Differential expression analysis

Alternative exon usage

Wolfgang Huber
EMBL

31 October 2013 - Recife
European Molecular Biology Laboratory (EMBL)

European Intergovernmental Research Organisation

- 20 Member States
- Founded in 1974
- Sites in Heidelberg (D), Cambridge (GB), Roma (I), Grenoble (F), Hamburg (D)
- ca. 1400 staff (≈1100 scientists) representing more than 60 nationalities
EMBL’s five missions

- Basic research
- Development of new technologies and instruments
- Technology transfer
- Services to the member states
- Advanced training
What can *you* do at EMBL?

**Biology**
**Chemistry**
**Physics**
**Mathematics**
**Informatics**
**Engineering**

[www.embl.org/phdprogramme](http://www.embl.org/phdprogramme)
[www.embl.org/postdocs](http://www.embl.org/postdocs)
[www.embl.org/jobs](http://www.embl.org/jobs)
Progress in science is driven by technology

Sequencing - DNA-Seq, RNA-Seq, ChiP-Seq, HiC

Microscopy & remote sensing - molecular interactions and life-cycles in single, live cells

Large scale perturbation libraries - RNAi, drugs

We work on the methods in statistical computing, integrative bioinformatics and mathematical modelling to turn these data into biology.
Research areas

Gene expression
• Statistics - differential expression; alternative exon usage
• 3D structure of DNA (HiC & Co.)
• Single-cell transcriptomics and noise
Simon Anders, Aleksandra Pekoswka, Alejandro Reyes, Jan Swedlow; Tibor Pakozdi

collaborations with L. Steinmetz, P. Bertone, E. Furlong, T. Hiiragi

Cancer Genomics & Precision Oncology
• Somatic mutation detection (incl subclonal)
• Phylogeny inference
Julian Gehring, Paul Pyl

collaborations with C.v.Kalle/M.Schmid, H. Glimm (NCT); J. Korbel

Genetic Interactions, pharmacogenetics (reverse genetics)
• Large-scale combinatorial RNAi & automated microscopy phenotyping
• Cancer mutations & drugs
Joseph Barry, Bernd Fischer, Felix Klein, Malgorzata Oles

collaborations with M.Boutros (DKFZ), T.Zenz (NCT), M. Knop (Uni)

Basics of statistics
• Tools & infrastructure for software ‘publication’
• Teaching
Bernd Klaus, Andrzej Oles

collaborations M.Morgan (FHCRC), R.Gentleman (Genentech)
Two applications of RNA-Seq

• Discovery
  • find new transcripts
  • find transcript boundaries
  • find splice junctions

• Comparison
  Given samples from different experimental conditions, find effects of the treatment on
  • gene expression strengths
  • isoform abundance ratios, splice patterns, transcript boundaries
## Count data in HTS

<table>
<thead>
<tr>
<th>Gene</th>
<th>G1iNS1</th>
<th>G144</th>
<th>G166</th>
<th>G179</th>
<th>CB541</th>
<th>CB660</th>
</tr>
</thead>
<tbody>
<tr>
<td>13CDNA73</td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>A2BP1</td>
<td>19</td>
<td>18</td>
<td>20</td>
<td>7</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>A2M</td>
<td>2724</td>
<td>2209</td>
<td>13</td>
<td>49</td>
<td>193</td>
<td>548</td>
</tr>
<tr>
<td>A4GALT</td>
<td>0</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AAAS</td>
<td>57</td>
<td>29</td>
<td>224</td>
<td>49</td>
<td>202</td>
<td>92</td>
</tr>
<tr>
<td>AACS</td>
<td>1904</td>
<td>1294</td>
<td>5073</td>
<td>5365</td>
<td>3737</td>
<td>3511</td>
</tr>
<tr>
<td>AADACL1</td>
<td>3</td>
<td>13</td>
<td>239</td>
<td>683</td>
<td>158</td>
<td>40</td>
</tr>
</tbody>
</table>

[...]

- RNA-Seq
- ChIP-Seq
- HiC
- Barcode-Seq
- Peptides in mass spec
- ...

Simon Anders
Counting rules

- Count reads, not bases
- Discard a read if
  - it cannot be uniquely mapped
  - its alignment overlaps with several genes
  - the alignment quality score is bad
  - (for paired-end reads) the mates do not map to the same gene
Two biological replicates

Treatment vs control

A = \log_2 \sqrt{N_1 N_2}
The Poisson distribution is used for counting processes.

\[ \lambda = 10 \]

\[ \lambda = 50 \]

\[ \sigma = \sqrt{\lambda} \]

\[ \frac{\sigma}{\mu} \equiv \text{c.v.} = \frac{1}{\sqrt{\lambda}} \]
Analysis method: ANOVA

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

Noise part

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

Systematic part

- \( \mu_{ij} \): expected count of region \( i \) in sample \( j \)
- \( s_j \): library size factor
- \( x_{kj} \): design matrix
- \( \beta_{ik} \): (differential) effect for region \( i \)
For Poisson-distributed data, the variance is equal to the mean.

No need to estimate the variance. This is convenient.

E.g. Wang et al. (2010), Bloom et al. (2009), Kasowski et al. (2010), Bullard et al. (2010), ...

\[
\text{NB: } v \sim \mu^2 \\
\text{Poisson: } v \sim \mu^1
\]

Data: Nagalakshmi et al. Science 2008
So we need a better way

data are discrete, positive, skewed
⇒ no (log-)normal model

small numbers of replicates
⇒ no rank based or permutation methods
⇒ want to use parametric stochastic model to infer tail behaviour (approximately) from low-order moments (mean, variance)

large dynamic range (0 ... $10^5$)
⇒ heteroskedasticity matters
The negative-binomial distribution

\[ P(K = k) = \binom{k + r - 1}{r - 1} p^r (1 - p)^k, \quad r \in \mathbb{R}^+, \ p \in [0, 1] \]

Alternative parameterisation

\[ \alpha = \frac{1}{r} \]
\[ \mu = \frac{pr}{1 - p} \]

Moments

mean = \mu

variance = \mu + \alpha \mu^2

Bioconductor package
DESeq, since 2010
The NB distribution models a Poisson process whose rate is itself randomly varying.

\[ \text{NB}(\mu, \sigma^2 + \mu) = \Lambda(\Gamma(\mu, \sigma^2)) \]
Two component noise model

\[ \text{var} = \mu + c \mu^2 \]

- **Large counts**
  - Biological noise dominant
  - Improve power: more biol. replicates

- **Small counts**
  - Sampling noise dominant
  - Improve power: deeper coverage
Generalised linear model of the negative binomial family

\[ N_{ij} \sim NB(\mu_{ij}, \alpha_{ij}) \]

Noise part

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

Systematic part

\( \mu_{ij} \) expected count of gene \( i \) in sample \( j \)

\( s_j \) library size effect

\( x_{kj} \) design matrix

\( \beta_{ik} \) (differential) expression effects for gene \( i \)
What is a generalized linear model?

\[ Y \sim D( m, s) \]

A GLM consists of three elements:

1. A probability distribution \( D \) (from the exponential family), with mean \( E[Y] = m \) and dispersion \( s \)

2. A linear predictor \( \eta = X \beta \)

3. A link function \( g \) such that \( g(m) = \eta \).

Ordinary linear model: \( g = \) identity, \( D = \) Normal

DESeq(2), edgeR, …: \( g = \log, D = \) Negative Binomial
design with a blocking factor

<table>
<thead>
<tr>
<th>Sample</th>
<th>treated</th>
<th>sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>S2</td>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>S3</td>
<td>no</td>
<td>male</td>
</tr>
<tr>
<td>S4</td>
<td>no</td>
<td>female</td>
</tr>
<tr>
<td>S5</td>
<td>no</td>
<td>female</td>
</tr>
<tr>
<td>S6</td>
<td>yes</td>
<td>male</td>
</tr>
<tr>
<td>S7</td>
<td>yes</td>
<td>male</td>
</tr>
<tr>
<td>S8</td>
<td>yes</td>
<td>female</td>
</tr>
<tr>
<td>S9</td>
<td>yes</td>
<td>female</td>
</tr>
<tr>
<td>S10</td>
<td>yes</td>
<td>female</td>
</tr>
</tbody>
</table>
GLM with blocking factor

\[ K_{ij} \sim NB\left(s_j \mu_{ij}, \alpha_{ij}\right) \]

full model for gene \( i \):

\[
\log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S + \beta_i^T x_j^T
\]

reduced model for gene \( i \):

\[
\log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S
\]
GLMs: Interaction

\[ K_{ij} \sim NB(s_j \mu_{ij}, \alpha_{ij}) \]

full model for gene \( i \):

\[
log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S + \beta_i^T x_j^T + \beta_i^I x_j^S x_j^T
\]

reduced model for gene \( i \):

\[
log \mu_{ij} = \beta_i^0 + \beta_i^S x_j^S + \beta_i^T x_j^T
\]
GLMs: paired designs

• Often, samples are paired (e.g., a tumour and a healthy-tissue sample from the same patient)

• Then, using pair identity as blocking factor improves power.

full model:

\[
\log \mu_{ijl} = \beta_i^0 + \begin{cases} 
0 & \text{for } l = 1(\text{healthy}) \\
\beta_i^T & \text{for } l = 2(\text{tumour}) 
\end{cases}
\]

reduced model:

\[
\log \mu_{ij} = \beta_i^0
\]

\begin{align*}
& \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad \quad i \quad \text{gene} \\
& j \quad \text{subject} \\
& l \quad \text{tissue state}
\end{align*}
Generalized linear models

Simple design:
Two groups, e.g. control and treatment

Common complex designs:
• Designs with blocking factors
• Factorial designs
• Designs with interactions
• Paired designs
GLMs: Dual-assay designs (e.g.: CLIP-Seq + RNA-Seq)

How does affinity of an RNA-binding protein to mRNA change under a (drug, RNAi) treatment?

For each sample, we are interested in the ratio of CLIP-Seq to RNA-Seq reads. How is it affected by treatment?

full model:
\[
\text{count} \sim \text{assayType} + \text{treatment} + \text{assayType} : \text{treatment}
\]

reduced model:
\[
\text{count} \sim \text{assayType} + \text{treatment}
\]

Zarnack et al., Cell 2013
Fotos:

wirtschaft

Schritt halten können. Sie z. B. die Treppen nicht mehr so rasch

Einschränkungen in ihrer Leistungsfähigkeit. Einseitige Bewegung
dem Alter kommt immer mehr
bei Kindern beginnen, die in ihrer freien
schnelllebigen Zeit treten. Atemein-

Atem stark durch unsere Lebensweise

Leistungsfähigkeit. Beeinflusst wird der

Atemsystem ist die Voraussetzung für

Atempädagogik

Einfache Atemübungen aus der Atempädagogik helfen, den Atem frei fließen zu lassen.

Ein gesundes und funktionierendes

Zwerchfell im wahrsten Sinne des
tigen Funktionen der Nase beim Atem-
erwärmt. Auch der Geruch der Atemluft
Luft wird befeuchtet, vorgereinigt und

Nase ein- und ausströmen. Die Nase
Ohne Belastung soll die Luft über die
mühelos zu atmen.

gane gesund zu erhalten und um
können selbst schon viel beachten, um
oft schränken wir uns selber ein. Sie

rechter Haltung. Unterschied bemerkt?

oder auch die Körperhaltung die Atmung

Rippen für die Atmung wichtig sind
Hätten Sie gedacht, dass Brustkorb und

so genannte Übungen unterstützen dies und

Bauch flach halten und einziehen erhöht

Um leichter zu atmen und dadurch leis-

muskulatur und verbraucht Energie.

Stand. Augen schließen. Daumen
holen. Finger wegnehmen und nachspüren.

Dehnung langsam lösen. Mehrmals wieder-

Nasenflügel anheben. Variation

loslassen. Mehrmals wiederholen.

Einatem spontan einströmt. Beim Ausatem

von unten an die Nasenflügel legen und

hüftbreiter Stand. Augen schließen. Daumen

holen. Finger wegnehmen und nachspüren.

Dehnung langsam lösen. Mehrmals wieder-

Nasenspitze legen und nach oben dehnen.

Übung: Nase dehnen

Nun Zeigefinger seitlich der Nasenflügel

hinunter zur Nasenspitze streichen.

Übung: Nase dehnen

– Duft und Hauch: Einatem

– ohne Hilfe der Daumen

Übung: Nase dehnen

– über die Nase von der Nasenwurzel

Nebenhöhlen tritt Erleichterung ein.

gängen und chronisch entzündeten

Tätigkeit an. Bei Schnupfen, engen Nasen-

Wirkung beider Übungen:

Ausatem über den Mund hauchen.

Übung: Nase dehnen

Einatem spontan einströmt. Beim Ausatem

Variation

Loslassen. Mehrmals wiederholen.
Why we discard non-unique alignments

gene A

color condition

treatment condition

gene B
Modelling Variance

To assess the variability in the data from one gene, we have

- the observed standard deviation for that gene
- that of all the other genes

⇒ ridge (Tikhonov) regularisation, empirical Bayes
Dispersion estimation: shrinkage

The dispersion plot in Figure 10 is typical, with the final estimates shrunk from the gene-wise estimates towards the fitted estimates. Some gene-wise estimates are flagged as outliers and not shrunk towards the fitted value. (This outlier detection is described in the man page for `estimateDispersion`). The amount of shrinkage can be more or less than seen here, depending on the sample size, the number of coefficients, the row mean and the variability of the gene-wise estimates.

**E.1 Local dispersion fit**

The local dispersion fit is available in case the parametric fit fails to converge. A warning will be printed that one should use `plotDispEsts` to check the quality of the fit, whether the curve is pulled dramatically by a few outlier points. In this case the two fit types appear to produce similar curves (Figure 11).

**Dispersion outliers:**

\[ \log(\alpha_{\text{gene-est}}) - \log(\alpha_{\text{fit}}) > 2 \sigma_{\text{rob}} \]
Beta (estimated effects): shrinkage
The mechanics: empirical Bayes shrinkage of gene-wise dispersion estimates and of (non-intercept) $\beta$s

$$\hat{\alpha}_{MLE} = \arg\max_{\alpha} \ell(\alpha| y, \hat{\mu})$$

$$CR(\alpha) = -\frac{1}{2} \log(\det(X^tWX))$$

$$\hat{\alpha}_{CR} = \arg\max_{\alpha} (\ell(\alpha| y, \hat{\mu}) + CR(\alpha))$$

$$\text{prior}(\alpha) = \log(f_N(\log(\alpha); \log(\alpha_{fit}), \sigma_{prior}^2))$$

$$\hat{\alpha}_{CR-MAP} = \arg\max_{\alpha} (\ell(\alpha| y, \hat{\mu}) + CR(\alpha) + \text{prior}(\alpha))$$
Outlier robustness

Cook’s distance: Change in fitted coefficients if the sample were removed
regularized log-transformation: visualization, clustering, PCA

Parathyroid data, Haglung et al. 2012
GSEA with shrunken log fold changes

Fly cell culture, knock-down of *pasilla* versus control (Brooks et al., 2011)

Turquoise circles:
Reactome Path “APC/C-mediated degradation of cell cycle proteins”
56 genes, avg LFC: -0.15, p value: 4·10^{-11} (t test)
Genes and transcripts

So far, we looked at read counts per gene.

A gene’s read count may increase because the gene produces more transcripts because the gene produces longer transcripts.

How to look at gene sub-structure?
Alternative isoform regulation

Data: Brooks, ..., Graveley, Genome Res., 2010
## Count table for a gene

number of reads mapped to each exon in a gene

<table>
<thead>
<tr>
<th></th>
<th>treated 1</th>
<th>treated 2</th>
<th>control 1</th>
<th>control 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>E01</td>
<td>398</td>
<td>556</td>
<td>561</td>
<td>456</td>
</tr>
<tr>
<td>E02</td>
<td>112</td>
<td>180</td>
<td>153</td>
<td>137</td>
</tr>
<tr>
<td>E03</td>
<td>238</td>
<td>306</td>
<td>298</td>
<td>226</td>
</tr>
<tr>
<td>E04</td>
<td>162</td>
<td>171</td>
<td>183</td>
<td>146</td>
</tr>
<tr>
<td>E05</td>
<td>192</td>
<td>272</td>
<td>234</td>
<td>199</td>
</tr>
<tr>
<td>E06</td>
<td>314</td>
<td>464</td>
<td>419</td>
<td>331</td>
</tr>
<tr>
<td>E07</td>
<td>373</td>
<td>525</td>
<td>481</td>
<td>404</td>
</tr>
<tr>
<td>E08</td>
<td>323</td>
<td>427</td>
<td>475</td>
<td>373</td>
</tr>
<tr>
<td>E09</td>
<td>194</td>
<td>213</td>
<td>273</td>
<td>176</td>
</tr>
<tr>
<td>E10</td>
<td>90</td>
<td>90</td>
<td>530</td>
<td>398</td>
</tr>
<tr>
<td>E11</td>
<td>172</td>
<td>207</td>
<td>283</td>
<td>227</td>
</tr>
<tr>
<td>E12</td>
<td>290</td>
<td>397</td>
<td>606</td>
<td>368</td>
</tr>
<tr>
<td>E13</td>
<td>33</td>
<td>48</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>E14</td>
<td>0</td>
<td>33</td>
<td>2</td>
<td>37</td>
</tr>
<tr>
<td>E15</td>
<td>248</td>
<td>314</td>
<td>468</td>
<td>287</td>
</tr>
<tr>
<td>E16</td>
<td>554</td>
<td>841</td>
<td>1024</td>
<td>680</td>
</tr>
</tbody>
</table>

[...]
Differential exon usage

msn - mishappen
Ten-m
DEXSeq

\[ K_{ijl} \sim \text{NB}(s_j \mu_{ijl}, \alpha_{il}) \]

counts in gene \( i \), sample \( j \), exon \( l \)

dispersion factor

\[ \log \mu_{ijl} = \beta_i^0 + \beta_{il}^E x_l^E + \beta_{ij}^T x_j^T + \beta_{ijl}^{ET} x_l^E x_j^T \]

expression strength in control

fraction of reads falling onto exon \( l \) in control

change in expression due to treatment

change to fraction of reads for exon \( l \) due to treatment

Bayes

\[ P(M | D) = \frac{P(D | M) P(M)}{P(D)} \]
DEXSeq

test for changes in the (relative) usage of exons:

\[
\frac{\text{number of reads mapping to the exon}}{\text{number of reads mapping to the other exons of the same gene}}
\]
PKCζ - PKMζ

long form: PKC-zeta
N-term.
truncated: PKM-zeta
Differential usage of exons or of isoforms?
<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>DEXSeq 1.1.5</th>
<th>cuffdiff 1.3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>proper comparison, PFC vs CB:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFC 1 – PFC 6</td>
<td>CB 1, CB 2</td>
<td>650</td>
<td>114</td>
</tr>
<tr>
<td>PFC 1, PFC 2</td>
<td>CB 1, CB 2</td>
<td>56</td>
<td>230</td>
</tr>
<tr>
<td>PFC 1, PFC 3</td>
<td>CB 1, CB 2</td>
<td>18</td>
<td>361</td>
</tr>
<tr>
<td>PFC 1, PFC 4</td>
<td>CB 1, CB 2</td>
<td>26</td>
<td>370</td>
</tr>
<tr>
<td>PFC 1, PFC 5</td>
<td>CB 1, CB 2</td>
<td>32</td>
<td>215</td>
</tr>
<tr>
<td>PFC 1, PFC 6</td>
<td>CB 1, CB 2</td>
<td>27</td>
<td>380</td>
</tr>
<tr>
<td>mock comparisons, PFC vs PFC:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PFC 1, PFC 3</td>
<td>PFC 2, PFC 4</td>
<td>3</td>
<td>405</td>
</tr>
<tr>
<td>PFC 1, PFC 2</td>
<td>PFC 3, PFC 4</td>
<td>0</td>
<td>399</td>
</tr>
<tr>
<td>PFC 1, PFC 4</td>
<td>PFC 2, PFC 3</td>
<td>244</td>
<td>590</td>
</tr>
<tr>
<td>PFC 1, PFC 3</td>
<td>PFC 2, PFC 5</td>
<td>2</td>
<td>628</td>
</tr>
<tr>
<td>PFC 1, PFC 2</td>
<td>PFC 3, PFC 5</td>
<td>1</td>
<td>499</td>
</tr>
<tr>
<td>PFC 1, PFC 5</td>
<td>PFC 2, PFC 3</td>
<td>2</td>
<td>555</td>
</tr>
<tr>
<td>PFC 1, PFC 4</td>
<td>PFC 2, PFC 5</td>
<td>2</td>
<td>460</td>
</tr>
<tr>
<td>PFC 1, PFC 2</td>
<td>PFC 4, PFC 5</td>
<td>2</td>
<td>504</td>
</tr>
<tr>
<td>PFC 1, PFC 5</td>
<td>PFC 2, PFC 4</td>
<td>2</td>
<td>308</td>
</tr>
<tr>
<td>PFC 1, PFC 4</td>
<td>PFC 3, PFC 5</td>
<td>10</td>
<td>497</td>
</tr>
<tr>
<td>PFC 1, PFC 3</td>
<td>PFC 4, PFC 5</td>
<td>5</td>
<td>554</td>
</tr>
<tr>
<td>PFC 1, PFC 5</td>
<td>PFC 3, PFC 4</td>
<td>0</td>
<td>353</td>
</tr>
<tr>
<td>PFC 2, PFC 4</td>
<td>PFC 3, PFC 5</td>
<td>1</td>
<td>476</td>
</tr>
<tr>
<td>PFC 2, PFC 3</td>
<td>PFC 4, PFC 5</td>
<td>10</td>
<td>823</td>
</tr>
<tr>
<td>PFC 2, PFC 5</td>
<td>PFC 3, PFC 4</td>
<td>0</td>
<td>526</td>
</tr>
</tbody>
</table>

Table S2: Results of the comparison for the Brawand et al. data.
Figure 1: Splicing graph representation of the four transcript variants of gene CIB3 (Entrez ID 117286). Left: transcript representation. Right: splicing graph representation. Orange arrows are edges corresponding to exons.
Noisy Splicing Drives mRNA Isoform Diversity in Human Cells

Joseph K. Pickrell1*, Athma A. Pai1*, Yoav Gilad1*, Jonathan K. Pritchard1,2*
1 Department of Human Genetics, The University of Chicago, Chicago, Illinois, United States of America, 2Howard Hughes Medical Institute, The University of Chicago, Chicago, Illinois, United States of America

Abstract
While the majority of multiexonic human genes show some evidence of alternative splicing, it is unclear what fraction of observed splice forms is functionally relevant. In this study, we examine the extent of alternative splicing in human cells using deep RNA sequencing and de novo identification of splice junctions. We demonstrate the existence of a large class of low abundance isoforms, encompassing approximately 150,000 previously unannotated splice junctions in our data. Newly-identified splice sites show little evidence of evolutionary conservation, suggesting that the majority are due to erroneous splice site choice. We show that sequence motifs involved in the recognition of exons are enriched in the vicinity of unconserved splice sites. We estimate that the average intron has a splicing error rate of approximately 0.7% and show that introns in highly expressed genes are spliced more accurately, likely due to their shorter length. These results implicate noisy splicing as an important property of genome evolution.

“... we extrapolate that the majority of different mRNA isoforms present in a cell are not functionally relevant, though most copies of a pre-mRNA produce truly functional isoforms.”

Figure 2. An example of splice junctions identified in a gene. In the top panel, we plot the average expression level at each base in a region surrounding HERPUD1. In blue are bases annotated as exonic, and in black are those annotated as not exonic. In the middle panel, we plot the positions of all splice junctions in the region identified in our data. In black are splice junctions that are present in gene databases; in red are those that are not. The number of sequencing reads supporting each junction is written to the right of each junction, and junctions are ordered from top to bottom of the plot according to their coverage. In the bottom panel, we show the gene models in the region from Ensembl. The blue boxes show the positions of exons, and the black lines the positions of introns.

doi:10.1371/journal.pgen.1001236.g002
Regulation of (alternative) exon usage

Data: multiple replicate samples each from:
- 6 primate species (hsa, ppa, ptr, ggo, ppy, mml) X
- 5 tissues (heart, kidney, liver, brain, cerebellum)

Brawand et al. Nature 2011 (Kaessmann Lab, Lausanne, CH)
Tissue and species dependence of relative exon usage

Drift and conservation of differential exon usage across tissues in primate species

Alejandro Reyes\textsuperscript{a,1}, Simon Anders\textsuperscript{a,1}, Robert J. Weatheritt\textsuperscript{b,2}, Toby J. Gibson\textsuperscript{b}, Lars M. Steinmetz\textsuperscript{a,3}, and Wolfgang Huber\textsuperscript{a,3}
Fig. 2. (a) Principal component analysis shows that gene expression (top) groups tightly by tissue, irrespective of species, while exon usage (bottom panels) shows prominent species-to-species variability. The interplay between species and tissue effects is explored further in the middle and right panels, which show PCA analyses of selected subsets.

(b) For each exon, the scatter plot depicts to what extent its variability across samples can be attributed to tissue differences (x axis), and to what extent to the different species origins of the samples (y axis). Exons with significant species effects are indicated by blue dots, those with significant tissue-dependence are shown in green, those in which the pattern of tissue-dependence is conserved are shown in red. The numbers of these different exon categories are shown (false discovery rate: 10%).

(c) The number of exons with conserved tissue-dependent regulation between the species depicted (y axis) are plotted as a function of the observed instances of speciation in million years (x axis).
Fig. 2. Tissue and species effects on exon usage. (A) Principal component analysis shows that gene expression (top) groups tightly by tissue, irrespective of species, while exon usage (bottom panels) shows prominent species-to-species variability. The interplay between species and tissue effects is explored further in the middle and right panels, which show PCA analyses of selected subsets. (B) Variance in the REUCs explained by tissue and by species, respectively. The numbers in the top right corner indicate how many exons, represented by dots, are in the four quadrants delineated by the dashed lines. Exons with CTDU between human and macaque are shown in magenta; exons with CTDU between all species pairs are shown in red. (C) Number of exons whose tissue-dependent usage pattern shows significant conservation between humans and the other primate species, plotted against the time since phylogenetic separation of the species from human, in million years. (D) Pearson correlation coefficient of REUCs across tissues between human and macaque versus tissue-dependent usage (TDU) strength in human. Exons with conserved TDU between human and macaque are plotted in red. The plot shows that if an exon's usage pattern shows strong differences across tissues in one species (here: human), this pattern tends to be the same in other species (here: macaque), indicating conservation of regulation and suggesting functional importance. However, as the histogram of TDU strengths on top shows, these exons represent only a small fraction of all exons.
Conservation: a core set of tissue-dependent exons across primates

Fig. 5. The number of exons with conserved tissue-dependent regulation between the species depicted (y-axis) are plotted as a function of the observed instances of speciation in million years (x-axis).
Strong patterns of tissue-dependent exon usage are frequently conserved.
Functional associations of conserved tissue-dependent exons

Fig. 3. Features of CTDR exons.
(A) Barplot indicating for each category of exons, their fraction overlapping with regions coding for protein disordered regions.
(B) Enrichment of CTDR exons in UTR regions.
(C) Heatmap representation of CDSR exons based on their mean REUC across tissues. The ordering of the rows is based on the exon classification according to their exons' relative inclusion in the different tissues. The four largest classes are colored differently from the rest of the classes (in gray) in the left bar of the heatmap.
(D) Biplot from a canonical correspondence analysis of the splicing factor binding motifs and the mean REUC's from the CTSDR exons across tissues. Each dot corresponds to an exon, the colors correspond to the different clusters of exons indicated in panel C.
Tissue-dependent usage patterns are associated with splicing factor binding motifs and suggest a cis-regulatory code.
Summary tissue-dependent exon usage

Detection of tissue-dependent regulation and its conservation across species at unprecedented scale and precision.

Most of tissue-dependent alternative exon usage in primates is

- low amplitude
- noise
- little evidence for conservation

However, a significant fraction is

- high amplitude
- conserved
- associated with function in mRNA life-cycle & localisation, translation regulation, protein interaction & function
Summary differential expression

• Text-book statistical concepts are (almost) sufficient for differential expression: ANOVA, hypothesis testing, generalized linear models
• In addition: small-n large-p - information sharing across genes, empirical Bayes, shrinkage
• In practice, visualisation (“drill down”) and quality control (batch effects) are very important
• Exon-level analysis
Simon Anders
Joseph Barry
Bernd Fischer
Julian Gehring
Bernd Klaus
Felix Klein
Michael Love
Malgorzata Oles
Aleksandra Pekowska
Paul-Theodor Pyl
Alejandro Reyes
Jan Swedlow

Collaborators
Lars Steinmetz
Robert Gentleman (Genentech)
Michael Boutros (DKFZ)
Martin Morgan (FHCRC)
Jan Korbel
Magnus Rattray (Manchester)

Special thanks
to all users who provided feed-back