Some Statistics for RNA-Seq

Paul Eilers

Erasmus Medical Center, Rotterdam
and
Biometris, Wageningen

RNA-seq Course, Wageningen, June 2013
My subject today: RNA-Seq

- Let others find the SNPs: it is “stable” knowledge
- Copy numbers might be interesting
- But results are rather unstructured
- Gene expression is a well-defined subject
- And we can learn a lot from it
- My experience and inspiration: mostly human data
- But also a plant example: Soybean
Why do you want sequencing?

• You have too much money?
• You love large hard disks?
• You crave for long computations times?
• You don’t want to look old-fashioned?
• No, you expect more bang for your bucks!
• Let’s see
Overview

• Basically, sequencing is counting (c)DNA fragments

• Statistics for counts are special

• You will get a little theory and three case studies

• Yeast: two samples, two technical replicates

• Kidney and liver: two samples, 7 technical replicates

• HapMap cell lines, 10 biological samples
Counting

- We count reads per gene (per exon)
- On average 20 - 100 reads per gene
- Number of genes in examples: 30k (human) or 7k (yeast)
- Consider one arbitrary gene
- Probability that one arbitrary read maps to this gene is quite low
  - It is 1/30000 or 1/7000, say 0.01%
- But we have many reads, say one million
- This is the province of Poisson statistics
The Poisson distribution

• If $p$ is the probability of a hit (a read mapped to a gene)

• And $n$ is the total number of reads

• Then you expect $\lambda = pn$ hits

• You will never get exactly $\lambda$ hits

• There will be a range of possible values near $\lambda$

• Each value has a certain probability of occurring

• There is a formula for it, the Poisson distribution

• It only depends on $\lambda$
Poisson pictures

\( \lambda = 5 \)

\( \lambda = 10 \)

\( \lambda = 20 \)

\( \lambda = 100 \)
Poisson formulas

- Formula for the Poisson distribution ($k \geq 0$):

$$\Pr(y = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- R has a function for it: `pr = dpois(k, lambda)`

- The average is $\mu$

- The standard deviation is $\sigma = \sqrt{\mu}$

- This is remarkable

- If you know $\mu$, you also know the uncertainty

- Compare to the normal distribution, where $\mu$ and $\sigma$ are unrelated
Consequences of low counts

- Let’s look at a comparison of two counts
- Rough approximation: $4\sigma$ difference gives a p-value below 0.05
- Compare counts $y$ to given $\lambda$
- Fold change needed: $f = y/\lambda > 1 + 4/\sqrt{\lambda}$
- Example $\lambda = 5, 10, 20, 50, 100$
- Then $f \approx 2.8, 2.3, 1.9, 1.6, 1.4$
- Only large fold changes can be detected with low counts
- And that is for a very optimistic p-value
Poisson in real life

- The Poisson distribution is an idealization.
- Do we see it in real life?
- Surprisingly (to statisticians), yes!
- We are used to seeing over-dispersion: $\sigma > \sqrt{\mu}$
- Examples: the yeast and kidney/liver data sets
The yeast data

- Bioconductor package `yeastRNASeq` contains the reads
- Bioconductor package `Genominator` shows how to map them
- Vignette “Working with the ShortRead Package”
- Resulting table is in the materials for the practical
- Two type of yeast (wild-type and mutant), duplicate runs
- A little over 400k reads per run; 7124 Ensembl gene IDs
Plots of the yeast data: large counts dominate
Plots of logarithms: noisy at low counts
Plots of square roots: a good compromise
How to judge (technical) reproducibility?

- Consider the two wild-type samples
- Put counts in a table, $Y = [y_{ij}]$, with two columns and 7124 rows
- Compute the following
  - sums per row: $a_i$ for row $i$
  - sums per column: $b_j$ for column $j$
  - total of the whole table: $t$
  - expected values for $y_{ij}$: $\mu_{ij} = a_i b_j / t$
  - standardized residuals: $r_{ij} = (y_{ij} - \mu_{ij}) / \sqrt{\mu_{ij}}$
  - $s = \sqrt{\sum_i \sum_j r_{ij}^2 / (m - 1)(n - 1)}.$
    - (table has $m$ rows, $n$ columns)
- Now $s$ should be close to 1, the Poisson limit.
More on reproducibility

- The procedure is like the $\chi^2$ test for contingency tables
- In fact $\chi^2 = \sum_i \sum_j r_{ij}^2$
- But the test is not quite meaningful here
- The size of $s$ matters
- You should first remove rows of $Y$ with all zero counts
- For wild-type yeast we get $s = 1.04$
- For the mutant it is $s = 1.03$
- Excellent results
How many genes are useful?

- Zero counts are useless: about 9% (of the genes)
- Counts below 10 are hardly useful
- For mutant: 29%, for wild-type: 44%
- Counts above 100 are reasonably useful
- For mutant: 12%, for wild-type: 12%
- So you lose a lot of genes
Histograms of yeast counts

Mutant 1
- Counts per gene, linear
- Frequency
0 1000 2000 3000
- Counts per gene, log(y + 1)
- Frequency
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5

Wild–type 1
- Counts per gene, linear
- Frequency
0 1000 3000 5000 7000
- Counts per gene, log(y + 1)
- Frequency
0 1 2 3 4

RNA-seq Course, Wageningen, June 2013
Groups with multiple samples

- The previous reasoning was based on single samples
- Often you will have groups with multiple samples
- If the Poisson assumption is true ...
- Add the counts (per gene) of all samples per group
- Treat them as single “super-samples”
- Check the assumption as for reproducibility
- With a table of counts per group
The Marioni et al. data

- One kidney sample, one liver sample
- Seven technical replicates of each
- Illumina Genome Analyzer
- Counts per gene given as supplementary file
- You will find it in the practice material
Histograms of Marioni data

Kidney
Counts per gene, linear
Frequency
0 10000 20000 30000
0 5000 15000 25000

Kidney
Counts per gene, log(y + 1)
Frequency
0 1 2 3 4
0 2000 6000 10000 14000

Liver
Counts per gene, linear
Frequency
0 20000 40000 60000 80000
0 5000 15000 25000

Liver
Counts per gene, log(y + 1)
Frequency
0 1 2 3 4 5
0 5000 10000 15000
Summary of Marioni data

- 32000 genes, 1.7M reads on average
- Very good technical reproducibility
- Not many useful genes
- Over 40% has zero counts
- About 20% has more than 100 counts
Montgomery et al. data


- Comparison of Affy arrays and sequencing

- HapMap (NA) cell lines

- Hansen et al. (2012) used part of data for Bioconductor `cqn` package

- The package contains a table: 23522 genes and 10 subjects

- It is also in the practice materials
Histogram of counts

NA06985
Counts per gene, log(y + 1)
Frequency
0 1 2 3 4 5
0 2000 4000 6000

NA06994
Counts per gene, log(y + 1)
Frequency
0 1 2 3 4
0 2000 4000 6000 8000

NA07037
Counts per gene, log(y + 1)
Frequency
0 1 2 3 4
0 2000 4000 6000

NA10847
Counts per gene, log(y + 1)
Frequency
0 1 2 3 4 5
0 2000 4000 6000
<table>
<thead>
<tr>
<th>Genes with top ten largest counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000211899</td>
</tr>
<tr>
<td>NA11920  1490</td>
</tr>
<tr>
<td>NA11918  24</td>
</tr>
<tr>
<td>NA11931  127366</td>
</tr>
<tr>
<td>NA12003  3657</td>
</tr>
<tr>
<td>NA12006  54321</td>
</tr>
<tr>
<td>NA12287  954</td>
</tr>
<tr>
<td>ENSG00000019582</td>
</tr>
<tr>
<td>NA11920  75796</td>
</tr>
<tr>
<td>NA11918  9167</td>
</tr>
<tr>
<td>NA11931  52518</td>
</tr>
<tr>
<td>NA12003  56673</td>
</tr>
<tr>
<td>NA12006  91131</td>
</tr>
<tr>
<td>NA12287  27518</td>
</tr>
<tr>
<td>ENSG00000167658</td>
</tr>
<tr>
<td>NA11920  56289</td>
</tr>
<tr>
<td>NA11918  10146</td>
</tr>
<tr>
<td>NA11931  56159</td>
</tr>
<tr>
<td>NA12003  78627</td>
</tr>
<tr>
<td>NA12006  61390</td>
</tr>
<tr>
<td>NA12287  64243</td>
</tr>
<tr>
<td>ENSG00000211895</td>
</tr>
<tr>
<td>NA11920  4575</td>
</tr>
<tr>
<td>NA11918  10482</td>
</tr>
<tr>
<td>NA11931  722</td>
</tr>
<tr>
<td>NA12003  52121</td>
</tr>
<tr>
<td>NA12006  6648</td>
</tr>
<tr>
<td>NA12287  12752</td>
</tr>
<tr>
<td>ENSG00000211890</td>
</tr>
<tr>
<td>NA11920  45407</td>
</tr>
<tr>
<td>NA11918  53</td>
</tr>
<tr>
<td>NA11931  9</td>
</tr>
<tr>
<td>NA12003  29</td>
</tr>
<tr>
<td>NA12006  86</td>
</tr>
<tr>
<td>NA12287  15</td>
</tr>
<tr>
<td>ENSG00000075624</td>
</tr>
<tr>
<td>NA11920  26068</td>
</tr>
<tr>
<td>NA11918  6905</td>
</tr>
<tr>
<td>NA11931  26267</td>
</tr>
<tr>
<td>NA12003  16048</td>
</tr>
<tr>
<td>NA12006  37149</td>
</tr>
<tr>
<td>NA12287  16675</td>
</tr>
<tr>
<td>ENSG00000136167</td>
</tr>
<tr>
<td>NA11920  14872</td>
</tr>
<tr>
<td>NA11918  7403</td>
</tr>
<tr>
<td>NA11931  29946</td>
</tr>
<tr>
<td>NA12003  14018</td>
</tr>
<tr>
<td>NA12006  34726</td>
</tr>
<tr>
<td>NA12287  20874</td>
</tr>
<tr>
<td>ENSG00000099875</td>
</tr>
<tr>
<td>NA11920  8648</td>
</tr>
<tr>
<td>NA11918  2967</td>
</tr>
<tr>
<td>NA11931  10543</td>
</tr>
<tr>
<td>NA12003  32190</td>
</tr>
<tr>
<td>NA12006  32795</td>
</tr>
<tr>
<td>NA12287  13347</td>
</tr>
<tr>
<td>ENSG00000100345</td>
</tr>
<tr>
<td>NA11920  7094</td>
</tr>
<tr>
<td>NA11918  3290</td>
</tr>
<tr>
<td>NA11931  9109</td>
</tr>
<tr>
<td>NA12003  11726</td>
</tr>
<tr>
<td>NA12006  28651</td>
</tr>
<tr>
<td>NA12287  5014</td>
</tr>
<tr>
<td>ENSG00000102962</td>
</tr>
<tr>
<td>NA11920  3981</td>
</tr>
<tr>
<td>NA11918  314</td>
</tr>
<tr>
<td>NA11931  7826</td>
</tr>
<tr>
<td>NA12003  20046</td>
</tr>
<tr>
<td>NA12006  26670</td>
</tr>
<tr>
<td>NA12287  2460</td>
</tr>
</tbody>
</table>
What do we see?

- Enormous variations in counts
- Ratios of 1000 to 1 between individuals
- These are cell lines
- Biological variation cannot be that large
- Reproducibility computation gives $s = 5.8$
- Many times larger than what Poisson says
- This is really worrisome
- About 68% (37%) of counts larger than 10 (100)
Some comparisons between samples
The Soybean data


- "RNA-Seq Atlas of *Glycine max*: A guide to the soybean transcriptome"

- All parts (14) of the plants, as well as seeds

- No replications

- Illumina Genome Analyzer II

- One to 5 million useful reads
Some typical Soybean histograms

- **seed.21DAF**
  - Total counts: 1M
  - Non-zero: 25.2%
  - Over 10: 25.2%
  - Over 100: 2.5%

- **seed.28DAF**
  - Total counts: 2.5M
  - Non-zero: 54.6%
  - Over 10: 31.7%
  - Over 100: 5.2%

- **pod.shell.14DAF**
  - Total counts: 2.8M
  - Non-zero: 58.4%
  - Over 10: 37.8%
  - Over 100: 8.6%

- **seed.42DAF**
  - Total counts: 3.4M
  - Non-zero: 51.7%
  - Over 10: 27.8%
  - Over 100: 3.9%

- **young_leaf**
  - Total counts: 4.1M
  - Non-zero: 59.4%
  - Over 10: 41.8%
  - Over 100: 13%

- **nodule**
  - Total counts: 5.1M
  - Non-zero: 56.1%
  - Over 10: 36.8%
  - Over 100: 11.5%
Soybean summary

• Low counts are the rule

• More than half are zero

• Only 10% or less above 100

• Remarkable and consistent distribution at low counts
Impression of the useful fraction

Number of reads

Percentage above threshold

Sample, ordered by total count

Reads (millions)

Number of reads

Percentage

Percentage above threshold

Sample, ordered by total count

RNA-seq Course, Wageningen, June 2013
The high dynamic range myth

- Severin et. al. (abstract): “… dynamic range of over six orders of magnitude …”

- This is nonsense

- Largest and second largest count: 1,102,650 and 300,802

- Only 0.002% of the counts is larger than 100,000

- And 0.02% is larger than 10,000

- A tiny minority
Conclusions

• Technical reproducibility looks very good
• Biological reproducibility (cell lines) is worrisome
• Low yield of genes with large enough reads
• My conclusion: RNA-Seq is hyped
• It might offer less then you expect