Biases in RNA-Seq data

October 15\textsuperscript{th}, 2012

NBIC Advanced RNA-Seq course

Prof. dr. Antoine van Kampen
Bioinformatics Laboratory
Academic Medical Center

Biosystems Data Analysis Group
Swammerdam Institute for Life Sciences

a.h.vankampen@amc.uva.nl
Aim: to provide you with a brief overview of literature about biases in RNA-seq data such that you become aware of this potential problem (and solutions)
Example of RNA-seq bias........
What is the problem?

- Experimental (and computational) biases affect expression estimates and, therefore, subsequent data analysis:
  - Differential expression analysis
  - Study of alternative splicing
  - Transcript assembly
  - Gene set enrichment analysis
  - Other downstream analysis

- We must attempt to avoid, detect and correct these biases
Types of bias

- Gene length

- Mappability of reads
  - e.g., due to sequence complexity

- Position
  - Fragments are preferentially located towards either the beginning or end of transcripts

- Sequence-specific
  - biased likelihood for fragments being selected
  - %GC
Few words about microarrays

- Are not free of bias

- It has taken a decade to understand these biases and to provide solutions
  - Recognition of biases (e.g., by the MicroArray Quality Control (MAQC) consortium) has led to the development of quality control standards

- For RNA-Seq it will also take some time to "understand the data".

- Comparison of microarrays and RNA-Seq may help to identify bias

Malone and Oliver (2011) BMC Biology, 9:34
Normalization for gene length and library size:

RPKM / FPKM
Within one sample

You can’t conclude that gene 2 has a higher expression than gene 1!
Comparison of two samples

transcript 1 (sample 1)

Count = 6, library size = 600

transcript 1 (sample 2)

Count = 12, library size = 1200

You can’t conclude that gene 1 has a higher expression in sample 2 compared to sample 1!
RPKM: Reads per kilobase per million mapped reads

Unit of measurement

\[ RPKM = \frac{\#MappedReads \times \frac{1000 \text{ bases} \times 10^6}{\text{length of transcript} \times \text{Total number of mapped reads}}} \]

- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total read number in the measurement

- This facilitates transparent comparison of transcript levels within and between samples

Rewriting the formula

\[ RPKM = \frac{\frac{\text{number of reads of the region}}{\text{total reads}}}{1,000,000} \times \frac{\text{region length}}{1,000} \]
Examples

\[ \text{RPKM} = \frac{\text{#MappedReads} \times 1000 \text{bases} \times 10^6}{\text{length of transcript} \times \text{Total number of mapped reads}} \]

Example 1:
- 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):
Examples

- RPKM = \#MappedReads \times \frac{1000 \text{ bases} \times 10^6}{\text{length of transcript} \times \text{Total number of mapped reads}}

Example 1:
- 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):
  - \text{RPKM} = 900 \times \frac{1000 \times 10^6}{2500 \times 8 \times 10^6} = 45
Examples

\[ RPKM = \frac{\text{#MappedReads} \times 1000 \text{bases} \times 10^6}{\text{length of transcript} \times \text{Total number of mapped reads}} \]

- **Example 1:**
  - 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):
    \[ RPKM = 900 \times \frac{1000 \times 10^6}{2500 \times 8 \times 10^6} = 45 \]

- **Example 2:**
  - Given a 40M read measurement, how many reads would we expect for a 1 RPKM measurement for a 2kb transcript?
Examples

\[ \text{RPKM} = \#\text{Mapped Reads} \times \frac{1000 \text{ bases} \times 10^6}{\text{length of transcript} \times \text{Total number of mapped reads}} \]

**Example 1:**
- 2500kb transcript with 900 alignments in a sample of 10 million reads (out of which 8 million reads can be mapped):
  \[ \text{RPKM} = 900 \times \frac{1000 \times 10^6}{2500 \times 8 \times 10^6} = 45 \]

**Example 2:**
- Given a 40M read measurement, how many reads would we expect for a 1 RPKM measurement for a 2kb transcript?
  \[ 1 = C \times \frac{1000 \times 10^6}{2000 \times 40 \times 10^6} \rightarrow C = 80 \]
FPKM: Fragments per K per M

What's the difference between FPKM and RPKM?

- Paired-end RNA-Seq experiments produce two reads per fragment, but that doesn't necessarily mean that both reads will be mappable. For example, the second read is of poor quality.

- If we were to count reads rather than fragments, we might double-count some fragments but not others, leading to a skewed expression value.

- Thus, FPKM is calculated by counting fragments, not reads.

Other normalization methods

- **Housekeeping genes** (Bullard et al, 2010)

- **Upper-quartile** (Bullard et al, 2010). Counts are divided by (75th) upper-quartile of counts for transcripts with at least one read

- **TMM** (Robinson and Oshlack, 2010). Trimmed Mean of M values

- **Quantile normalization** (Irizarry et al, 2003; developed for microarrays)

The total RNA production, $S_k$, cannot be estimated directly
- we do not know the expression levels and true lengths of every gene.

Relative RNA production of two samples, $f_k = S_k / S_{k'}$, essentially a global fold change, can more easily be determined.

→ empirical strategy that equates the overall expression levels of genes between samples under the assumption that the majority of them are not differentially expressed.

Weighted trimmed mean of the log expression ratios (trimmed mean of M values (TMM); remove upper and lower x% of data):

$$M_g = \log_2 \frac{Y_{gk}}{Y_{gk'}} \frac{N_k}{N_{k'}}$$

$$A_g = \frac{1}{2} \log_2 \left( \frac{Y_{gk}}{N_k} \cdot \frac{Y_{gk'}}{N_{k'}} \right) \text{ for } Y_g \neq 0$$

normal M,A values:
Quantile Normalization

http://www.people.vcu.edu/~mreimers/OGMDA/normalize.expression.html

Schematic representation of quantile normalization: the value $x$, which is the $\alpha$-th quantile of all probes on chip 1, is mapped to the value $y$, which is the $\alpha$ quantile of the reference distribution $F_2$. 
Gene length bias
Gene length bias

This bias (a) affects comparison between genes or isoforms within one sample and (b) results in more power to detect longer transcripts.

33% of highest expressed genes
33% of lowest expressed genes

Oshlack and Wakefield (2009) Biology Direct, 16, 4
Question: does this bias disappear when we use RPKM?
Mean-variance relationship

- **Red line**: for the one third of shortest genes
- **Blue line**: for the longest genes.
- **Black line**: line of equality

- Plot A: blue/red lines close to line of equality between mean and variance which is what would be expected from a Poisson process.
Mean-variance relationship

- Plot B: Counts divided by gene length (which you do when using RPKM). The short genes have higher variance for a given expression level than long genes.

- Because of the change in variance we are still left with a gene length dependency. → Thus, RPKM does not fully correct
Just to refresh your memory

![Graph showing control and experimental distributions with critical t = 2.03 and power 1-β.]

Reality

<table>
<thead>
<tr>
<th>Measured/Perceived</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct (😊)</td>
<td>Type I False Positive</td>
</tr>
<tr>
<td>False</td>
<td>Type II False Negative</td>
<td>Correct (😊)</td>
</tr>
</tbody>
</table>
Power and gene length bias

- More power to detect longer differentially expressed transcripts

\[ D = X_1 - X_2 \]
\[ S.E.(D) = \sqrt{cN_1L + cN_2L} \]

\[ \delta = \frac{cN_1L - cN_2L}{\sqrt{cN_1L + cN_2L}} \propto \sqrt{L}. \]

\[ \frac{E(D)}{S.E.(D)} = \delta \]

Effect size.

Power is related to \( \delta \)

In the paper they show that \( \delta \) is still related to \( L \) after accounting for gene length.
Gene set enrichment analysis and gene length bias

- This bias affects Gene Set Enrichment analysis
  - In GSEA we compare sets of transcripts that are potentially of different length

- For gene length corrections in this context see:

- Correction at gene level or gene set level
Mappability bias
Mappability bias

- **Uniquely** mapping reads are typically summarized over genomic regions
  - Regions with lower sequence complexity will tend to end up with lower sequence coverage

- **Test:** generate all 32nt fragments from hg18 and align them back to hg18
  - Each fragment that cannot be uniquely aligned is unmappable and its first position is considered an unmappable position
  - Expect: uniform distribution

Result of test

Unexpected because introns are assumed to have lower sequence complexity in general

- Since in RNA-seq we align reads prior to further analysis, this step may already introduce a (slight) bias.
Mappability: dependency on transcript length

Reads corresponding to longer transcripts have a higher mappability.
Mappability: evolutionary conservation and expression level

Note: expression level in lung fibroblasts
Sequence-specific bias 1
RNA-Seq protocol

- Current sequencers require that cDNA molecules represent partial fragments of the RNA

- cDNA fragments are obtained by a series of steps (e.g., priming, fragmentation, size selection)
  - Some of these steps are inherently random
  - Therefore: we expect fragments with starting points approximately uniformly distributed over transcript.
Biases in Illumina RNA-seq data caused by hexamer priming

- Generation of double-stranded complementary DNA (dscDNA) involves priming by random hexamers
  - To generate reads across entire transcript length

- Turns out to give a bias in the nucleotide composition at the start (5’-end) of sequencing reads

- This bias influences the uniformity of coverage along transcript

Hansen et al (2010) NAR, 38(12), e131
but also see:
Li et al (2010) Genome Biology, 11(5), R50
Roberts et al(2011) Genome Biology, 12: R22
Determine the nucleotide frequencies considering all reads: What do we expect??

<table>
<thead>
<tr>
<th>Nucl</th>
<th>Position in read</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1   2   3   4   5   ........  35</td>
</tr>
<tr>
<td>C</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

We would expect that the frequencies for these nucleotides at the different positions are about equal.
Bias

- Frequencies slightly deviate between experiments but show identical relative behavior.
- Distribution after position 13 reflects nt composition of transcriptome.
- Effect is independent of study, laboratory, and organisms.

Apparently, hexamer priming is not completely random.

These patterns do not reflect sequencing error and cannot be removed by 5’ trimming!
Re-weighting scheme 1

- **Aim**
  - Adjust biased nucleotide frequencies at the beginning of the reads to make them similar to distribution of the end of the reads (which is assumed to be representative for the transcriptome)

- **Approach**
  1. Associate weight with each read such that they
  2. down- or up-weight reads with heptamer at beginning of read that is over/under-represented
  ➔ Determine expression level of region by adjusting counts by multiplying them with the weights
Re-weighting scheme 2

(assume read of at least 35nt)

\[ \text{Read} \]

heptamers \( (h) \) at positions i=1,2 of reads

heptamers \( (h) \) at positions i=24..29 of reads

Weights are determined over all possible \( 4^7 = 16.384 \) heptamers

\[
\begin{align*}
\log(w) &= \frac{1}{6} \sum_{i=24}^{29} \hat{p}_{\text{hep}:i}(h) \\
&= \frac{1}{2} \left( \hat{p}_{\text{hep}:1}(h) + \hat{p}_{\text{hep}:2}(h) \right)
\end{align*}
\]
## Application of re-weighting scheme

E.g., indicates that this heptamer (TTGGTTCG) was under-represented, thus count is up-weighted.
Example: gene YOL086C in yeast for WT experiment

Base level counts at each position

Extreme expression values are removed but coverage is still far from uniform
Sequence-specific bias 2
Explanation of next figure

- (A) Sequence logos showing the distribution of nucleotides in a 23bp window surrounding the ends of fragments from an experiment primed with hexamers. The 3’-end sequences are complemented. Counts were taken only from transcripts mapping to single-isoform genes.

- (B) Sequence logo showing normalized nucleotide frequencies after reweighting by initial (not bias corrected) FPKM in order to account for differences in abundance.

- (C) The background distribution for the yeast transcriptome, assuming uniform expression of all single-isoform genes. The difference in 5’ en 3’ distributions are due to the ends being primed from opposite strands.

- Comparing (C) to (A) and (B) shows that while the bias is confounded with expression in (A), the abundance normalization reveals the true bias to extend from 5bp upstream to 5bp downstream of the fragment end.

- Taking the ratio of the normalized nucleotide frequencies (B) to the background (C) for the NNSR dataset gives bias weights (D), which further reveal that the bias is partially due to selection for upstream sequences similar to the strand tags, namely TCCGATCTCT in first-strand synthesis (which selects the 5’ end) and TCCGATCTGA in second-strand synthesis (which selects the 3’-end).

Roberts et al (2011) Genome Biology, 12:R22
5' Fragment End

A. Count Density
B. Normalized Density
C. Expected Density

3' Fragment End

D. Ratio (Bias Weight)

Offset from 5' Fragment End

Offset from 3' Fragment End

Raw counts

FPKM correction

Background distribution

internal to fragment
Bias correction

- Use of statistical model that takes expression and nucleotide bias into account

![Graph showing NM_004684]

- Non-uniform coverage of raw read counts along transcript
- Bias weights. Used to correct raw read counts.
- Large weights correspond to positions with high count.
Use of spike-in standards
Synthetic spike-in standards

- External RNA Control Consortium (ERCC)

- ERCC RNA standards
  - range of GC content and length
  - minimal sequence homology with endogenous transcripts from sequenced eukaryotes

- FPKM normalization

Jiang (2011) Synthetic spike-in standards for RNA-seq experiments. Genome Research
Results suggest systematic bias:
better agreement between the observed read counts from replicates than
between the observed read counts and expected concentration of ERCC’s
within a given library.

Count versus concentration*length (mass)
per ERCC. Pool of 44 2% ERCC spike-in H.
Sapiens libraries

Read counts for each ERCC transcript
in two different libraries of human
RNA-seq with 2% ERCC spike-ins
Transcript-specific sources of error

Fold deviation between observed and expected read count for each ERCC in the 100% ERCC library

- Fold deviation between observed and expected depends on
  - Read count
  - GC content
  - Transcript length

Transcript specific biases affect comparisons of read counts between different RNAs in one library
Read coverage biases: single ERCC RNA

The ERCC RNAs are single isoform with well-defined ends Ideal for measuring transcript coverage

Position effect
Read coverage biases: 96 ERCC RNAs

Average relative coverage along all control RNAs for ERCC spiked in 44 H. sapiens libraries. Dashed lines represent 1 SD around the average across different libraries.

Suggested: drop in coverage at 3’-end due to the inherently reduced number of priming positions at the end of the transcript.
Read coverage biases: sequence-specific heterogeneity

- Could be due to
  - RNA structure (single vs double-stranded template regions) and/or
  - Preparation of the RNA (e.g., nonrandom hydrolysis) or
  - cDNA synthesis (e.g., nonrandomness in “random” hexamer)
Read coverage biases: account for sequence-specific bias through statistical models

These models result in a more even coverage
GC bias
GC-bias in DNA-seq

- Correlation of the Solexa read coverage and GC content. 27mer reads generated from Beta vulgaris BAC ZR-47B15. Each data point corresponds to the number of reads recorded for a 1-kbp window.

- This genome is GC poor

~linear relationship

GC-bias in DNA-seq (1)

Models for GC bias:

- **Fragmentation model**
  - locally, GC counts could be associated with the stability of DNA and the modify the probability of a fragmentation point in the genome

- **Read model**
  - GC content primarily modifies the base-sequencing process (GC explains read count)

- **Full-fragment model**
  - GC content of full fragment determines which fragments are selected or amplified

- **Global model**
  - GC effects on scales larger than the fragment length (e.g., higher-order DNA structure)
GC-bias in DNA-seq (2)

Conclusions from their study

- Not a linear relationship (compare to Dohm et al, and Jiang et al). Instead unimodel relationship

- Dependency between count and GC originates from a biased representation of possible DNA fragments (both high GC and high AT fragments being underrepresented)
  - PCR is the most important cause for GC bias
  - Not the GC content of the reads.

- This dependency is consistent but the exact shape varies considerably across samples, even matched samples

- They argue that models taking GC content and fragment length into account is also important for RNA-seq
GC-bias in DNA-seq (3)

- Single position models
  - Estimate 'mean fragment count' (rate) for individual locations rather than bins.
  - Link fragment count to GC content
Single Position Model (1)

- Mappaple positions along genome are randomly sampled (n~10 million)

Determine GC count in corresponding sliding window $W_{0.4}$

- count #fragments with 5'-end in sampled positions
**Single Position Model (2)**

\[ S_{gc} = \text{stratum with } gc=GC(x+a,l) \quad (x=\text{position, } a=\text{shift, } l=\text{length}) \]

\[ N_{gc} = \text{number of sample positions assigned to } S_{GC} \]

\[ F_{gc} = \text{number of fragments starting (5'-end) at the } x's \text{ in } S_{gc} \]

Estimate \( \lambda_{gc} \) by

\[ \hat{\lambda}_{gc} = \frac{F_{gc}}{N_{gc}}. \]

\[
\begin{array}{c|c|c|c|c|c}
\text{Posi} (N_{gc}) & 0 & 1 & 2 & 3 & 4 \\
\hline
\text{Frag} (F_{gc}) & - & - & 1 & 2 & - \\
\hline
\text{Rate} (\lambda_{gc}) & - & - & 1 & 0.66 & - \\
\end{array}
\]
Single Position Model (3)

\[ \hat{\lambda}_{gc} = \frac{F_{gc}}{N_g} \]

\(W_{0,4}\)

<table>
<thead>
<tr>
<th>gc</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Posi ((N_{gc}))</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Frag ((F_{gc}))</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Rate ((\lambda_{gc}))</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.66</td>
<td>-</td>
</tr>
</tbody>
</table>

unimodel
Model comparison

- Estimated model $W_{a,l}$ (i.e., choice of GC window)
  - Used to generate predicted counts for any genomic region

- Comparison of models (i.e., different $a$, $l$) through $TV$ ($0 \leq TV \leq 1$)
  - **normalized total variation distance**
  - distance between stratified estimated rates $W_{a,l}$ and uniform rate ($U$)
    - global mean rate $\hat{\lambda} = \frac{F}{n}$ ($n=$sampled positions, $F=$total number of mapped fragments)
  - We look for high TV: counts are strongly dependent on GC

$$TV(W_{a,l}, U) = \frac{1}{2\hat{\lambda}} \sum_{gc=0}^{l} \frac{N_{gc}}{n} |\hat{\lambda}_{gc} - \hat{\lambda}|$$
Fragment length models

- To measure effect of fragment lengths
- Fragment length model = single position model
  - but counting fragments of length $s$ only
- Determine model $W_{a,l}^s$
  - only count fragments of length $s$ starting at $x$'s in $S_{gc}$

$$\hat{\lambda}_{gc}^s = \frac{F_{gc}^s}{N_{gc}}$$

- Model the count of fragments using GC in the fragment (not in fixed window of length $l$)
- To reduce impact of local biases a few base pairs from the ends of the fragments are removed: $W_{a,s-a-m}^s$
- GC window grows with fragment length
- In this model we have parameters GC and fragment length
## Predicted rates

- For example, \( W_{a=0,l=25} \)

<table>
<thead>
<tr>
<th>window l and gc=5</th>
<th>3 reads</th>
<th>5 reads</th>
<th>4 reads</th>
</tr>
</thead>
</table>

\[ \hat{\lambda}_{gc} = \frac{F_{gc}}{N_{gc}} \]

- \( F = 12 \)
- \( N = 4 \)
- mean fragment rate = \( \frac{12}{4} = 3 \) (we are just averaging)

\[ \hat{\mu}_x = \begin{cases} \hat{\lambda}_{GC(x+a, l)} & \text{if } x \text{ is uniquely mappable} \\ 0 & \text{otherwise.} \end{cases} \]
Evaluation

- The success of a model \((W_{a,l})\) is evaluated by comparing its predictions with the observed fragment counts.

- For robust evaluation use Mean Absolute Deviation (MAD)

\[ \text{B is set of bins (i.e., non-overlapping windows on genome)} \]
\[ \text{F}_b \] the count of fragments for which 5'-end is inside bin \(b\).

\[
\text{MAD}(F, \hat{\mu}) = \text{avg}_{b \in B} |F_b - \hat{\mu}_b|
\]

\[
\text{CN}_b(F_b, \hat{\mu}_b) = \frac{F_b + \epsilon}{\hat{\mu}_b + \epsilon}
\]

Normalization. (CN=copy number, \(\epsilon=0.1\) account for small counts)
Results (1) – Bin counts

- Two different libraries from same starting DNA
- 10 kb bins
- Unimodal relation
- Same trend but the curves are not aligned.
- This makes a case for single sample normalization
Results (2) – Single position models

- Compare different GC windows through TV
- \( a=0 \), different lengths \( l \)

- Expectation:
  - strongest effect after a few bp (fragmentation effect)
  - after 30-75 bp (read effect)
  - at the fragment length (full-fragment effect)
Strongest effect (TV score) coincides with fragment length ($W_{0.180}$ and $W_{0.295}$).
Corresponding GC curve is very sharp.

Bars on the bottom (left panel): median and 0.05/0.95 quantiles.
Results (3) – Single position models

- Smaller scales ($l=50\text{bp}$) allows to compare GC window that overlaps with read with a GC window that does not
  - $W_{0,50}$ versus $W_{75,50}$
  - Effect of 'read' model is not as large as 'fragment center' model
    - May imply that bias is not driven by base calling or sequencing effects but by the composition of the full fragment
Results (4) – Results of fragment length

- Length of fragment influences shape of the GC curve
  - Interaction between GC and length

- Long fragments tend to have a higher GC count
GC-content bias
Results (5) – Fragmentation effect

- Sequence specific bias (as also observed in RNAseq experiments)
Results (6) – read count correction

- The authors conclude that the fragment model best explains the counts (see figure 7 in paper).
  - However, not including fragment length (thus $W(a,l)$ instead of $W(a,s)$).
Some other stuff.......
Technical bias: cDNA library preparation

DNA library preparation: RNA fragmentation and cDNA fragmentation compared. Fragmentation of oligo-dT primed cDNA (blue line) is more biased towards the 3′ end of the transcript. RNA fragmentation (red line) provides more even coverage along the gene body, but is relatively depleted for both the 5′ and 3′ ends.

RNAseq may detect other RNA species (1)

Tarazona et al (2011) Genome research, 21(12), 2213–23
RNAseq may detect other RNA species (2)

median transcript length in Brain and UHR samples (MAQC) versus read depth

protein coding  pseudo gene  processed transcript  lincRNA

Tarazona et al (2011) Genome research, 21(12), 2213–23
Incorrect base quality values

- Work of DePristo et al. in context of genotyping (exome/genome sequencing)

And more....


Tools

- Picard tools
To conclude

- Be aware of different types of bias
- Try to avoid
- Try to detect
- Try to correct