University of Zurich[UZH]

Institute of Molecular Life Sciences

Statistical models for count data analysis

Mark D. Robinson, Institute of Molecular Life Sciences

- Reminder of tricks used:
  - conditional likelihood
  - (local) weighted likelihood
- More general framework – GLMs
- Extension to “differential splicing”

http://www.rna-seqblog.com/youve-been-doing-your-rna-seq-all-wrong/

You've been doing your RNA-Seq all wrong

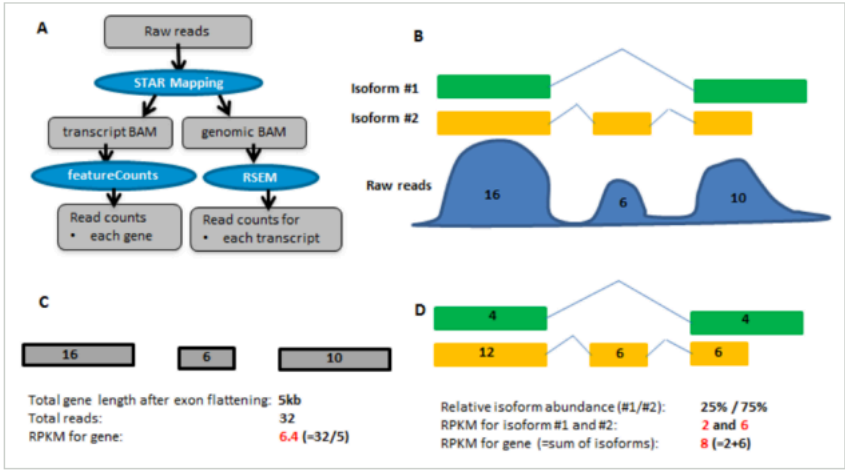
Posted by: RNA-Seq Blog In Expression and Quantification November 12, 2015 5,307 Views

In recent years, RNA-seq is emerging as a powerful technology in estimation of gene and/or transcript expression, and RPKM (Reads Per Kilobase per Million reads) is widely used to represent the relative abundance of mRNAs for a gene. In general, the methods for gene quantification can be largely divided into two categories: transcript-based approach and 'union exon'-based approach. Transcript-based approach is intrinsically more difficult because different isoforms of the gene typically have a high proportion of genomic overlap. On the other hand, 'union exon'-based approach method is much simpler and thus widely used in RNA-seq gene quantification. Biologically, a gene is expressed in one or more transcript isoforms. Therefore, transcript-based approach is

based approach. Despite the fact that gene quantification i

however, it remains unclear whether 'union exon'-based ap

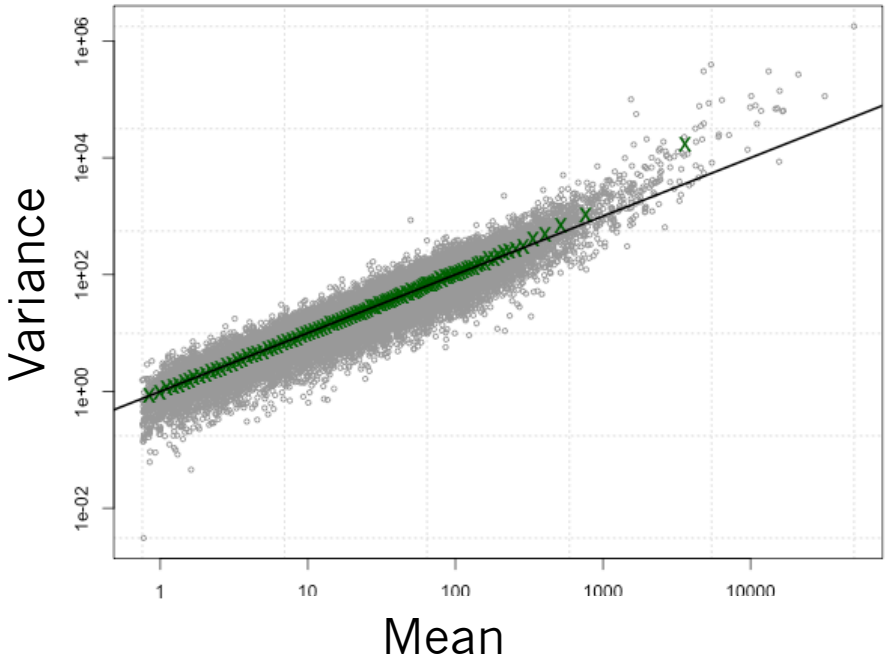
good practice or not.



→ Transcript-based versus "union exon"-based quantification

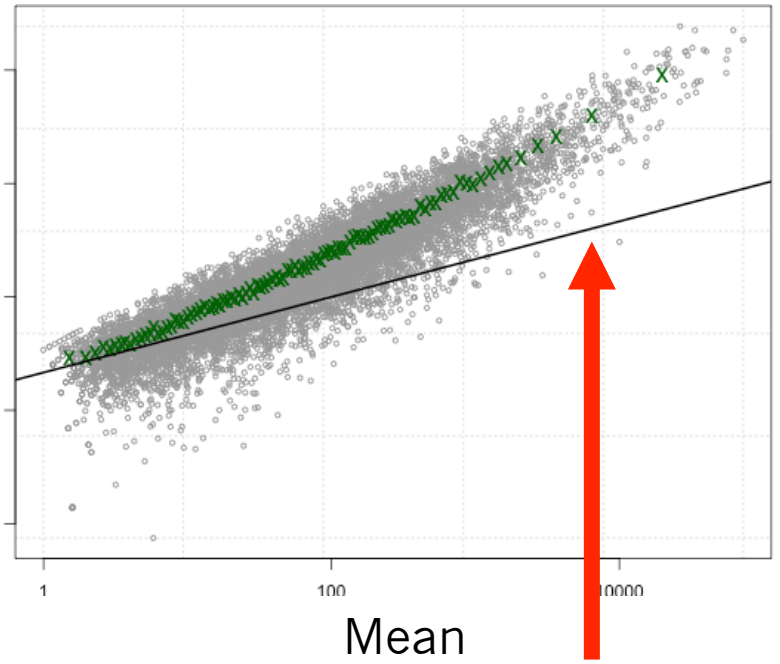
Mean-Variance plots: What we see in real data

Technical replicates



Data from Marioni et al. Genome Research 2008

Biological replicates



Data from Parikh et al. *Genome Biology* 2010

mean=variance
(Poisson assumption)

Moderate dispersion estimate

Weighted likelihood -- individual log-likelihood plus a weighted version of the **common log**-likelihood:

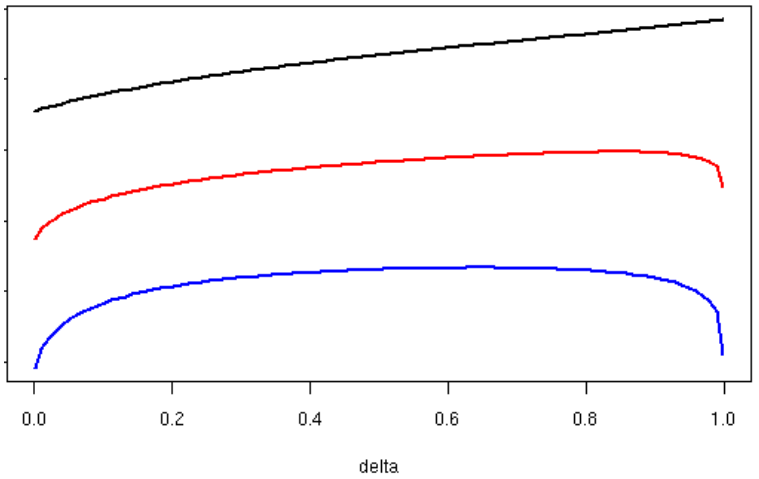
$$\mathrm{WL}(\phi_g) = l_g(\phi_g) + \alpha\, l_C(\phi_g)$$

$(1\text{-}\alpha)$

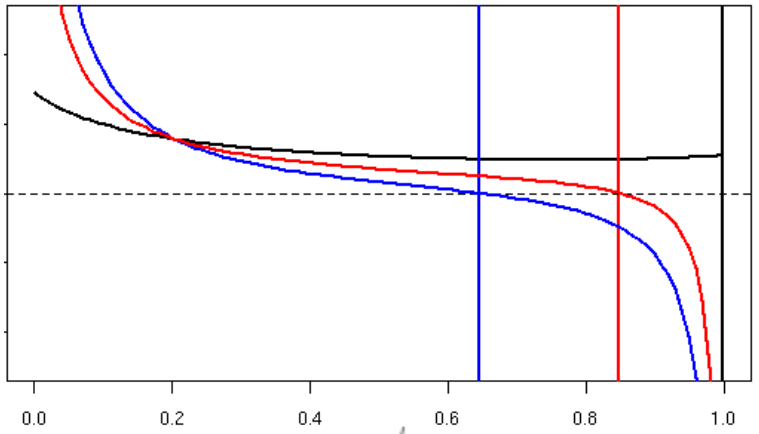
L_g - quantile-adjusted conditional likelihood

Black: single tag
Blue: common dispersion
Red: Linear combination of the two

Log-Likelihood



Score (1st derivative of LL)



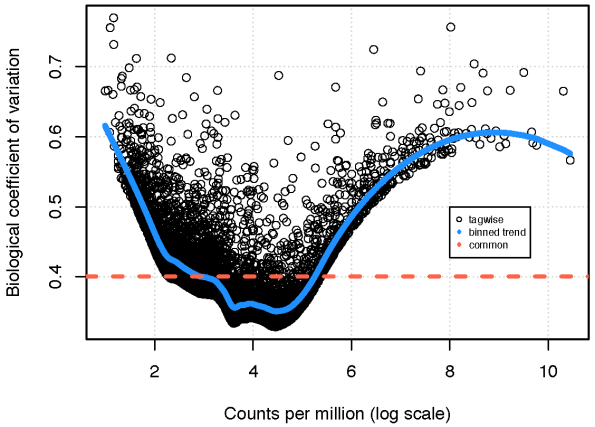
$$\delta = \frac{\phi}{\phi+1}$$

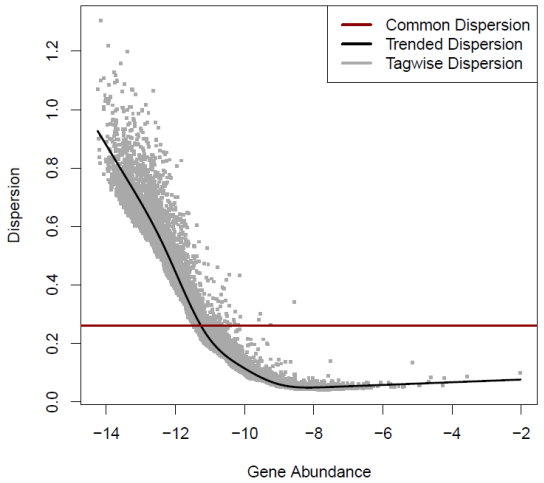
Dispersion varies with mean: moderate (e.g., weighted likelihood) dispersion towards trend

$$\mathrm{WL}(\phi_g) = l_g(\phi_g) + \alpha\, l_C(\phi_g)$$

$(1\text{-}\alpha)$

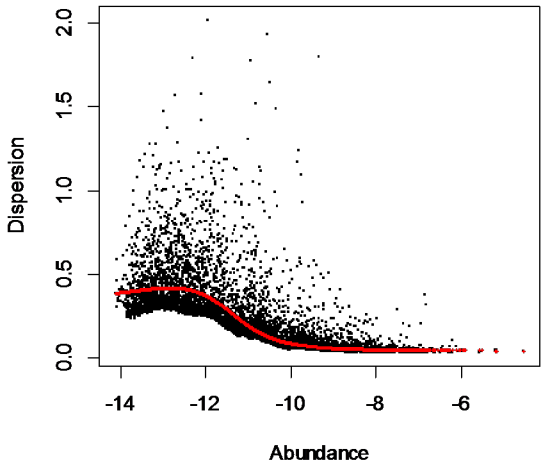
Data:
Tuch et al., 2008





Mouse hemapoeitic stem cells

Advantage: genes are allowed to have their own variance.



Mouse lymphomas

Linear Models (microarray setting)

In general, need to specify:

- Dependent variable
- Explanatory variables (experimental design, covariates, etc.)

More generally:

$$y = X\beta + \epsilon$$

vector of observed data

design matrix

Vector of parameters to estimate

ANOVA as a linear model

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix}$$

b_0 = Group1 expression

α_1 = Group2-Group1 expression

α_2 = Group3-Group1 expression

$E[y_1]=\beta_0$ $E[y_3]=\beta_0+\alpha_1$ $E[y_5]=\beta_0+\alpha_2$

$E[y_2]=\beta_0$ $E[y_4]=\beta_0+\alpha_1$ $E[y_6]=\beta_0+\alpha_2$

Parameter of interest: α_1, α_2 (simultaneously)

Applications: Pairing, multi-factor designs, interactions

→ **This setting only valid for continuous response**

Generalized linear models: a framework

Gaussian (normal) distributed response → various other (common) types.

Three components:

1. probability distribution (in exponential family)
2. Linear predictor (covariates; design matrix)
3. Link function (link mean to linear predictor)

Link function and linear predictor

Design matrix

$$\eta_i = g(\mu_i) = \mathbf{x}_i'\boldsymbol{\beta}.$$

$$\mu_i = g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

Provides a way to link the mean of response to a linear predictor.

Data is not transformed.

Common distributions, "Canonical" link functions

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential Gamma	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = \mu^{-2}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $[0, +\infty)$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $[0, 1]$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $[0, N]$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences			

http://en.wikipedia.org/wiki/Generalized_linear_model

RNA-seq setting – Negative binomial regression

Response is negative binomial (dispersion "fixed" to make it in the exponential family).

Link function (relate mean of response to linear combination of parameters)

For example:

$$Y_i \sim \mathrm{NB}(\mu_i, \phi)$$

$$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$$

X – design matrix
g() – link function (here: log)
β – parameters

`glmFit()`

Same challenge as last time: getting a good estimate of dispersion

$$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$$

$$Y_i \sim \mathrm{NB}(\mu_i, \phi)$$

$$(\hat{\lambda}_{MLE}, \hat{\phi}_{MLE}) = \arg\max_{\lambda,\phi} l(\lambda, \phi)$$

Several choices here:

- Maximum Likelihood (MLE)
- Pseudo-Likelihood (PL)

$$X^2 = \sum_{gij} \frac{(y_{gij} - \hat{\mu}_{gi})^2}{\hat{\mu}_{gi}(1 + \hat{\phi}_{PL}\hat{\mu}_{gi})} = G(n_1 + n_2 - 2)$$

- Quasi-Likelihood (QL)

$$D = 2\sum_{gij}\left\{ y_{gij} \log\left[\frac{y_{gij}}{\mu_{gi}}\right] - (y_{gij} + \phi_{QL}^{-1}) \log\left[\frac{y_{gij} + \phi_{QL}^{-1}}{\mu_{gi} + \phi_{QL}^{-1}}\right]\right\}$$

- ~~Conditional Maximum Likelihood (CML)~~
- Approximate Conditional Inference (Cox-Reid)
- ~~*quantile-adjusted Maximum Likelihood (qCML)*~~

$$Y_i \sim \text{NB}(\mu_i, \phi)$$

$$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$$

Estimation of dispersion parameter

In this setting, we are trying to get an estimate of dispersion, so the beta (regression) parameters are the “nuisance” parameters.

We turn the problem around later to make inferences about the regression parameters.

J. R. Statist. Soc. B (1987)
49, *No.* 1, *pp.* 1–39

Parameter Orthogonality and Approximate Conditional Inference

D. R. COX† and N. REID

Imperial College, London *University of British Columbia, Vancouver*

[*Read before the* Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, 8th October, 1986*, Professor A. F. M. Smith *in the Chair*]

SUMMARY

We consider inference for a scalar parameter ψ in the presence of one or more nuisance parameters. The nuisance parameters are required to be orthogonal to the parameter of interest, and the construction and interpretation of orthogonalized parameters is discussed in some detail. For purposes of inference we propose a likelihood ratio statistic constructed from the conditional distribution of the observations, given maximum likelihood estimates for the nuisance parameters. We consider to what extent this is preferable to the profile likelihood ratio statistic in which the likelihood function is maximized over the nuisance parameters. There are close connections to the modified profile likelihood of Barndorff-Nielsen (1983). The normal transformation model of Box and Cox (1964) is discussed as an illustration.

Keywords: ASYMPTOTIC THEORY; CONDITIONAL INFERENCE; LIKELIHOOD RATIO TEST; NORMAL TRANSFORMATION MODEL; NUISANCE PARAMETERS; ORTHOGONAL PARAMETERS

Cox-Reid adjusted profile likelihood

The adjusted profile likelihood (APL) for ϕ_g is the penalized log-likelihood

$$\mathrm{APL}_g(\phi_g) = \ell(\phi_g; \mathbf{y}_g, \hat{\beta}_g) - \frac{1}{2}\log\det \mathcal{I}_g.$$

where $\mathbf{y}_g$ is the vector of counts for gene g, $\hat{\beta}_g$ is the estimated coefficient vector, $\ell()$ is the log-likelihood function and $\mathcal{I}_g$ is the Fisher information matrix. The

Simply another likelihood, so WL still works

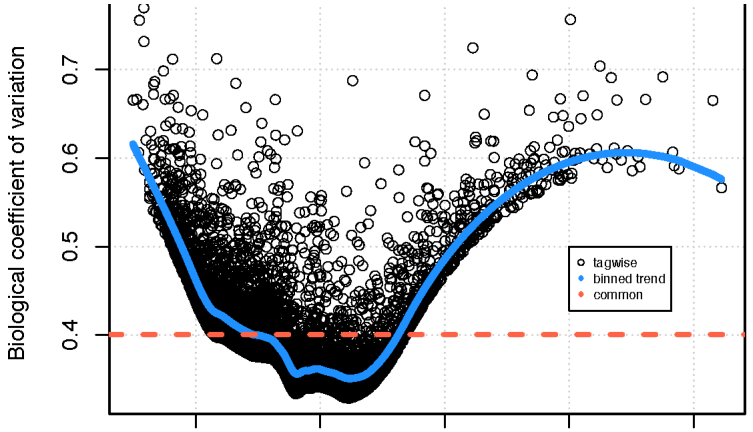
WL is the individual log-likelihood plus a weighted version of the **common log**-likelihood:

$$\mathrm{WL}(\phi_g) = l_g(\phi_g) + \alpha\, l_C(\phi_g)$$

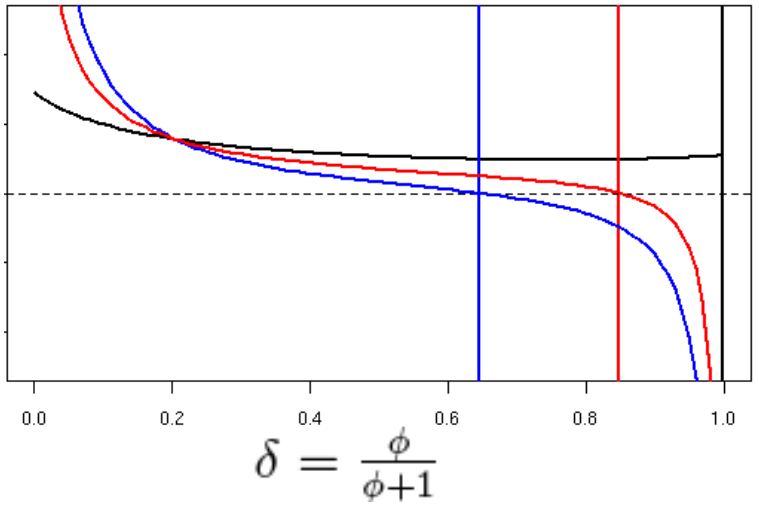
$(1\text{-}\alpha)$

L_g - adjusted profile likelihood (or trended version)

Black: single tag
Blue: common dispersion
Red: Linear combination of the two



Score (1st derivative of LL)



Given dispersion estimates: estimation, statistical testing of regression parameters

Estimation follows a pretty standard framework (details follow)

Options for statistical testing: **Wald**, **Score**, **likelihood ratio**. All of these are based on asymptotics ("large" sample approximations) – how to choose one that works well in practice?

Exponential family (normal distribution)

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}.$$

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{1}{2}\frac{(y_i - \mu_i)^2}{\sigma^2}\}.$$

$$E(Y_i) = b'(\theta_i) = \theta_i = \mu_i,$$
$$\text{var}(Y_i) = b''(\theta_i)a_i(\phi) = \sigma^2.$$

Note: **negative binomial is NOT in exponential family** unless dispersion parameter is treated as fixed.

Optional exercise: what are a(), b() and c() for negative binomial?

A few more GLM details: fitting

Initial estimate: $\hat{\boldsymbol{\beta}}$

can calculate:

$$\hat{\eta}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$$

$$\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$$

Can form "working dependent variable":

$$z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i},$$

$$w_i = p_i / [b''(\theta_i)(\frac{\mathrm{d}\eta_i}{\mathrm{d}\mu_i})^2]$$

Effectively turning the problem into weighted regression, which can follow the steps of IRLS, as seen before:

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{z}$$

Large sample theory – Result 1 (Regression parameter estimates are asymptotically normal)

The Wald test follows immediately from the fact that the information matrix for generalized linear models is given by

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{W}\mathbf{X}/\phi, \tag{B.9}$$

so the large sample distribution of the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$ is multivariate normal

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\phi). \tag{B.10}$$

http://en.wikipedia.org/wiki/Score_(statistics)

$$\mathcal{I}(\theta) = \mathbb{E}\left\{ \left[\frac{\partial}{\partial\theta} \log L(\theta; X) \right]^2 \middle| \theta \right\}.$$

Large sample theory – Result 2 (score is asymptotically normal)

$$\dot{\ell}_1 = \frac{\partial \ell}{\partial \boldsymbol{\theta}_1}$$

$$\dot{\ell}_2 = \frac{\partial \ell}{\partial \boldsymbol{\theta}_2}$$

The "score" function is the first derivative (gradient) of the log-likelihood function, is (asymptotically) normally distributed with mean 0 and variance(-covariance) Fisher information.

Say, we to test H_0: θ_2=0, θ_1 is/are "nuisance" parameter(s)

$$\mathcal{I}_{2.1} = \mathcal{I}_{22} - \mathcal{I}_{21}\mathcal{I}_{11}^{-1}\mathcal{I}_{12}.$$

$$\mathcal{I} = \begin{pmatrix} \mathcal{I}_{11} & \mathcal{I}_{12} \\ \mathcal{I}_{21} & \mathcal{I}_{22} \end{pmatrix}$$

$$S = \dot{\ell}_2^T \mathcal{I}_{2.1}^{-1} \dot{\ell}_2$$

Large sample theory – Result 3 (likelihood ratio test)

$$D = -2 \ln \left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}} \right)$$
$$= -2 \ln(\text{likelihood for null model}) + 2 \ln(\text{likelihood for alternative model})$$

http://en.wikipedia.org/wiki/Likelihood-ratio_test

General form (exponential family)

$$-2 \log \lambda = 2 \sum_{i=1}^{n} \frac{y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)}{a_i(\phi)}.$$

`glmLRT()`

Again, large sample theory says this is approx. χ^2 with degrees of freedom according to the difference in the number of parameters between null and alternative (assuming they are nested).

Likelihood ratio test

edgeR/DESeq use likelihood ratio tests. DESeq2 uses Wald test.

> Genewise tests are conducted by computing likelihood-ratio statistics to compare the null hypothesis that the coefficient or contrast is equal to zero against the two-sided alternative that it is different from zero. The log-likelihood-ratio statistics are asymptotically chi square distributed under the null hypothesis that the co-efficient or contrast is zero. Simulations show that the likelihood ratio tests hold their size relatively well and generally give a good approximation to the exact test (23) when the latter is available (data not shown). Any

McCarthy et al. NAR 2012

Some generalizations of edgeR/DESeq (1)

TweeDESeq

$$Y_{ijk} \sim \text{NBP}(\pi_{ik} m_{jk}, \phi, \alpha).$$

if Y_{ijk} has an NBP distribution, then $\text{Var}(Y_{ijk}) = \mu_{ik}(1 + \phi \mu_{ik}^{\alpha-1})$.

Some generalizations of edgeR/DESeq (2)

$$\lambda = 2(l(\hat{\beta}) - l(\tilde{\beta})),$$

$$r = \text{sign}(\hat{\psi} - \psi_0)\sqrt{\lambda}$$

Higher order approximations

For testing a one-dimensional parameter of interest ($q = 1$), Barndorff-Nielsen (1986, 1991) showed that a *modified directed deviance*

$$r^* = r - \frac{1}{r}\log(z) \qquad (5)$$

is, in wide generality, asymptotically standard normally distributed to a higher order of accuracy than the directed deviance r itself, where z is an adjustment term to be discussed below. Tests based on high-order asymptotic adjustment to the likelihood ratio statistic, such as r^* or its approximation (explained below), are referred to as higher-order asymptotic (HOA) tests. They generally have better accuracy than corresponding unadjusted likelihood ratio tests, especially in situations where the sample size is small and/or when the number of nuisance parameters ($p - q$) is large.

University of Zurich[UZH]

Institute of Molecular Life Sciences

Offset ξ explicitly in GLM:

$$E[Y] = \mu = g^{-1}(\eta) = g^{-1}(X\beta + \xi)$$

Some generalizations of edgeR/DESeq (3): normalization

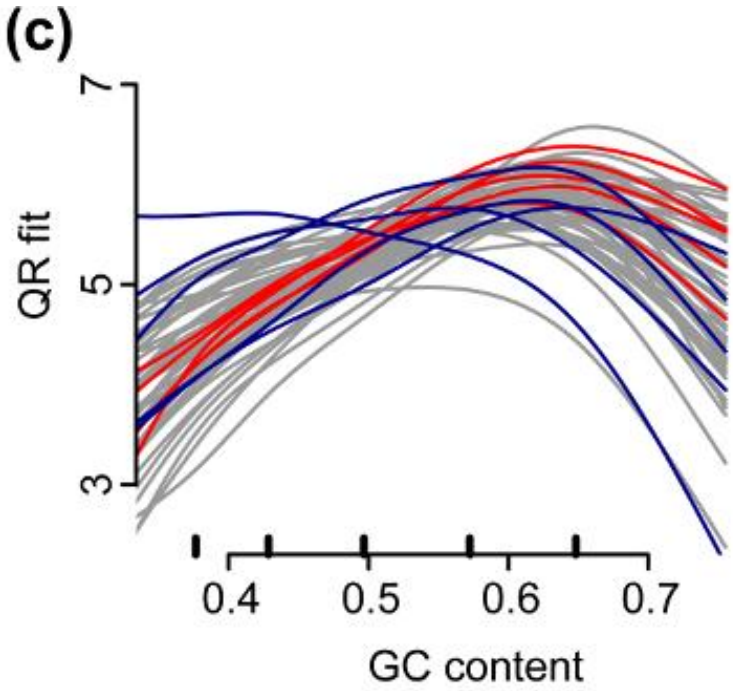


Figure 3. Results from normalizing 60 samples. In these plots we only show genes with a length greater than 100bp and an average (across all 60 samples) standard $\log_2$-RPKM of 2 or greater. (a) Empirical density estimates of $\log_2$-RPKM for five different biological replicates from the Montgomery data are shown. (b) As (a) but CQN normalized expression values on the $\log_2$-scale are shown. (c) The estimated GC-content effect are shown as curves for all 60 biological replicates in the Montgomery study. We created a five versus five comparison using the samples highlighted in blue (group 1) and red (group 2). (d) as (c) but curves are shown for the gene length effect instead of GC-content. (e) Average log-fold-change is plotted against GC-content. Here we used RPKM values and compared group 2 to group 1. (f) Average log-fold-change is plotted against GC-content using CQN normalized expression measures.

Quantile Normalization

Profile vary from sample to sample:
GC content
Gene length

Again, **DOES NOT** change data, use offsets to modify expected mean

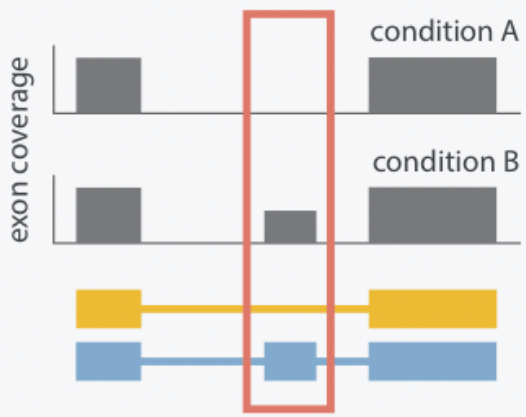
Give a sample-specific offset to edgeR/DESeq

Hansen, Irizarry, Wu
Biostatistics 2012

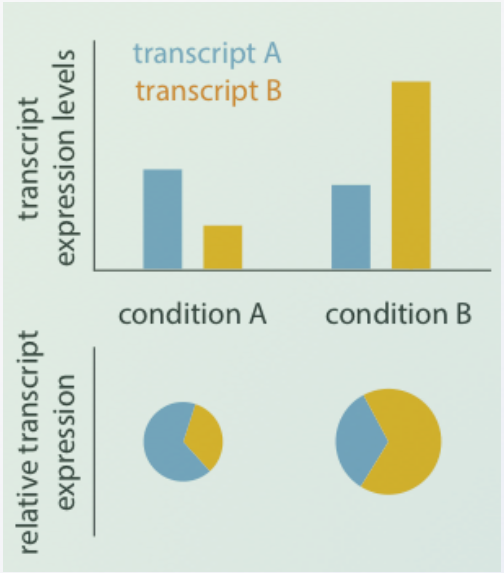
Some terms: DTE, DEU, DTU

Differential transcript expression (DTE)

Differential exon usage (DEU)



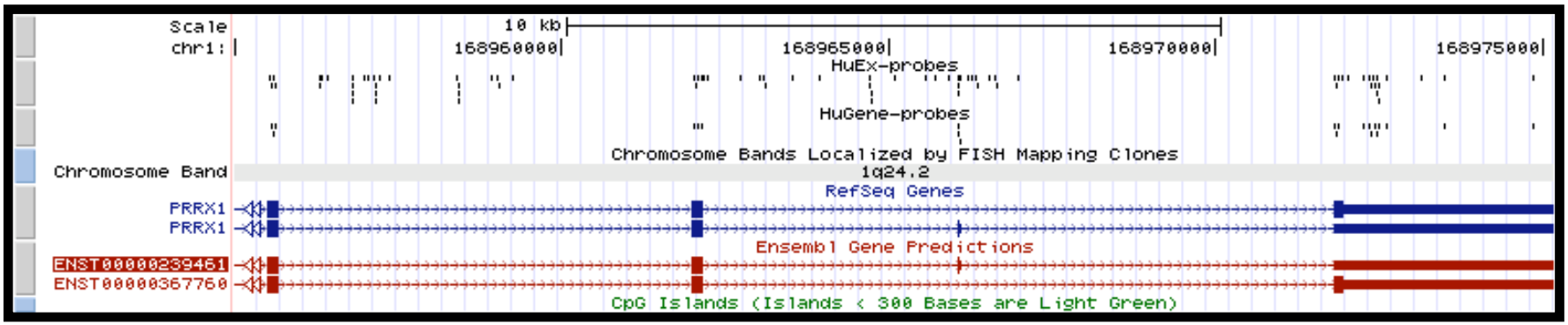
Differential transcript usage (DTU)



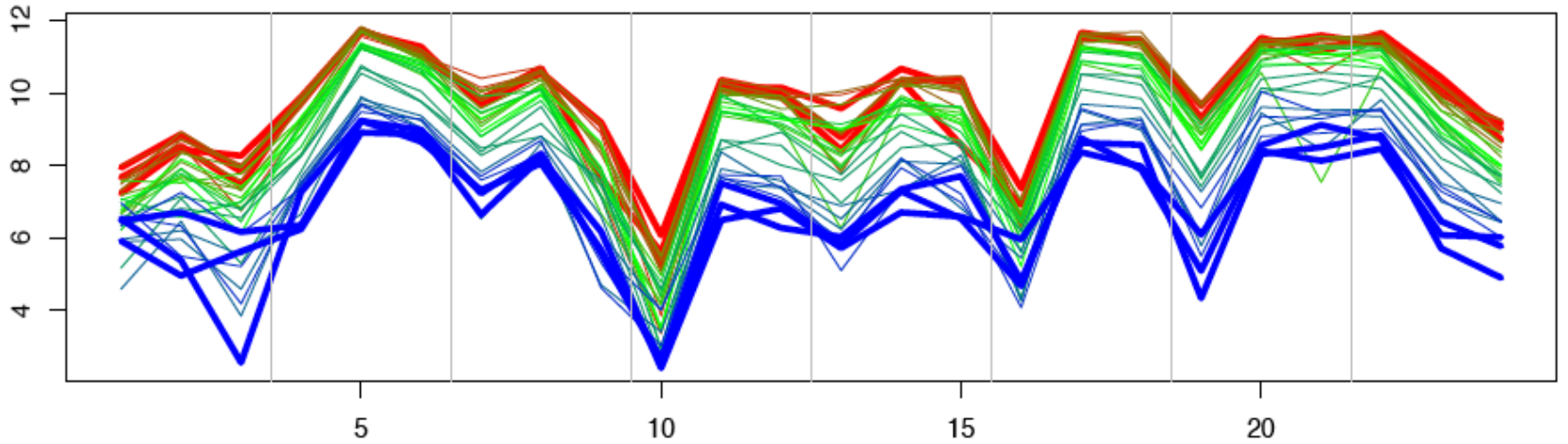
differential splicing

Slide adapted from Mar Gonzàlez-Porta's talk at ECCB 2014
http://radiant-project.eu/ECCB/gonzalez-porta-140907065638-phpapp01.pdf

Digression 1/3: The nature of Affymetrix Probe Level Data

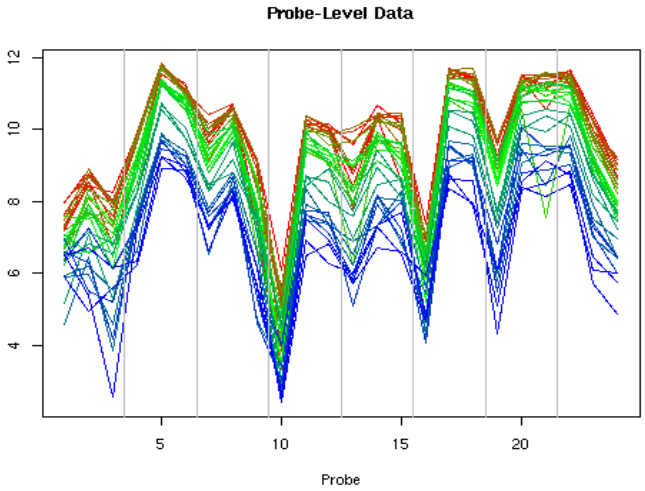


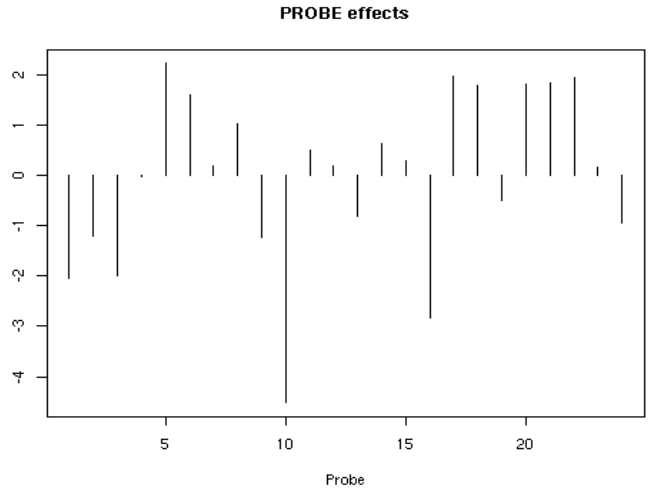
HuGene data [red–heart,blue–brain,mixtures] 10 ENSG00000116132

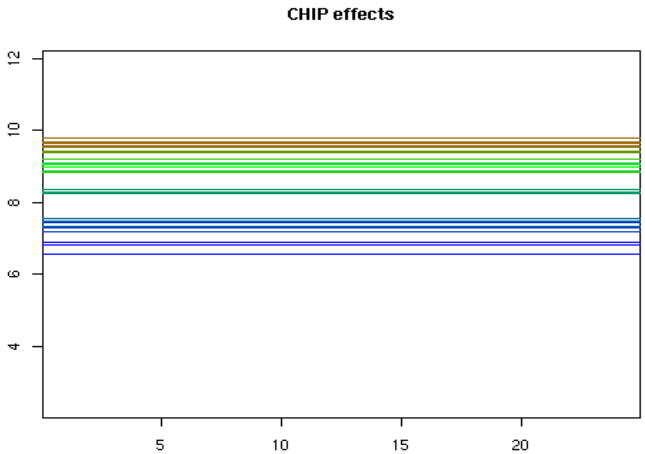


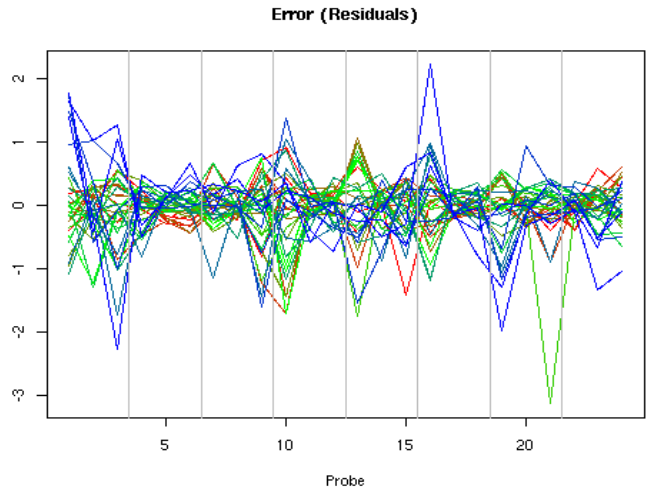
- Data for one gene that is differentially expressed between heart (red is 100% heart) and brain (blue is 100% brain).
- 11 mixtures x 3 replicates = 33 samples (33 lines)
- Note the parallelism: probes have different **affinities**

(Digression 2/3) Differential expression: Affy microarrays



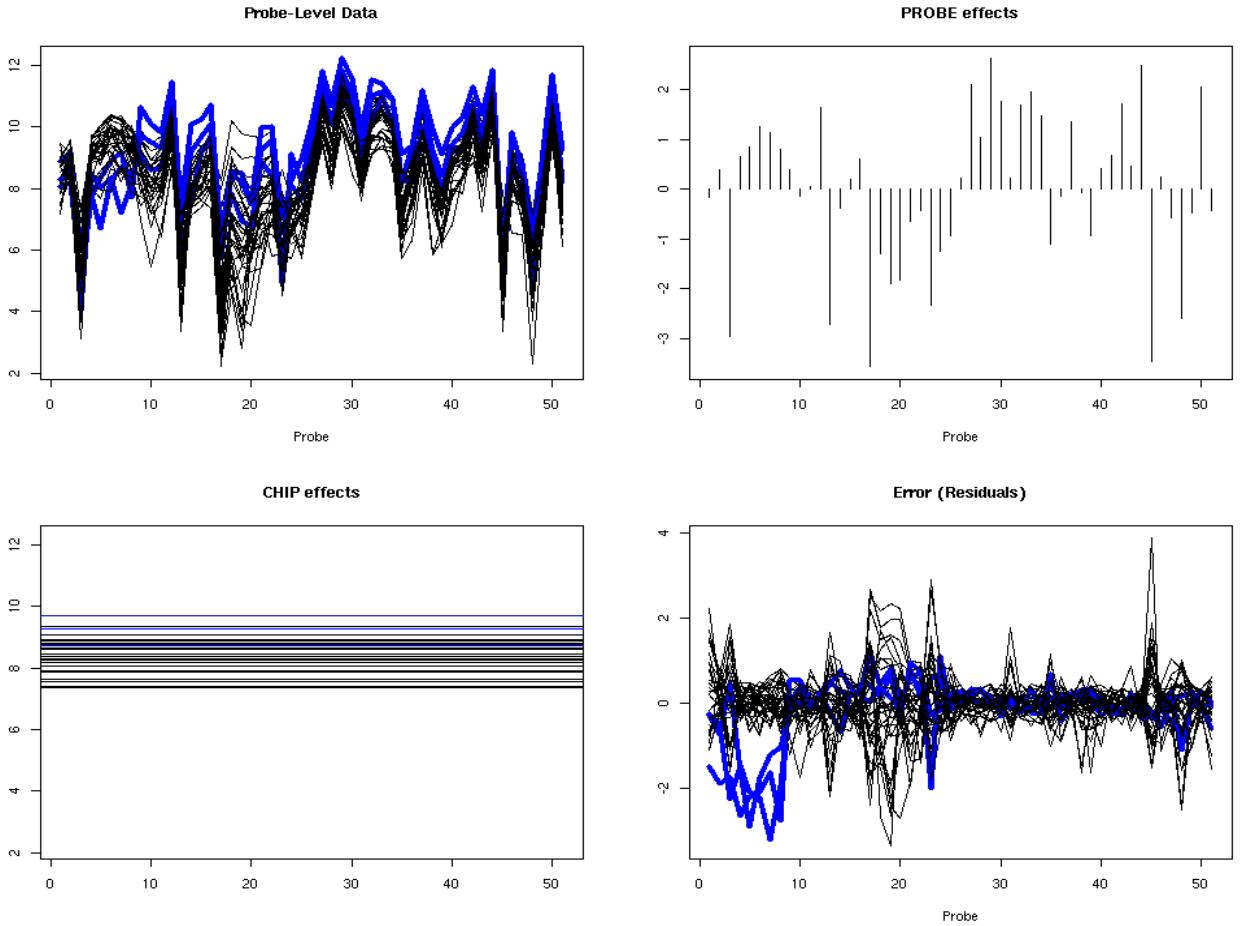






$$y_{ik} = g_i + p_k + e_{ik}$$

Digression 3/3: "Differential splicing" or "Differential isoform usage": Affy microarrays



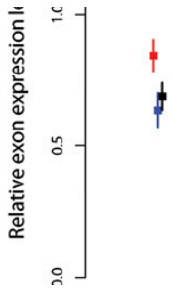
$$y_{ik} = g_i + p_k + e_{ik}$$

(back to RNA-seq) Beyond differential expression: differential splicing

Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments

Hugues Richard[1,*], Marcel H. Schulz[1,2], Marc Sultan[3], Asja Nürnberger[3], Sabine Schrinner[3], Daniela Balzereit[3], Emilie Dagand[3], Axel Rasche[3], Hans Lehrach[3], Martin Vingron[1], Stefan A. Haas[1] and Marie-Laure Yaspo[3]

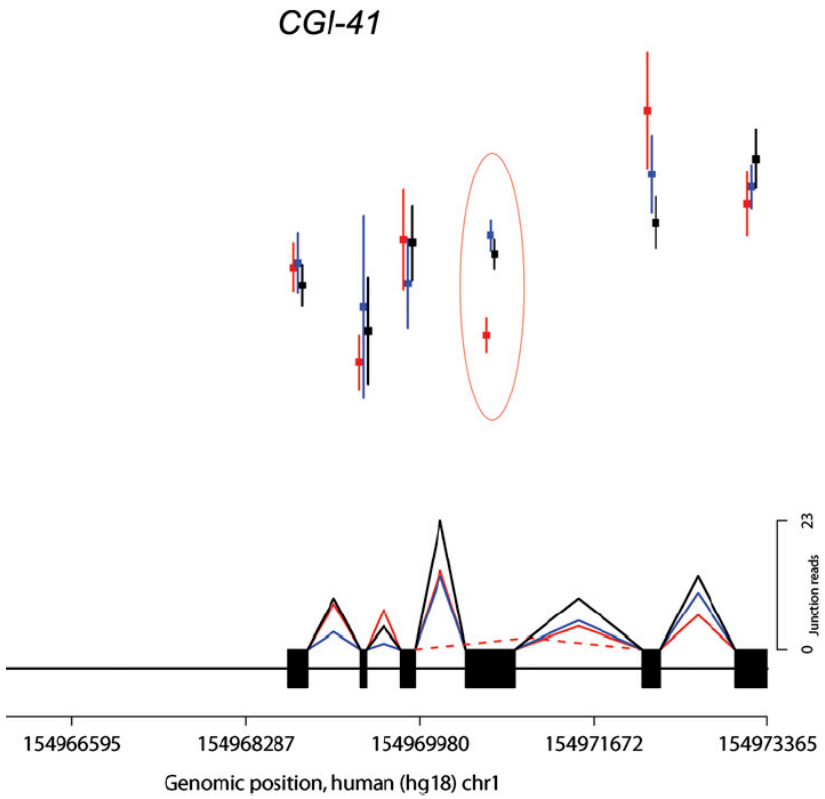
[1]Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestr. 73, [2]International Max Planck Research School for Computational Biology and Scientific Computing, and [3]Department of Vertebrate Genomics, Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195 Berlin, Germany



Sex-specific and lineage-specific alternative splicing in primates

Ran Blekhman,[1,4,5] John C. Marioni,[1,4,5] Paul Zumbo,[2] Matthew Stephens,[1,3,5] and Yoav Gilad[1,5]

[1]Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA; [2]Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA; [3]Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA



Counting: a few considerations (exon-level)

All the downstream statistical methods start with a count table.

How to get one?

- annotation-based? What about novel genes?
- gene-level versus transcript-level? versus exon-level?
- ambiguities
- **junctions?**

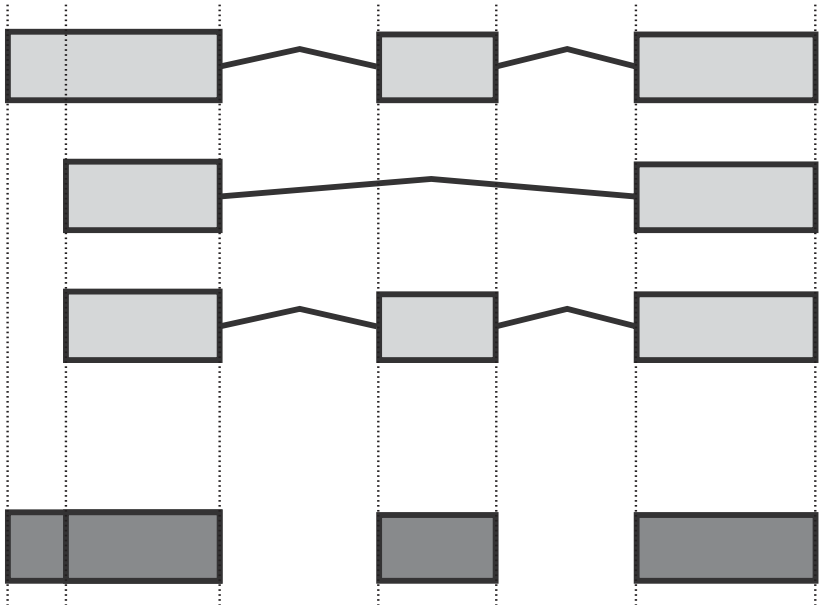


Figure 1. Flattening of gene models: This (fictional) gene has three annotated transcripts involving three exons (light shading), one of which has alternative boundaries. We form counting bins (dark shaded boxes) from the exons as depicted; the exon of variable length gets split into two bins.

Anders et al. 2012 Genome Research

DEXSeq

Method

Detecting differential usage of exons from RNA-seq data

Simon Anders,[1,2] Alejandro Reyes,[1] and Wolfgang Huber

European Molecular Biology Laboratory, 69111 Heidelberg, Germany

Transcript inventory versus differential expression

Shotgun RNA-seq data can be used both for identification of transcripts and for differential expression analysis. In the former, one annotates the regions of the genome that can be expressed, i.e., the exons, and how the pre-mRNAs are spliced into transcripts. In differential expression analysis, one aims to study the regulation of these processes across different conditions. For the method described here, we assume that a transcript inventory has already been defined, and focus on differential expression.

DEXSeq – general structure

We use generalized linear models (GLMs) (McCullagh and Nelder 1989) to model read counts. Specifically, we assume K_{ijl} to follow a negative binomial (NB) distribution:

$$K_{ijl} \sim NB\left(\text{mean} = s_j \mu_{ijl}, \text{dispersion} = \alpha_{il}\right), \qquad (1)$$

where α_{il} is the dispersion parameter (a measure of the distribution's spread; see below) for counting bin (i, l), and the mean is predicted via a log-linear model as

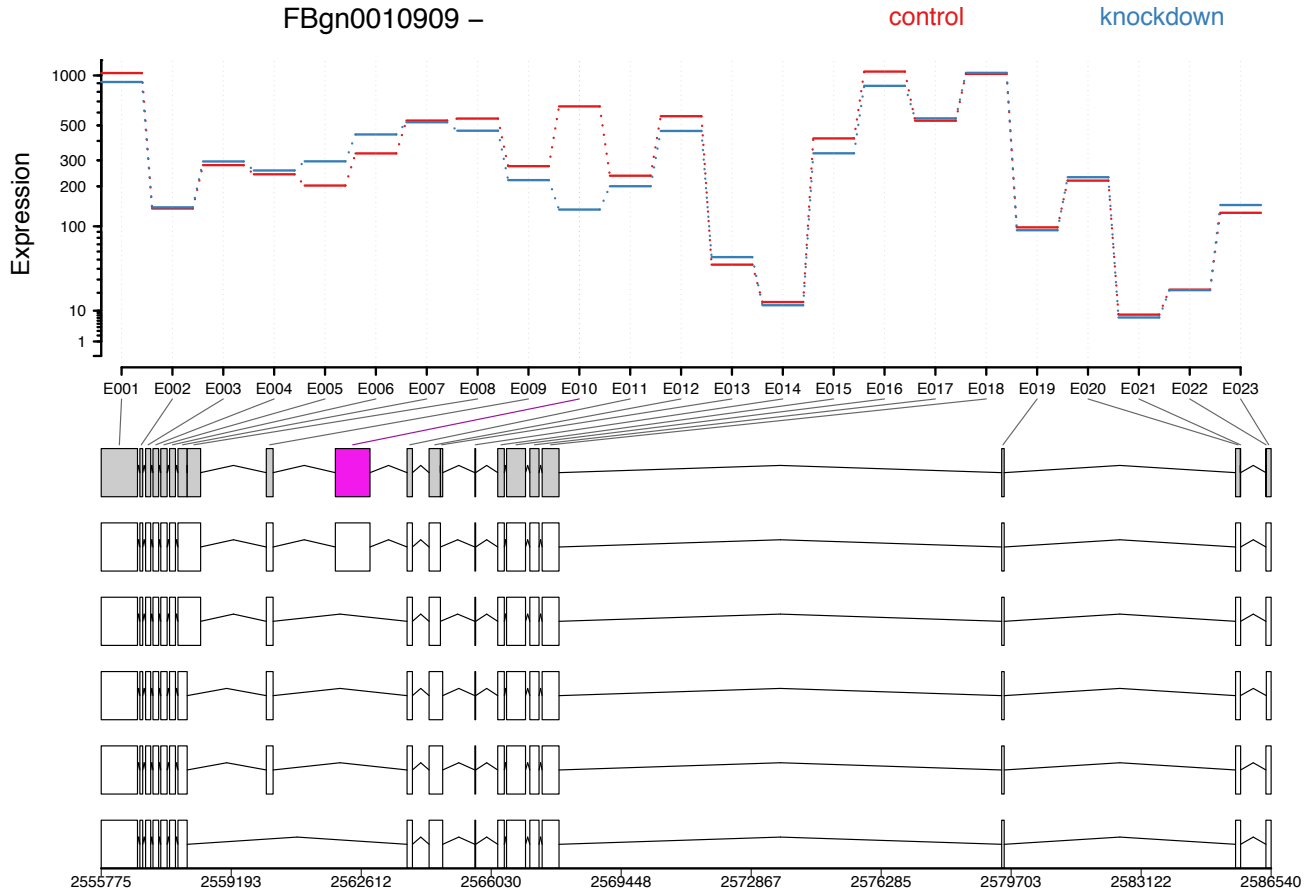
$$\log \mu_{ijl} = \beta_i^{\mathrm{G}} + \beta_{il}^{\mathrm{E}} + \beta_{i\rho_j}^{\mathrm{C}} + \beta_{i\rho_j l}^{\mathrm{EC}}. \qquad (2)$$

i – gene
j – sample … ρ_j is condition (categorical)
l – bin

β^G – baseline "expression strength"
β^E – "exon" (bin) effect
β^C – condition effect
β^{EC} – condition x "exon" interaction

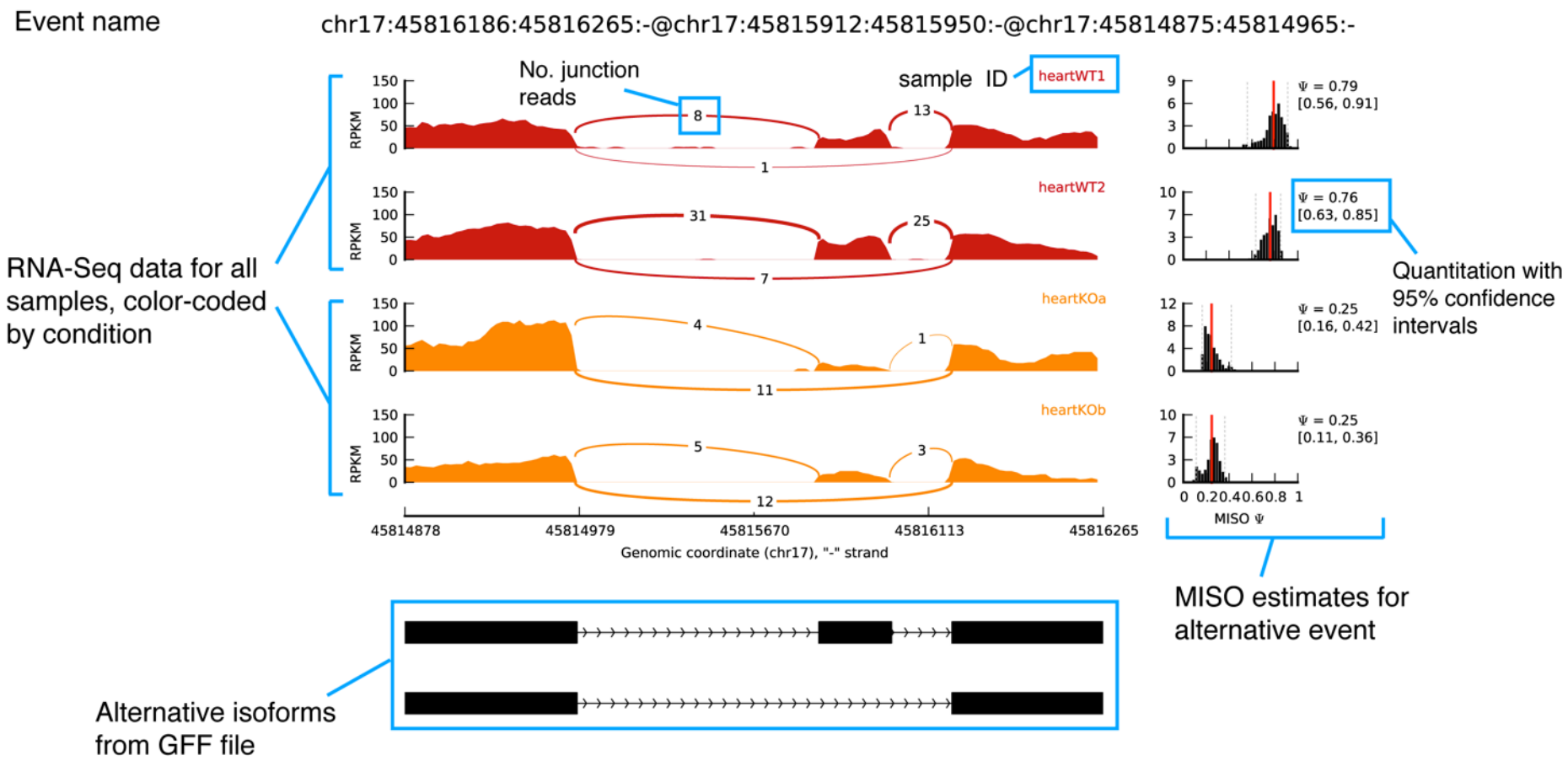
DEXSeq: sig. interaction terms = differential exon usage

(DEXSeq vignette)



Katz et al., Nature Methods, 2010

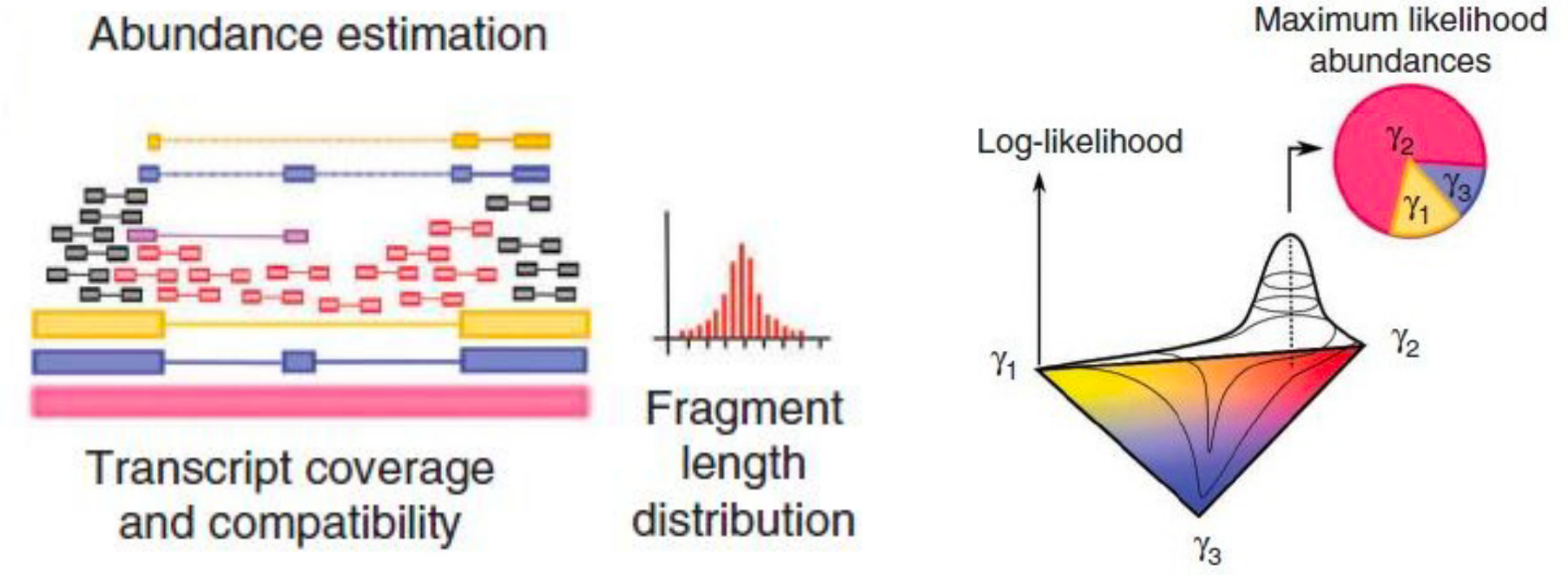
Percent spliced in (psi) -- MISO



http://genes.mit.edu/burgelab/miso/docs/: "currently, MISO does not handle replicates / groups of samples in any special way" → rMATs (Shen et al., PNAS, 2014)

Isoform-level estimation: cufflinks (kallisto, salmon, RSEM), cuffdiff2; many others

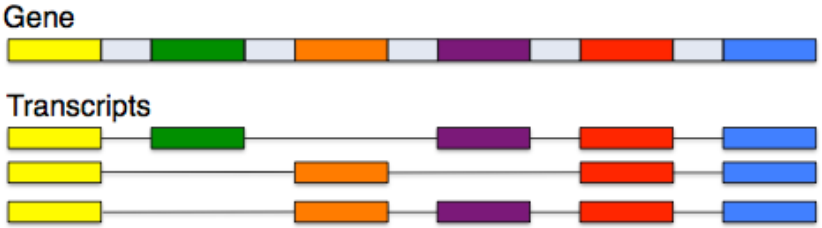
Trapnell et al. 2012, Nature Biotechnology



From estimated isoform abundance from set of (assembled) transcripts, use Jenson-Shannon (JS) divergence to determine change in the mix of transcripts between conditions.

University of Zurich[UZH]
Institute of Molecular Life Sciences

DTU → dirichlet-multinomial distribution



Observed:

- transcript counts $Y = (y_1, y_2, y_3)$
- gene expression $n = \sum_{j=1}^{k} y_j$

Estimated:

- transcript ratios $\Pi = (\pi_1, \pi_2, \pi_3)$

Multinomial: $$P(\mathbf{Y} = \mathbf{y} | \mathbf{\Pi} = \pi) = \binom{n}{\mathbf{y}} \prod_{j=1}^{k} \pi_j^{y_j}$$

Dirichlet: $$P(\mathbf{\Pi} = \pi) = \frac{\Gamma(\gamma_+)}{\prod_{j=1}^{k} \Gamma(\gamma_j)} \prod_{j=1}^{k} \pi_j^{\gamma_j - 1}, \gamma_+ = \sum_{j=1}^{k} \gamma_j$$

Dirichlet-multinomial: $$P(\mathbf{Y} = \mathbf{y}) = \binom{n}{\mathbf{y}} \frac{\Gamma(\gamma_+)}{\Gamma(n + \gamma_+)} \prod_{j=1}^{k} \frac{\Gamma(y_j + \gamma_j)}{\Gamma(\gamma_j)}, \gamma_j = \pi_j \gamma_+$$

Nowicka and Robinson 2016, F1000 Research