Reducing technical variability and bias in RNA-seq data

Francesca Finotello
RNA-Seq is a recent methodology (Nagalakshmi, Science 2008) for transcriptome profiling that is based on Next-Generation Sequencing.

RNA-seq methodology is widely adopted in quantitative transcriptomics and seen as a valuable alternative to microarrays.
RNA-seq data
RNA-seq data
RNA-seq data

RNAs

fragmentation + size selection
RNA-seq data

RNAs → fragmentation + size selection → retrotranscription → cDNAs
RNA-seq data

RNAs → fragmentation + size selection → retrotranscription → cDNAs → amplification
RNA-seq data

1. RNAs
2. Fragmentation + size selection
3. Retrotranscription
4. Amplification
5. Sequencing
6. Reads
Counts
number of reads aligned on a gene
digital measure of gene expression
RNA-seq data

Counts
number of reads aligned on a gene
digital measure of gene expression

<table>
<thead>
<tr>
<th>gene 1</th>
<th>Condition 1</th>
<th>Condition 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene 1</td>
<td>27</td>
<td>80</td>
</tr>
<tr>
<td>gene 2</td>
<td>15</td>
<td>56</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>gene N</td>
<td>50</td>
<td>20</td>
</tr>
</tbody>
</table>

RNAseq data

RNAs → fragmentation + size selection → retrotranscription → cDNAs → amplification → sequencing → reads → mapping → DE analysis
RNA-seq [...] can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets.

- Wang, Nat Methods. 2008
RNA-seq biases

RNA-seq [...] can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets.

- Wang, Nat Methods. 2008

- Read coverage is not uniform along genes/transcripts
**RNA-seq biases**

RNA-seq [...] can capture transcriptome dynamics across different tissues or conditions **without sophisticated normalization** of data sets.

- *Wang, Nat Methods. 2008*

- Read coverage is not uniform along genes/transcripts
- Different samples can be sequenced at different sequencing depths
RNA-seq biases

RNA-seq [...] can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets.

- Wang, Nat Methods. 2008

- Read coverage is not uniform along genes/transcripts
- Different samples can be sequenced at different sequencing depths
- Longer genes are more likely to have higher counts
RNA-seq biases

- RNA-seq [...] can capture transcriptome dynamics across different tissues or conditions without sophisticated normalization of data sets.
  - Wang, Nat Methods. 2008

- Read coverage is not uniform along genes/transcripts
- Different samples can be sequenced at different sequencing depths
- Longer genes are more likely to have higher counts
- Most of reads arise from a restricted subset of highly expressed genes
Outline

• Definition of an alternative approach for computing counts
• Assessment of bias with standard and novel approach
• Evaluation of effects on quantification and differential expression analysis
• Conclusions and future developments
Outline

• Definition of an alternative approach for computing counts
  • Assessment of bias with standard and novel approach
  • Evaluation of effects on quantification and differential expression analysis
  • Conclusions and future developments
New approach \texttt{maxcounts}

- Consider the reads aligned to an exon
- For each exon $i$, in sample $j$
  \[ N_{jip} \] are the number of reads covering exon base $p$
- \texttt{maxcounts} are computed as the maximum of per-base counts:

\[ M_{ji} = \max(N_{jip}) \]
New approach maxcounts

- Consider the reads aligned to an exon
- For each exon $i$, in sample $j$
  \[ N_{jip} \] are the number of reads covering exon base $p$
- \textit{maxcounts} are computed as the maximum of per-base counts:

\[
M_{ji} = \max(N_{jip})
\]

Methods

Reads mapped on reference genomes with TopHat, \textit{not allowing multiple alignments} (\texttt{--g 1} option)

Counts (\textit{totcounts}) and per-base counts computed with bedtools (Quinlan, 2010)

\textit{maxcounts} computed with custom scripts (C++ and Perl)

Differences in sequencing depths corrected via TMM (Robinson, 2010)
Outline

• Definition of an alternative approach for computing counts
• Assessment of bias with standard and novel approach
• Evaluation of effects on quantification and differential expression analysis
• Conclusions and future developments
Data set: Griffith, 2010

Smoothed scatter plot of counts vs. exon length (log-log)
Cubic-spline fit of mean log-counts, bins of 100 exons each

- Length bias also at exon level
Biases exon length

Data set: Griffith, 2010

Smoothed scatter plot of counts vs. exon length (log-log)
Cubic-spline fit of mean log-counts, bins of 100 exons each

\[ r = 0.43 \]

- Length bias also at exon level

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th>Exp. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1 [100 bp]</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>e2 [95 bp]</td>
<td>120</td>
<td>115</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>e100 [2000 bp]</td>
<td>2120</td>
<td>2000</td>
</tr>
<tr>
<td>( \sum ) counts</td>
<td>15 000</td>
<td>10 000</td>
</tr>
</tbody>
</table>

**RPKM**

Reads Per Kilobase of exon model per Million mapped reads

\[
RPKM_{ij} = \frac{N_{ij}}{N_j \cdot 10^6 \cdot L_i \cdot 10^3}
\]
Biases in exon length

Data set: Griffith, 2010

Smoothed scatter plot of counts vs. exon length (log-log)
Cubic-spline fit of mean log-counts, bins of 100 exons each

- Length bias also at exon level
- RPKMs overcorrect

<table>
<thead>
<tr>
<th></th>
<th>Exp. 1</th>
<th>Exp. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>e1</td>
<td>100</td>
<td>80</td>
</tr>
<tr>
<td>e2</td>
<td>120</td>
<td>115</td>
</tr>
<tr>
<td>e100</td>
<td>2120</td>
<td>2000</td>
</tr>
</tbody>
</table>

\[
\text{RPKM}_{ij} = \frac{N_{ij}}{N_j/10^6 \cdot L_i/10^3}
\]
Biases on exon length

Data set: Griffith, 2010

Smoothed scatter plot of counts vs. exon length (log-log)
Cubic-spline fit of mean log-counts, bins of 100 exons each

\[ \text{RPKM} = \frac{N_{ij}}{N_j/10^6 \cdot L_i/10^3} \]

- Length bias also at exon level
- RPKMs overcorrect
- maxcounts strongly reduce length bias
Counts distribution across exons

Data set: Bullard, 2010
Counts distribution across exons

- 3-5% exons contain 50% of counts
- 27-32% exons contain 90% of counts

Data set: Bullard, 2010
Counts distribution across exons

- 3-5% exons contain 50% of counts
- 27-32% exons contain 90% of counts
- 1-3% exons contain 50% counts
- 15-34% exons contain 90% counts

Data set: Bullard, 2010

Data set: Marioni, 2008
Counts distribution across exons

- 3-5% exons contain 50% of counts
- 27-32% exons contain 90% of counts
- 1-3% exons contain 50% counts
- 15-34% exons contain 90% counts

- maxcounts have a less steep curve than totcounts and RPKMs
- i.e. counts are more evenly distributed across exons
Outline

• Definition of an alternative approach for computing counts

• Assessment of bias with standard and novel approach

• Evaluation of effects on quantification and differential expression analysis

• Conclusions and future developments
Quantification spike-in RNAs

Data set: Jiang, 2011

Spike-in RNAs (ERCC Consortium)
- Single-isoforms
- Known sequence and concentration

 totcounts
 RPKMs
 maxcounts
Quantification of spike-in RNAs

Data set: Jiang, 2011
Spike-in RNAs (ERCC Consortium)
- Single-isoforms
- Known sequence and concentration

- All measures have high concordance with concentrations
- Transcripts length 270-2000 nt (performance on shorter transcripts?)
DE analysis log-fold-changes

Data set: Griffith, 2010

DE analysis with edgeR (Robinson, 2010) → log-fold-changes (logFC)
Negative Binomial distribution of data required (no RPKMs)

![Scatter plots showing RNA-seq logFC vs qRT-PCR logFC for totcounts and maxcounts.](image)

- totcounts: RMSD = 1.619
- maxcounts: RMSD = 1.422
DE analysis log-fold-changes

*Data set: Griffith, 2010*

DE analysis with edgeR (Robinson, 2010) → log-fold-changes (logFC)
Negative Binomial distribution of data required (no RPKMs)

\[
\text{RMSD} (\hat{\theta}) = \sqrt{E(\hat{\theta} - \theta)^2}
\]

*maxcounts* have a lower RMSD → higher concordance with qRT-PCR
Outline

• Definition of an alternative approach for computing counts
• Assessment of bias with standard and novel approach
• Evaluation of effects on quantification and differential expression analysis
• Conclusions and future developments
## Conclusions & future developments

<table>
<thead>
<tr>
<th></th>
<th>length bias</th>
<th>count distrib.</th>
<th>tech. variance</th>
<th>spike-in quant.</th>
<th>DE analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>totcounts (std approach)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RPKM</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>maxcounts</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
</tbody>
</table>
Conclusions & future developments

<table>
<thead>
<tr>
<th></th>
<th>length bias</th>
<th>count distrib.</th>
<th>tech. variance</th>
<th>spike-in quant.</th>
<th>DE analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>totcounts (std approach)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RPKM</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>maxcounts</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
</tbody>
</table>

Work in progress and future developments

- Benchmark on more data sets (biological replicates, spike-in RNAs)
Work in progress and future developments

- Benchmark on more data sets (biological replicates, spike-in RNAs)
- Use other DE methods downstream
Conclusions & future developments

- Benchmark on more data sets (biological replicates, spike-in RNAs)
- Use other DE methods downstream
- Aggregate exon *maxcounts* to have a measure at gene/transcript level
Conclusions & future developments

<table>
<thead>
<tr>
<th></th>
<th>length bias</th>
<th>count distrib.</th>
<th>tech. variance</th>
<th>spike-in quant.</th>
<th>DE analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>totcounts (std approach)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RPKM</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>maxcounts</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
</tbody>
</table>

Work in progress and future developments

- Benchmark on more data sets (biological replicates, spike-in RNAs)
- Use other DE methods downstream
- Aggregate exon maxcounts to have a measure at gene/transcript level
- Define a robust pre-processing pipeline to avoid artifacts
### Work in progress and future developments

- Benchmark on more data sets (biological replicates, spike-in RNAs)
- Use other DE methods downstream
- Aggregate exon `maxcounts` to have a measure at gene/transcript level
- Define a robust pre-processing pipeline to avoid artifacts
- Develop an alternative strategy for computing `maxcounts` and implement all versions in a bedtools module

<table>
<thead>
<tr>
<th></th>
<th>length bias</th>
<th>count distrib.</th>
<th>tech. variance</th>
<th>spike-in quant.</th>
<th>DE analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>totcounts (std approach)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>RPKM</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>maxcounts</td>
<td>++</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td>++</td>
</tr>
</tbody>
</table>
Aknowledgements

Enrico Lavezzo
Luisa Barzon
Stefano Toppo
Paolo Fontana
Paolo Mazzon
Barbara Di Camillo
Thank you for your attention!
Annex technical variance

Variance vs. mean of log-counts/RPKMs across technical replicates

Data set: Bullard, 2010

- maxcounts' variance is always lower than totcounts' variance
- RPKMs' variance depends on data set
- Assessment on other data sets

Data set: Griffith, 2010
Annex positional counts
Annex totcount distribution
Annex: maxcount distribution

MIP101 replicates - mean VS variance of counts

log(var) vs log(mean)

log(var) vs log(mean)