Introduction to RNA-Seq Data Analysis

Dr. Benilton S Carvalho
Department of Medical Genetics
Faculty of Medical Sciences
State University of Campinas
• Material:

• http://tiny.cc/rnaseq

• Slides:

• http://tiny.cc/slidesrnaseq
Tools of Choice

• R and BioConductor:
  – Both created by Robert Gentleman;
  – Open-source tools;
  – Easy to prototype;
  – Communicate with C/C++/Fortran;
About R

• Cross-platform;
• Data analysis and visualization;
• Fast deployment to users;
• Able to interact with C/C++/Fortran;
• Thousands of packages:
  – Descriptive analyses;
  – Clustering and classification;
  – Regression Models and Trees;
  – Visualization;
  – Reproducible research;
  – Etc;
About Bioconductor

• Software infra-structure that uses R;
• Designed for biological data;
• Hundreds of packages:
  – Mass spectrometry;
  – Microarrays;
  – Next Generation Sequencing (NGS);
• Active community:
  – Heavily used by industry;
  – Releases in April and October;
  – Cutting-edge methods.
Illumina Products

MiSeq

HiSeq
Illumina Products

MiSeq
- 2 x 75bp ~ 24h : 3.8Gb
- 2 x 300bp ~ 65h : 15Gb

HiSeq
- 1 x 36bp ~ 29h : 144Gb
- 2 x 50bp ~ 60h : 400Gb
- 2 x 100bp ~ 120h : 800Gb
- 2 x 150bp ~ 144h : 1Tb
Illumina HiSeq X Ten

• Considering the Human Genome @ 30x;
• 320 Genomes per week;
• 1500 Genomes per month;
• 18000 Genomes per year;

• Note: HiSeq 2500 ~ 10 Genomes per week
How does RNA-Seq work?

Pepke et. al. (2009)
How does RNA-Seq work?
Pipeline for Analysis

Raw Data
• (ShortRead)

Quality Assessment
• (Rqc) to be published

Mapping
(Rsubread/gmapr)
• Aligned Reads
• Non-aligned Reads

Downstream Analysis
• goseq

Statistical Modelling
• DESeq2
• edgeR

Count Table
• Rsubread
• GenomicFeatures
Relatively Large Files

• In our pilot experiment (per sample):
  – FastQ: 20GB per strand;
  – BAM: 8GB;
  – Counts: 250KB;
  – Temporary Files: 2 x 20GB per strand;
  – Total: ~ 130GB!

• The example above: RNA-Seq on Rats;

• For Human samples, when sequencing DNA, files are in average 10x bigger;
RAW DATA
Inside a FASTQ File

Instrument
Run ID
Flowcell ID
Lane
Tile number
X in tile
Y in tile
Mate
Fail filter
Control bits
Index seq

[benilton@bioinf1 tmp]$ head -n 4 *
⇒ IC01_GCCAAT_L001_R1.fastq <=
@HMI-ST932:92:C1EU1ACXX:1:1101:1206:2174 1:N:0:GCCAAT
GAAGGCAGCAGGCAGCGCAGCAATTACCACCTCCGCACCCGGAGGATGTGACGAA
+
@@DD3DBFH8?DCGEHIIIIGIIICHGDGGHEGIIIBEDCB>5@CCACB@B
⇒ IC01_GCCAAT_L001_R2.fastq <=
@HMI-ST932:92:C1EU1ACXX:1:1101:1206:2174 2:N:0:GCCAAT
CTGCGGTATCCAGGGCGTCGGCGCTTGAACACTGCTAATTATTTCAGAT
+
@@DDDDDDFBFHGGGGGAAGGHB>FF@FIG@FGEEGIEHE;CEHDEECC
[benilton@bioinf1 tmp]$
The Mystery of the Quality Scores

S - Sanger
   Phred+33, raw reads typically (0, 40)

X - Solexa
   Solexa+64, raw reads typically (-5, 40)

I - Illumina 1.3+
   Phred+64, raw reads typically (0, 40)

J - Illumina 1.5+
   Phred+64, raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).

L - Illumina 1.8+
   Phred+33, raw reads typically (0, 41)
The Mystery of Quality Scores

- Base 1:
  - G/@
- @ = 31
- PHRED = 31
- \(-10\cdot\log_{10}(1-P) = 31\)
- P = 0.9992057
QUALITY ASSESSMENT
FastQC

• We have experience with FastQC, but we are developing our own tool;
• FastQC is Java-based;
• Includes the option of pointing and clicking;
FastQC – Per Base Seq Quality

**Good**

**Poor**
FastQC – Quality Score over All Seqs

**Good**

**Poor**
FastQC – Sequence Content

Good

Poor
FastQC – Sequence Duplication

Good

Poor
Principles of Mapping

• Obtain the reference (genome or transcriptome) for the organism of interest:
  • Mapping to the genome:
    – Allows for identification of novel genes/isoforms
    – Must allow for gaps (really hard)
  • Mapping to the transcriptome:
    – Fast(er)
    – No need for spliced alignments
    – Can’t find novel genes/isoforms
Principles of Mapping

Genome alignment (e.g. align to 23 chromosomes):

Transcriptome alignment (e.g. align to 150,000 known transcripts):
Result of Mapping: SAM/BAM

<table>
<thead>
<tr>
<th>op</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>Alignment match (can be a sequence match or mismatch)</td>
</tr>
<tr>
<td>I</td>
<td>Insertion to the reference</td>
</tr>
<tr>
<td>D</td>
<td>Deletion from the reference</td>
</tr>
<tr>
<td>N</td>
<td>Skipped region from the reference</td>
</tr>
<tr>
<td>S</td>
<td>Soft clip on the read (clipped sequence present in &lt;seq&gt;)</td>
</tr>
<tr>
<td>H</td>
<td>Hard clip on the read (clipped sequence NOT present in &lt;seq&gt;)</td>
</tr>
<tr>
<td>P</td>
<td>Padding (silent deletion from the padded reference sequence)</td>
</tr>
</tbody>
</table>
COUNT TABLE
The BAM isn’t the final file

- BAM files give the location of mapped reads;
- But, per individual, how many reads should be considered as from any particular gene?
- The count table represents this;
- It can be obtained through *GenomicAlignments, HTSeq, Rsubread and EasyRNASeq;*
Count-table Example

<table>
<thead>
<tr>
<th></th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSRNOG00000010603</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENSRNOG000000033787</td>
<td>4289</td>
<td>7831</td>
<td>12489</td>
<td>5904</td>
<td>5033</td>
<td>4619</td>
</tr>
<tr>
<td>ENSRNOG000000014887</td>
<td>3</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>ENSRNOG000000045753</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ENSRNOG000000048290</td>
<td>9</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>ENSRNOG00000001689</td>
<td>233</td>
<td>375</td>
<td>466</td>
<td>489</td>
<td>405</td>
<td>266</td>
</tr>
</tbody>
</table>
STATISTICAL MODELING
What is a model?
Different Transcripts, Rates and Probabilities

Number of fragments: Poisson Distribution
Different Transcripts, Rates and Probabilities

Number of fragments: Poisson Distribution
Characteristics of a Poisson Distribution

- $X \sim \text{Poisson}(\lambda p)$

$P(X = k) = \frac{(\lambda p)^k e^{-\lambda p}}{k!}$

- Mean: $\lambda p$
- Variance: $\lambda p$
Analysis method: GLM

\[
N_{ij} \sim \text{Poisson} \left( \mu_{ij} \right)
\]

\[
\log \mu_{ij} = s_j + \sum_k \beta_{ik} x_{kj}
\]

Noise Part

Expected count of region i in sample j

Deterministic Part

Library size effect

(Differential) effect for region i

Design matrix
Need to account for extra variability

Based on the data of Nagalakshmi et al. Science 2008; slide adapted from Huber;
Characteristics of a Negative Binomial (NB) Distribution

- $X \mid \lambda_p \sim \text{Poisson}(\lambda_p)$
- $\lambda_p \sim \text{Gamma}(a, b)$
- Mean: $\mu$
- Variance: $\mu/\nu$
  \[ 0 < \nu < 1 \]

Current methods for DE use NB model!
Sequencing – Rationale

Biological Replicates

• For subject j, on transcript i:
  \[ Y_{ij} \mid \lambda_{ij} \sim P(\lambda_{ij}) \]

• Different subjects have different rates, which we can model through:
  \[ \lambda_{ij} \sim \Gamma(\alpha, \beta) \]

• This hierarchy changes the distribution of \( Y \):
  \[ Y_{ij} \sim NB \left( \alpha, \frac{1}{1 + \beta} \right) \]
An additional source of variation

\[ N_{ij} | \eta_{ij} \sim \text{Poisson} (\eta_{ij}) \]
\[ \eta_{ij} | \mu_{ij} \sim \text{Gamma} (\beta_1(\mu_{ij}), \beta_2(\mu_{ij})) \]
\[ N_{ij} \sim \text{NB} (\mu_{ij}, \alpha(\mu_{ij})) \]

\[ \log \mu_{ij} = s_j + \sum_k \beta_{ik} x_{kj} \]

Deterministic Part
Library size effect
(Differential) effect for region i
Smooth dispersion-mean relation \( \alpha \)
Summary of the Poisson and Negative Binomial Models

• Poisson($\lambda$):
  – Mean: $\lambda$
  – Variance: $\lambda$

• Negative Binomial ($\alpha, 1/(1+\beta)$):
  – Mean: $\alpha/\beta$
  – Variance: $\alpha(1+\beta)/\beta^2$
    • $= \alpha/\beta + \alpha/\beta^2 = \text{mean} + 1/\alpha * \text{mean}^2$

Shot noise

Biological noise
Example: DE / DEU

FBgn0010909 -

Normalized counts

E023  E021  E019  E017  E015  E013  E011  E009  E007  E005  E003  E001

2555775 2559193 2562612 2566030 2569448 2572867 2576285 2579703 2583122 2586540

treated  untreated
Summary of Models

Treatment ($x_j$) as Covariate

Gene Expression / DESeq

$N_{ij} \sim NB(s_j \mu_{ij}, \alpha(\mu_{ij}))$

$log \mu_{ij} \sim \beta_i^0 + \beta_i^T x_j^T$

Change for treatment

Alternative Exon Usage / DEXSeq

$N_{ijl} \sim NB(s_j \mu_{ijl}, \alpha(\mu_{ijl}))$

$log \mu_{ijl} \sim \beta_i^0 + \beta_{il}^E x_j^E + \beta_i^T x_j^T + \beta_{ijl}^{ET} x_l^E x_j^T$

Expression in control

Fraction of reads falling onto exon l in control

Change to fraction of reads for exon l due to treatment
Variance Shrinkage

Dispersion estimation: shrinkage

\[ \log(\alpha_{\text{gene-est}}) - \log(\alpha_{\text{fit}}) > 2 \sigma_{\text{rob}} \]
Downstream Effect of Shrinkage
Remember the variance effect!

- Variance changes as mean changes...
- This seriously affects visualization;
- It also interferes with comparisons;
- One needs to adjust variance before performing clustering, visualization, PCA;
- DESeq2 has a “regularized log-transformation” method designed for that.
The Truth Statistical Models

- There is no “correct model”;
- Models are approximations of the truth;
- There is a “useful model”;
- Understand the mechanisms of the system for better choices of model alternatives;
THINGS THAT STATISTICIAN SAYS...
The Experiment

• A procedure used to answer the questions;
• Comprised of multiple items:
  – Population;
  – Sample;
  – Hypotheses;
  – Test statistic;
  – Rejection criteria;
Population

- Superset of subjects of interest;
- Ideally, every subject in the population is surveyed;
- Issues with the “census approach”;
Sample

- Select some subjects from the population;
- We refer to this subset as sample;
- Subject in a sample can be called replicate;
- Replicate: technical vs. biological;
Hypotheses

• Sets that define the “underlying truth”;
• Null Hypothesis (H0): default situation.
  – Cannot be proven;
  – Reject (in favor of H1) vs. fail to reject;
• Alternative Hypothesis (H1): alternative (duh!)
  – Complements H0 on the parametric space;
  – Assists on the definition of the rejection criteria.
Examples of Hypotheses

• Comparing expression: Tumor vs. Normal:
  – Expressions on tumor and normal are the same;
  – Expressions on tumor and normal are different;

\[ H_0 : \mu_T = \mu_N \]
\[ H_{1}^a : \mu_T > \mu_N \]
\[ H_{1}^b : \mu_T < \mu_N \]

\[ H_0 : \mu_T = \mu_N \]
\[ H_1 : \mu_T \neq \mu_N \]
Test Statistic

• Summary of the data;
• Built “under H0”;
• Independent of unknown parameters;
• Known distributions;
• Compatibility between data and H0;
Test Statistic

- What the statistician see...

\[
X_{T,i} \sim N(\mu_T, \sigma^2) \quad \bar{X}_T \sim N(\mu_T, \sigma^2/n) \\
X_{N,i} \sim N(\mu_N, \sigma^2) \quad \bar{X}_N \sim N(\mu_N, \sigma^2/n)
\]

If \[ H_0 : \mu_T = \mu_N \]

Then \[ Z = \frac{\bar{X}_T - \bar{X}_N}{\sqrt{2\sigma^2/n}} \sim N(0, 1) \]
Rejection Criteria

• Function of three factors:
  – Test statistic;
  – Hypotheses;
  – Type I Error (False Positive), $\alpha$;

• Determines thresholds used to reject H0:

• Defines what is “extreme” for the experiment;
Rejection Criteria

\[ H_0 : \mu_T = \mu_N \]
\[ H_1 : \mu_T \neq \mu_N \]

\[ Z = \frac{\bar{X}_T - \bar{X}_N}{\sqrt{2\sigma^2/n}} \sim N(0, 1) \]
From Rejection Criteria to P-value!

\[ Z = \frac{\bar{X}_T - \bar{X}_N}{\sqrt{2\sigma^2/n}} \sim N(0, 1) \]
What if we look at multiple p-values at a time?

• On a Gene Expression study, we test often 20K genes for differential expression;
• Each test leads to one p-value;
• Should we trust the p-values in order to make decisions?
What if we look at multiple p-values at a time?

• Can we simulate this?
• Choose an $\alpha$–level;
• Generate two populations with the same pars;
• Run t-test;
• Is the result smaller than $\alpha$?
  – Yes: reject;
  – No: don’t reject;
Multiple Testing

• We are doing high-throughput experiments;
• Comparing thousands of units simultaneously;
• At this scale, we can observe several instances of rare events just by chance:
  – Event A: 1 in 1000 chance of happening;
  – Event B: 999 in 1000 chance of happening;
  – And the experiment is tried 20,000 times;
  – We expect 20 occurrences of Event A to be observed, although Event B is much more likely;
Multiple Testing

- Similar scenario, for example, with DE;
- Most genes are not differentially expressed;
- High-throughput experiments;
- Differential expression is tested for 20K genes;
- Need to protect against false positives;
- Suggestion:
  - use non-specific filtering;
  - use adjusted p-values;
Type I and Type II Errors

Type I Error: You're pregnant

Type II Error: You're not pregnant
Non-Specific Filtering

• The majority of the genes are not differentially expressed – this is the basic hypothesis for normalization;

• If we reduce the number of genes to be tested, the chance of making a wrong decision is reduced;

• Non-Specific filtering refers to removing genes that are clearly not DE without looking at the phenotypic information of the samples;
Using Variance as a Filter

Differentially Expressed

Not-Differentially Expressed
FDR – Benjamini Hochberg (BH)

• Sort the p-values by magnitude;
• Get the adjusted values by

\[ j^* = \max \left\{ j : p_j \leq \frac{j}{m} \alpha \right\} \]
ADDITIONAL STUFF TO REMEMBER!
Useful Facts

- The Law of the Large Numbers guarantees that the larger the sample size is, the closer the sample average is to the actual mean;
- Normality assumption isn’t that important with large sample size;
- The Central Limit Theorem states that the average is asymptotically normal;
Useful Facts

• The Z-score depends on the precise knowledge of the variance term:

\[ Z = \frac{\bar{X} - \mu_0}{\sqrt{\sigma^2/n}} \sim N(0, 1) \]

• Estimating the variance changes the distribution of the test statistic:

\[ T = \frac{\bar{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}} \sim t_n \]
Useful Facts

• The Student’s $t$ distribution is similar to the Normal distribution, but has heavier tails;
• Larger sample size, more d.f.;
• More d.f., closer to Normal;
DO I REALLY NEED A STATISTICIAN BEFORE I EVEN RUN MY EXPERIMENT?
Sample size is crucial

• The larger, the better;
• Ideal $N = (\$ \text{ I have}) / (\$ \text{ it costs})$
• With differential expression, one can observe this more easily;
• RNASeqPower BioConductor package;
About Technology

• Is RNA-Seq really worth it when we consider:
  – Cost,
  – Strategies for analysis, and
  – Technical requirements?
Can my experiment answer the question of interest?

Flow Cell 1
- Group A

Flow Cell 2
- Group B

Flow Cell 3
- Group C

Flow Cell 4
- Group D
Differential Expression Across Groups

Flow Cell Confounded With Group

Flow Cell 1: Group A
Flow Cell 2: Group B
Flow Cell 3: Group C
Flow Cell 4: Group D
Differential Expression Across Groups
Randomize Samples wrt Flow Cell

Flow Cell 1
Flow Cell 2
Flow Cell 3
Flow Cell 4
Differential Expression Across Groups

Barcoding vs. Lane Effect

Flow Cell 1

Flow Cell 2

Flow Cell 3

Flow Cell 4