BioRxiv (2017)
4. Functional outcome of alternative splicing: Transcript2GO
T2GO combines SQANTI classification with Genomic, transcript and protein annotation to maximize analytical possibilities

Transcript2GO: assessing the functional outcome of alternative splicing manuscript in preparation
Differential splicing linked to the appearance of sequence motifs on a transcriptome wide scale

**Gene Expression**

**Transcripts Expression**

**ORF collapse**

**Proteins View** - Aligned for Project 'pl1'

<table>
<thead>
<tr>
<th>INPUT</th>
<th>Nuclear Fraction</th>
<th>Cytosolic Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>NPC1 NPC2 OPC1 OPC2</td>
<td>M NPC1 NPC2 OPC1 OPC2</td>
</tr>
</tbody>
</table>

GAPDH

H3-Ac

**Transcript2GO: assessing the functional outcome of alternative splicing**

*manuscript in preparation*
Thanks!

UF
William Farmerie
Eric Triplet
Lauren McIntyre

UCI
Ali Mortazavi
Liz Tseng

Pacbio
Victoria Moreno
Susana Rodriguez

UF UNIVERSITY OF FLORIDA

PRINCIPE FELIPE CENTRES DE INVESTIGACION

SEVENTH FRAMEWORK PROGRAMME

MARIE CURIE

COST

USDA

Ministerio de Educación