Project

From Lecture 1:

Expectations: **project**

- ~10-15 page report, with R code in line (e.g. **knitR** / **Rmarkdown**)
- Describe the biological setting, statistical analysis, exploratory analysis with publication-quality graphics embedded
- Three possibilities:
  - Comparison of statistical methods (simulation / independent reference data + metrics)
  - Reproduce an analysis from a paper from the raw data
  - Real collaborative project with FGCZ or a local laboratory
- Be strategic: work on something related to your interests!
- Typically due at end of first working week of January

Notes:

- can work in groups of 2-3 or individually
- I would like to have a *plan* from you by **01.12.2025**
- Best if you separate JC and project, but possible if good case is made
- Plan: topic and 3-4 bullet points of what you will do; list the GitHub names of the group members

In preparation for “hands-on” session in Lecture 9

- Install docker

https://docs.docker.com/engine/install/

- Download the image I created (from command line):

```
docker pull markrobinsonuzh/sta426hs2025:latest
```

- Get to command line prompt inside container

```
docker run -it markrobinsonuzh/sta426hs2025:latest /bin/bash
```

If that for some reason this does not work:

1. post some details to the #exercises channel.
2. Make an account on renkulab.io

This will not work yet.

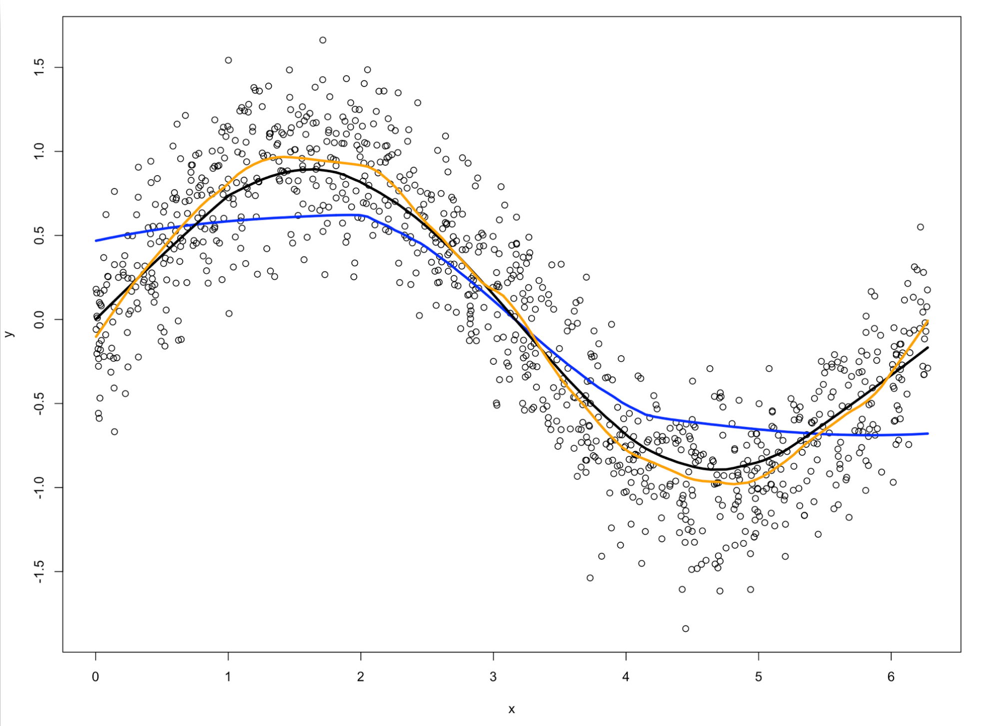
Statistical models for count data analysis (part 2)

- Short aside: lowess/loess
- Reminder of tricks used / material already presented:
  - conditional likelihood (what about normal distribution?)
  - (local) weighted likelihood
  - linear models
- Bringing them together: a more general framework – GLMs
- Beyond differential expression: “differential splicing”, DTE, DTU, DGE, DTE -> G

lowess/loess

- several variations
- typically used to fit smooth curves to data
- non-parametric
- long history in bioinformatics e.g., microarray normalization, mean-variance relationships, limma trend, etc.

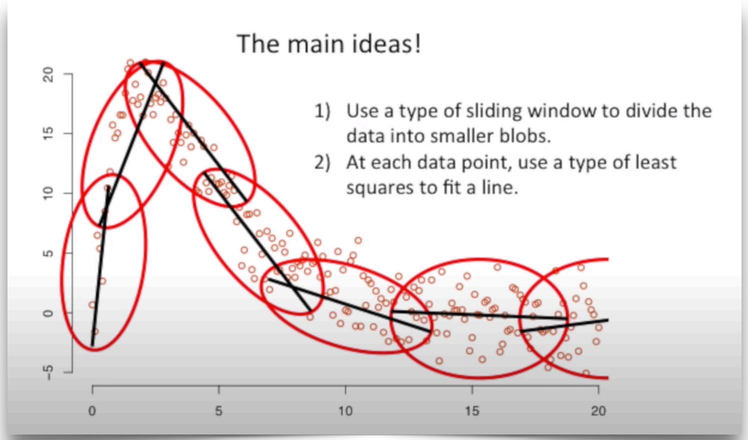
lowess = locally weighted scatterplot smoother



Fit successive local (linear/polynomial) regressions on subsets of the data

Key parameters:

- percentage of data to use (f)
- order of the polynomial fit (usually 2)



The traditional weight function used for LOESS is the tri-cube weight function,

$$w(x) = (1 - |d|^3)^3$$

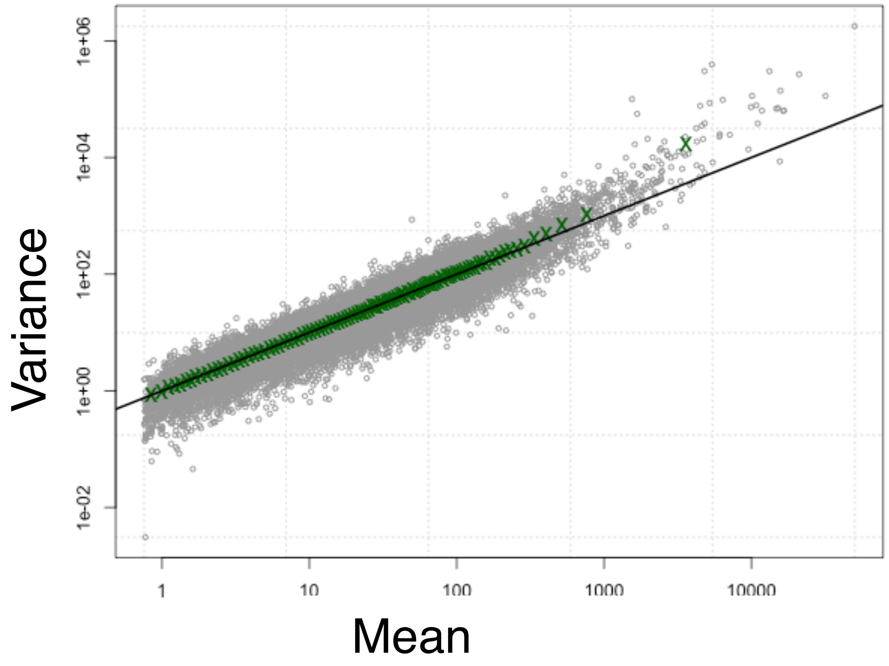
```
x <- runif(1000, 0, 2*pi)
y <- sin(x) + rnorm(1000, sd = .3)

plot(x,y)
lines(lowess(y~x, f=2/3), col="blue", lwd=3)
lines(lowess(y~x, f=1/3), col="black", lwd=3)
lines(lowess(y~x, f=1/10), col="orange", lwd=3)
```

$$Y_i \sim NB(\ \mu_i = N_i * \lambda_i\ ,\ \phi_i\ )$$

$$\text{variance}(\ Y_i\ ) = \mu_i\ (\ 1 + \mu_i\ \varphi_i\ )$$

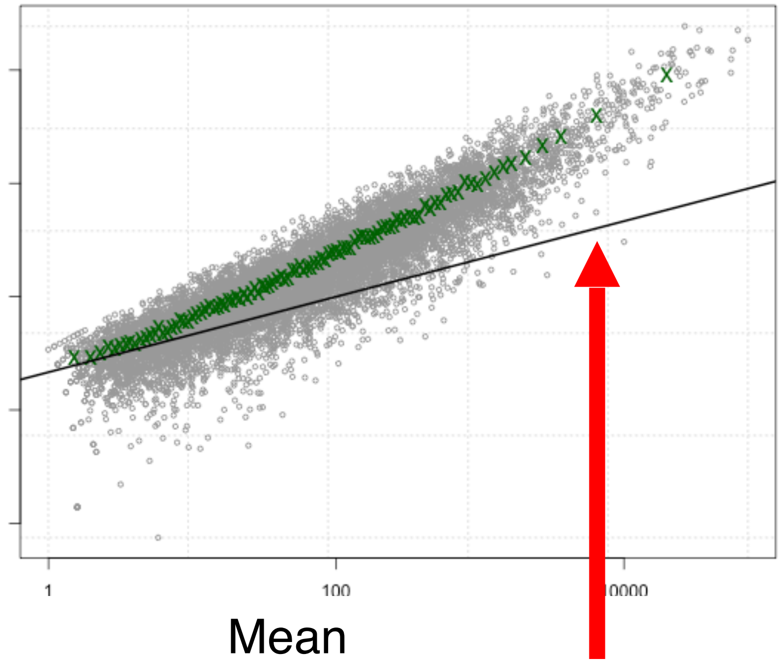
Mean-Variance plots: What we see in real data

Technical replicates



Data from Marioni et al. Genome Research 2008

Biological replicates



mean=variance
(Poisson assumption)

Data from Parikh et al.
Genome Biology 2010

$Y_i \sim NB(\mu_i = Ni * \lambda_i\ , \phi_i\ )$

Conditional likelihood

Likelihood for single **negative binomial** observation:

$$f(y;\mu,\phi) = P(Y=y) = \frac{\Gamma(y+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y+1)}\left(\frac{1}{1+\mu\phi}\right)^{\phi^{-1}}\left(\frac{\mu}{\phi^{-1}+\mu}\right)^{y}$$

If all libraries are the same size (i.e. $m_i \equiv m$), the sum $Z = Y_1 + \cdots + Y_n \sim \mathrm{NB}(nm\lambda, \phi n^{-1})$

Thus, can form conditional likelihood:

$$l_{Y|Z=z}(\phi) = \left[\sum_{i=1}^{n} \log\Gamma(y_i+\phi^{-1})\right] + \log\Gamma(n\phi^{-1}) - \log\Gamma(z+n\phi^{-1}) - n\log\Gamma(\phi^{-1})$$

https://newonlinecourses.science.psu.edu/stat414/node/191/
https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood

Maximum likelihood for normal distribution

likelihood:

$$L(\mu,\sigma^2;x_1,\ldots,x_n) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2\right)$$

log-likelihood:

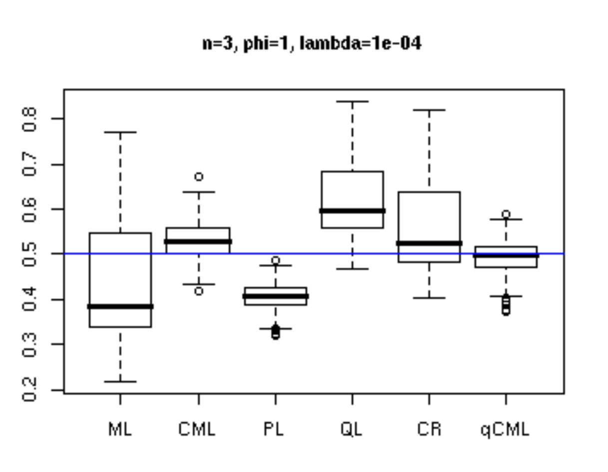
$$l(\mu,\sigma^2;x_1,\ldots,x_n) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2$$

maximize likelihood w.r.t. parameters mu, sigma

$$\hat{\mu}_n = \frac{1}{n}\sum_{j=1}^{n}x_j$$

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j-\hat{\mu})^2$$

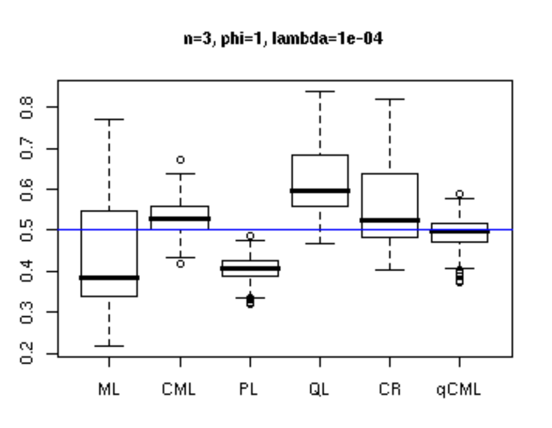
From Lecture 7 (estimation of NB dispersion):



Main point: MLE is good for many models and generally works in larger samples (asymptotically unbiased), but often **biased** in small samples —> other approximations https://www.stats.net.au/Maths_REML_manual.pdf

Maximum conditional likelihood for normal distribution: look at Y_1, .., Y_n | Y-bar

From Lecture 7 (estimation of NB dispersion):



For y_1,.., y_n:

$$logLikelihood = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{2\sigma^2} - \frac{n(\bar{y}-\mu)^2}{2\sigma^2}$$

y-bar:

$$logLikelihood \text{ for } \bar{y} = \frac{1}{2}ln(n) - \frac{1}{2}ln(2\pi) - \frac{1}{2}ln(\sigma^2) - \frac{n(\bar{y}-\mu)^2}{2\sigma^2}$$

maximize likelihood w.r.t. parameters **sigma**

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Moderated dispersion estimate

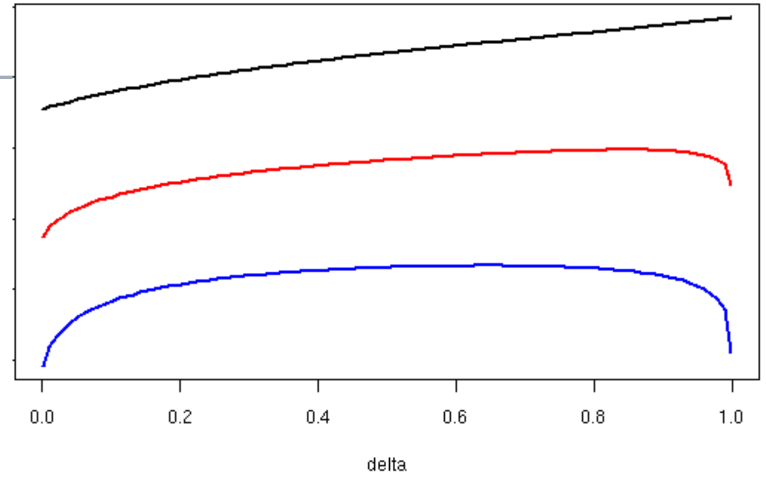
Weighted likelihood -- individual log-likelihood plus a weighted version of the common log-likelihood:

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

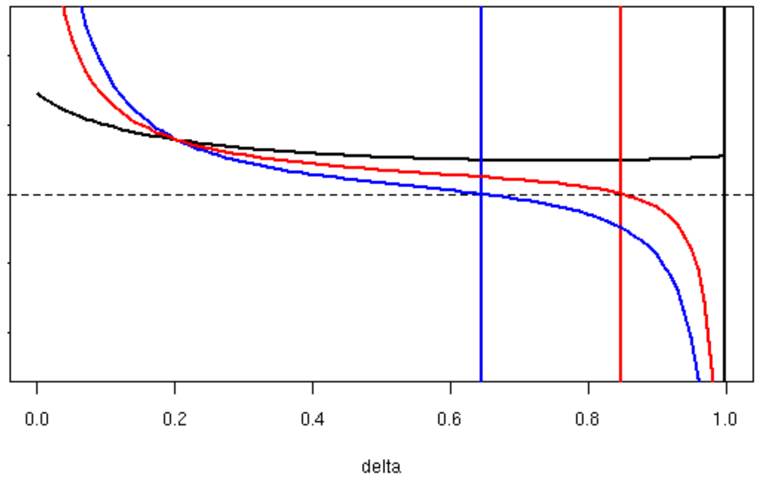
l_g - quantile-adjusted conditional likelihood

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

Log-Likelihood



Score (1st derivative of LL)



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

University of Zurich[UZH]

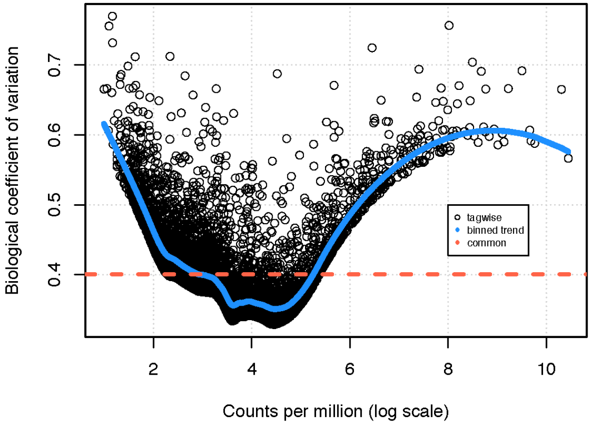
Institute of Molecular Life Sciences

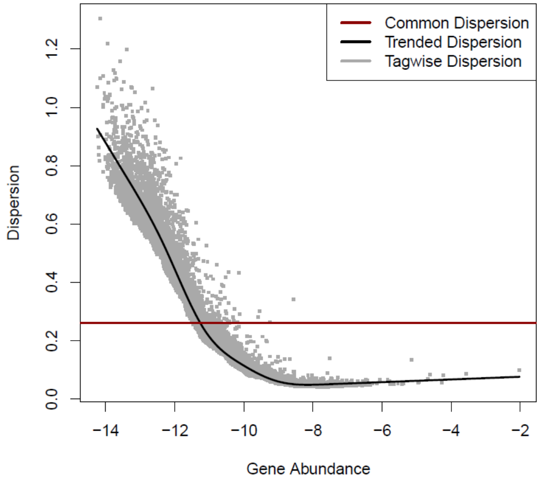
Dispersion varies with mean: moderate (e.g., weighted likelihood) dispersion towards trend

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

(1-α)

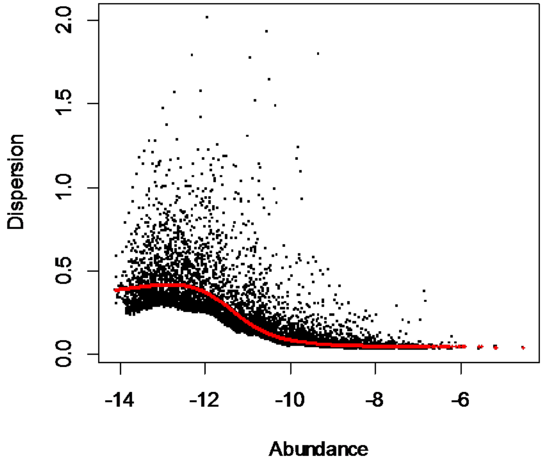
Data:
Tuch et al., 2008





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:

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

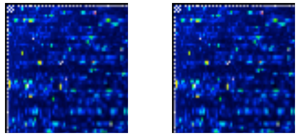
vector of observed data

design matrix

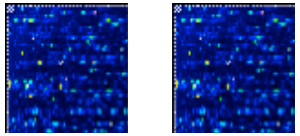
Vector of parameters to estimate

Analysis of Variance → Linear model

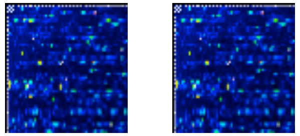
WT x 2



Cond A x 2



Cond B x 2



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

α_1 = wt log-expression

α_2 = Cond A - wt

α_3 = Cond B - wt

$E[y_1]=E[y_2]=\alpha_1$ $E[y_3]=E[y_4]=\alpha_1+\alpha_2$ $E[y_5]=E[y_6]=\alpha_1+\alpha_3$

Applications: paired designs, multi-factor designs, interactions

—> This particular model only valid for continuous response

Generalized linear models: a more general framework

Gaussian (normal) distributed response —> various other (common) types.

Three components:

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

Link function and linear predictor

Link function

$$E(Y_i) = \mu_i \qquad\qquad g(\mu_i) = \eta_i$$

$$\eta_i = \beta_0 + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

Linear predictor (covariates)

$$\mathrm{var}(Y_i) = \phi V(\mu)$$

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

Data is not transformed.

Variance is a function of mean.

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	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = \mu^{-1}$	$\mu = (\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
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: $[0, K)$	outcome of single K-way occurrence			
	K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1				
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

`edgeR::glmFit()`

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

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

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

Several choices here:

- Maximum Likelihood (MLE)
- Pseudo-Likelihood (PL)
- Quasi-Likelihood (QL)
- Conditional Maximum Likelihood (CML)
- ~~Approximate Conditional Inference (Cox-Reid~~)
- *quantile-adjusted Maximum Likelihood (qCML)*

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

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

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

“Cox Reid adjusted profile likelihood” —> Estimation of dispersion parameter

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

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

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

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.

$$Y_i \sim \mathrm{NB}(\mu_i, \phi)$$
$$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$$

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.

In this approach, ϕ_g is estimated by maximizing

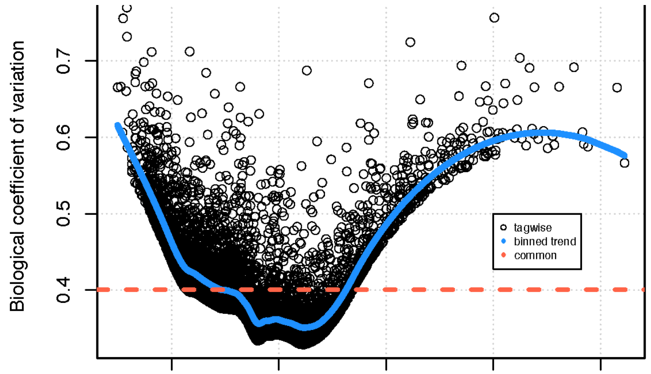
$$\mathrm{APL}_g(\phi_g) + G_0\,\mathrm{APL}_{Sg}(\phi_g),$$

where G_0 is the weight given to the shared likelihood and $\mathrm{APL}_{Sg}(\phi_g)$ is the local shared log-likelihood.

APL is simply another likelihood, so weighted likelihood still works

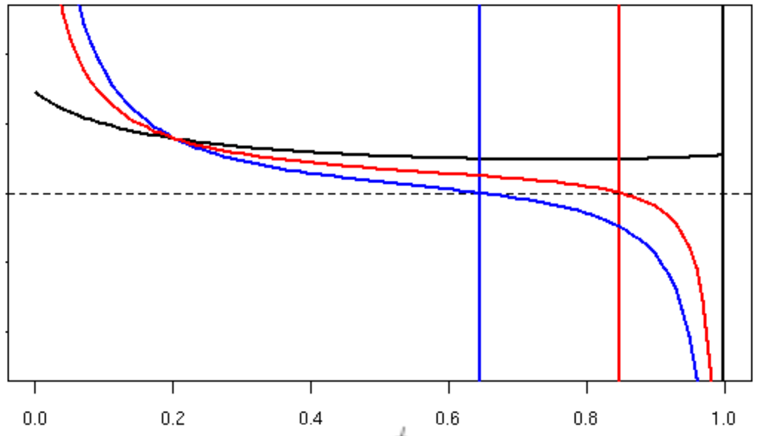
WL is the individual log-likelihood plus a weighted version of the **common log-**likelihood:

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



L_g - adjusted profile likelihood (or trended version)

Score (1st derivative of LL)



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

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

Exponential family

“natural parameter”

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

Distribution	Natural parameter	c	d
Poisson	$\log \theta$	$-\theta$	$-\log y!$
Normal	$\dfrac{\mu}{\sigma^2}$	$-\dfrac{\mu^2}{2\sigma^2} - \dfrac{1}{2}\log\left(2\pi\sigma^2\right)$	$-\dfrac{y^2}{2\sigma^2}$
Binomial	$\log\left(\dfrac{\pi}{1-\pi}\right)$	$n\log\left(1-\pi\right)$	$\log\binom{n}{y}$

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

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

— from Introduction to Generalized Linear Models, Annette Dobson, 2nd edition.

Given dispersion estimates (Cox-Reid APL): estimation, statistical testing of regression parameters

Generalized linear model comes with many advantages:

1. Estimation is the same for all response types (so-called Fisher scoring, which effectively turns likelihood maximization into an iteratively re-weighted estimation problem)
2. Asymptotic theory that lead to i) Wald; ii) Score; or, iii) likelihood ratio tests for parameters of interest (more details). All of these are based on asymptotics ("large" sample approximations) – how to choose one that works well in practice?

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

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

with mean $\boldsymbol{\beta}$ and variance-covariance matrix $(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\phi$.

Tests for subsets of $\boldsymbol{\beta}$ are based on the corresponding marginal normal distributions.

(Wald test used in DESeq2 package)

http://data.princeton.edu/wws509/

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)

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

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

General form (exponential family)

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

`edgeR::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).

Interesting generalizations of NB modeling for RNA-seq (1)

$$LRT_k = 2(\ell_k(\hat{\mu}_k|y_k) - \ell_k(\tilde{\mu}_k|y_k)) \longrightarrow LRT_k \sim \Phi_k \chi_q^2 + O_p(n^{-1/2})$$

$$\hat{\Phi}_k = \frac{2(\ell_k(y_k|y_k) - \ell_k(\hat{\mu}_k|y_k))}{n-p}$$

$$F_{QL} = \frac{LRT_k/q}{\hat{\Phi}_k}$$

Accounting for the uncertainty in estimating dispersion

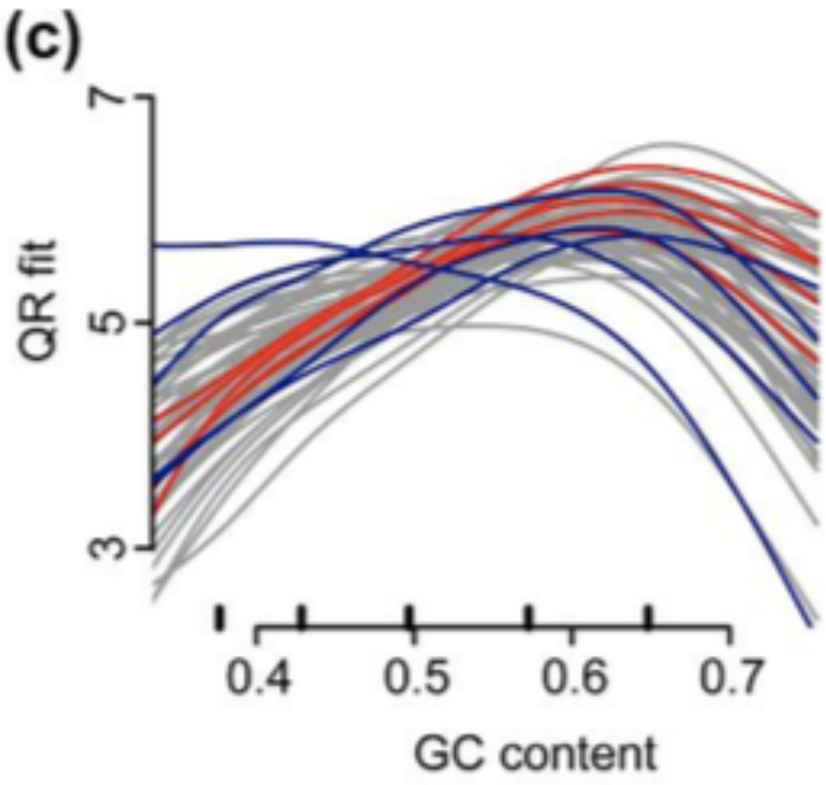
`edgeR::glmQLF`

Lund et al., SAGMB 2012; 11(5):8

Offset ξ explicitly in GLM:

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

Interesting generalizations of NB modeling for RNA-seq (2)



Integrate sample-specific normalization via offset

Profiles vary from sample to sample:
GC content
Gene length

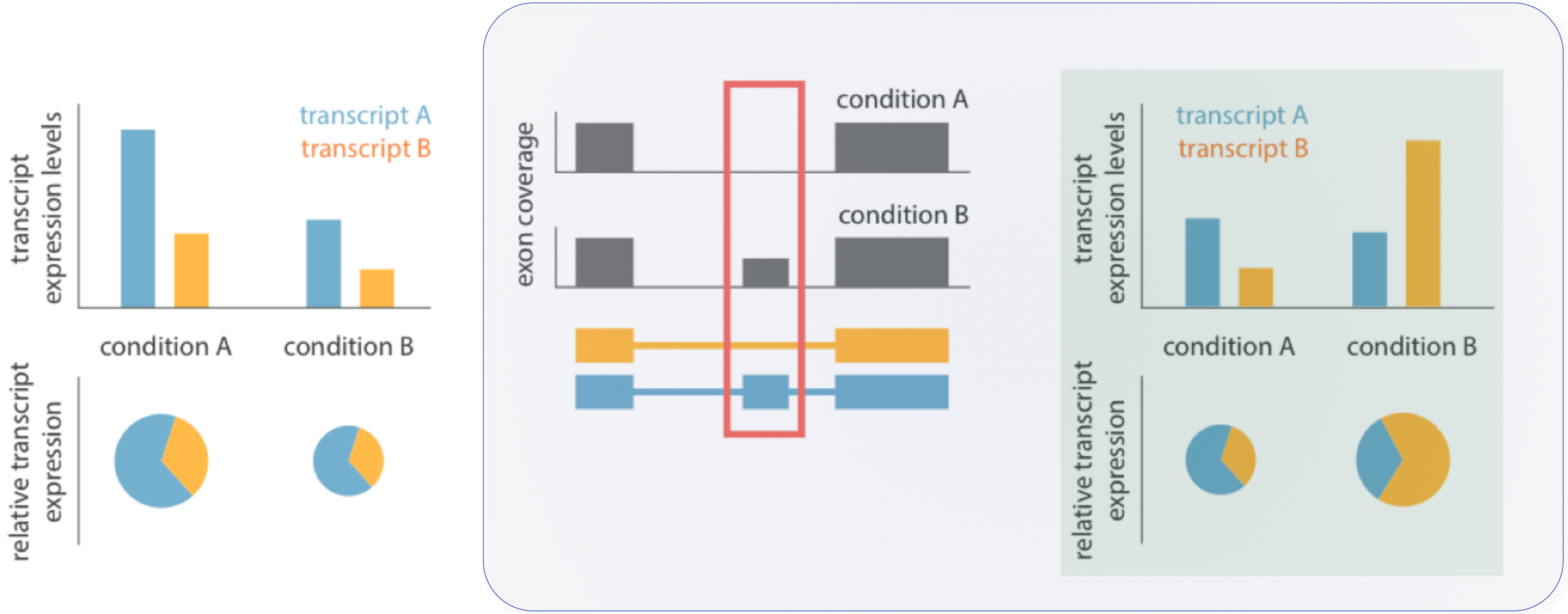
DOES NOT change data, use offsets to modify expected mean

Give a sample (or gene)-specific offset to edgeR/DESeq2

Hansen, Irizarry, Wu Biostatistics 2012

Some terms: DTE, DEU, DTU

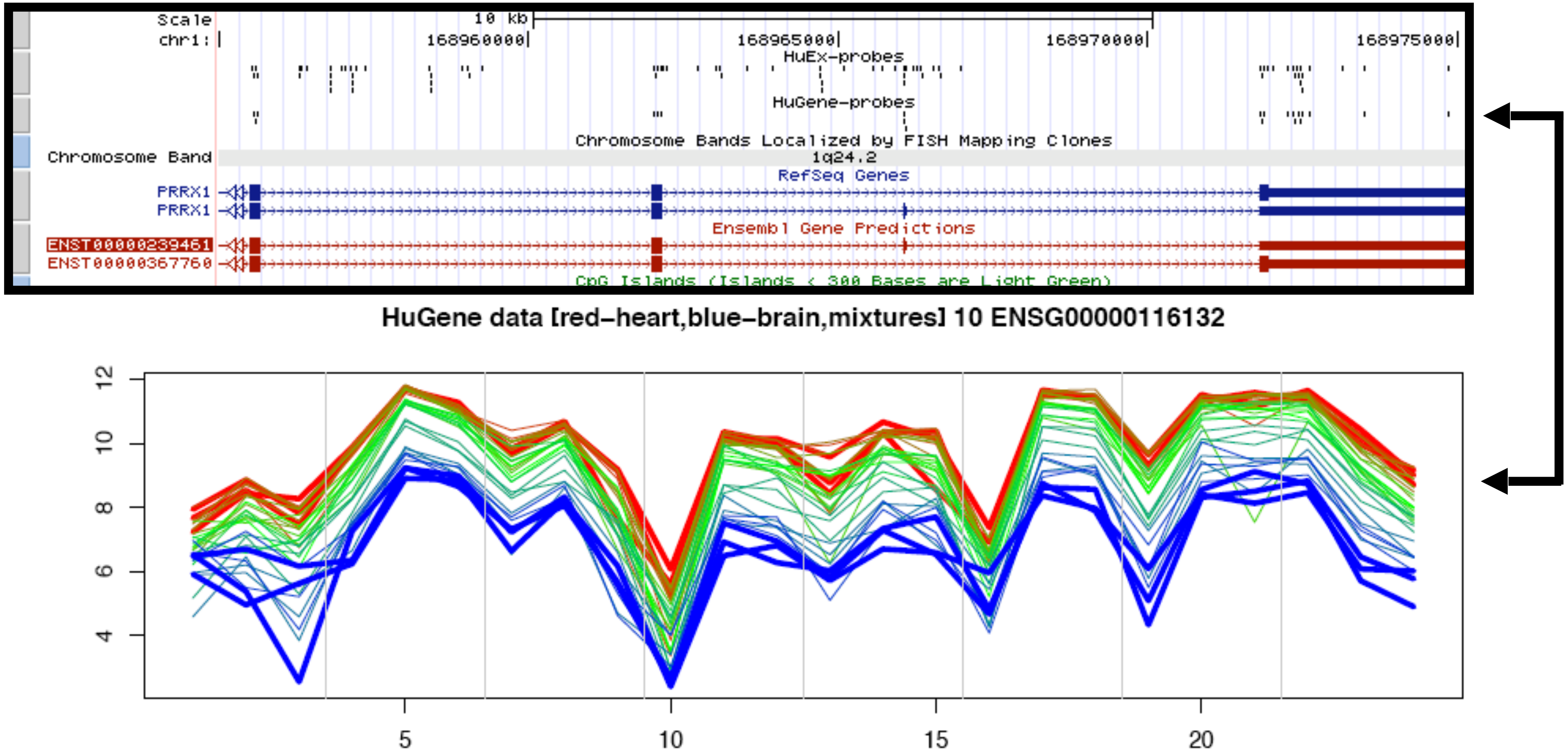
Differential transcript expression (DTE) | **Differential exon usage (DEU)** | **Differential transcript usage (DTU)**



differential splicing

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

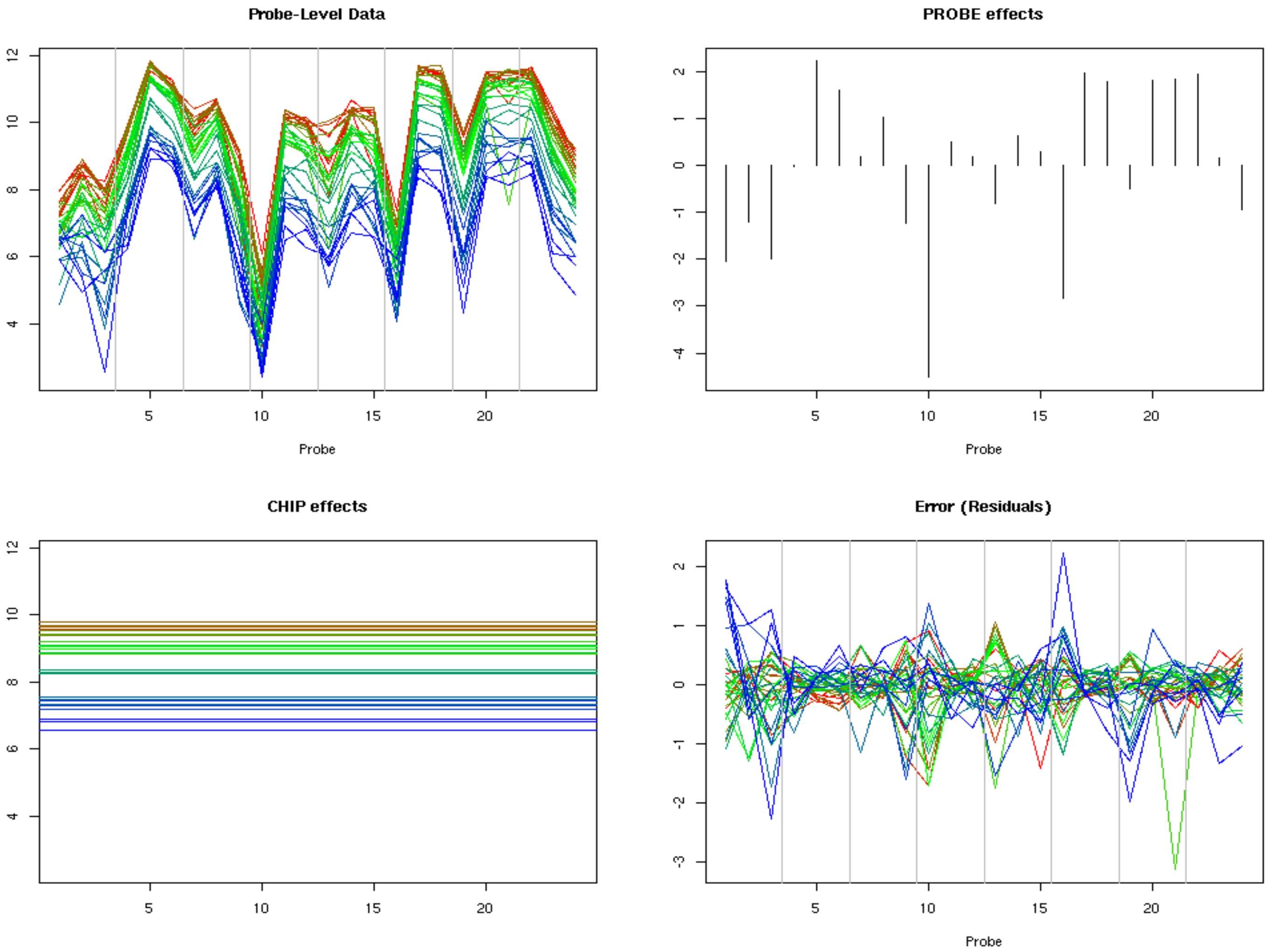
Digression 1/3: The nature of Affymetrix Probe Level Data



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

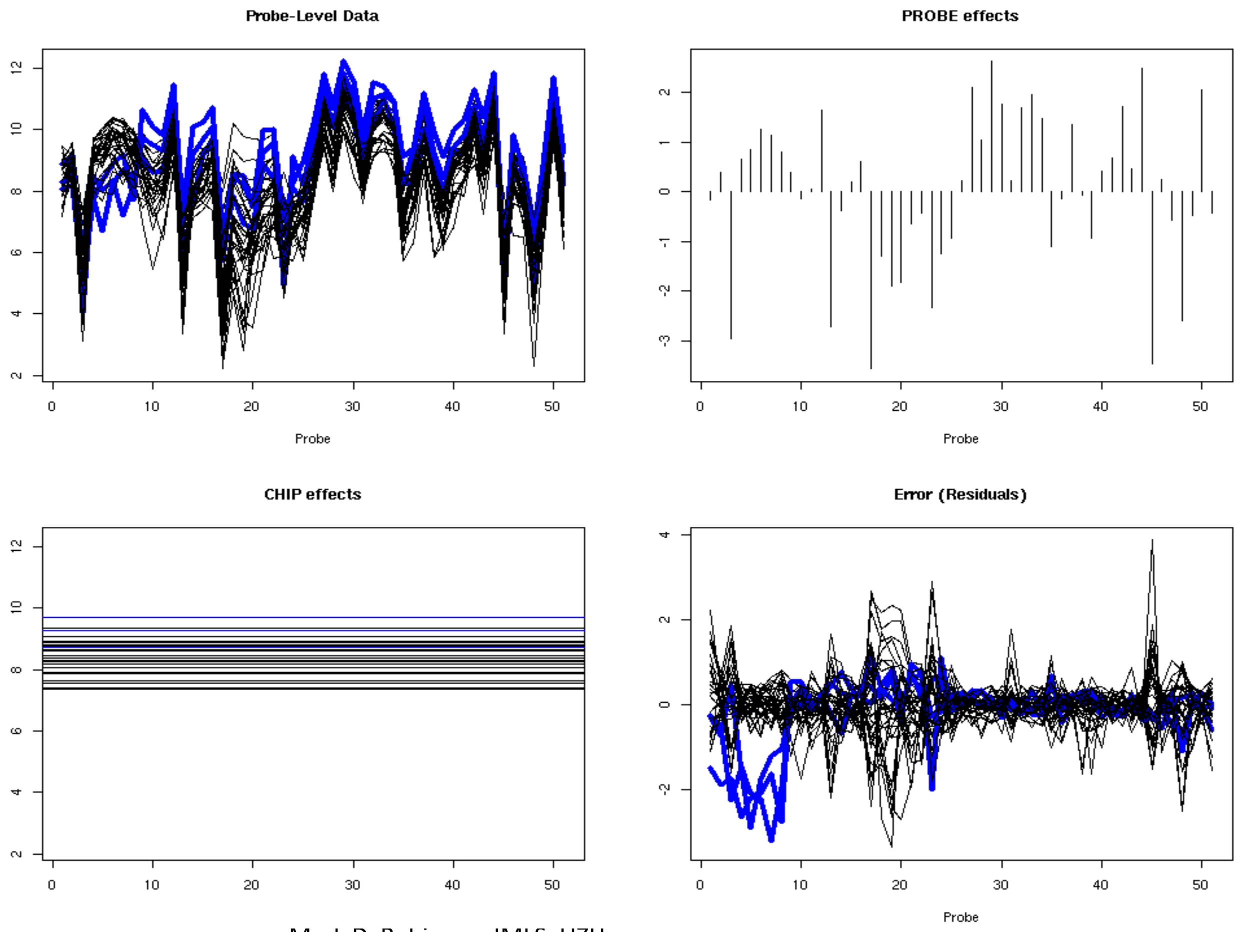
Tissue mixture dataset

(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



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

University of Zurich[UZH]

Statistical Bioinformatics // Institute of Molecular Life Sciences

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

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

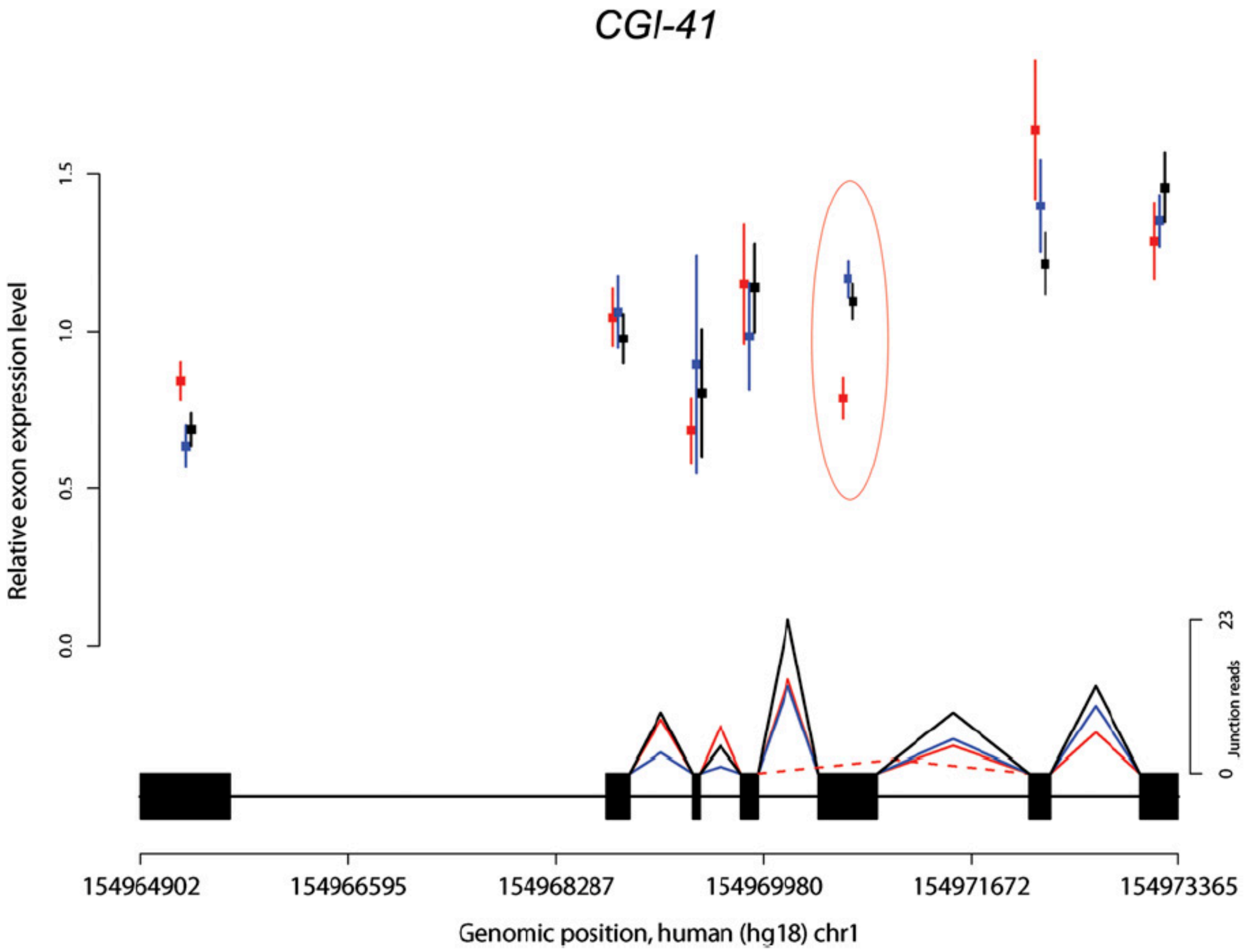
Hugues Richard[1,*], Marcel H. Schulz[1,2], Marc Sultan[3], Asja Nürnberger[3], Sabine Schrinner[3], Daniela Balzereit[3], Emilie Dagand[3], Axel Rasche[3], Hans Lehrach[3], Martin Vingron[1], Stefan A. Haas[1] and Marie-Laure Yaspo[3]

[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?

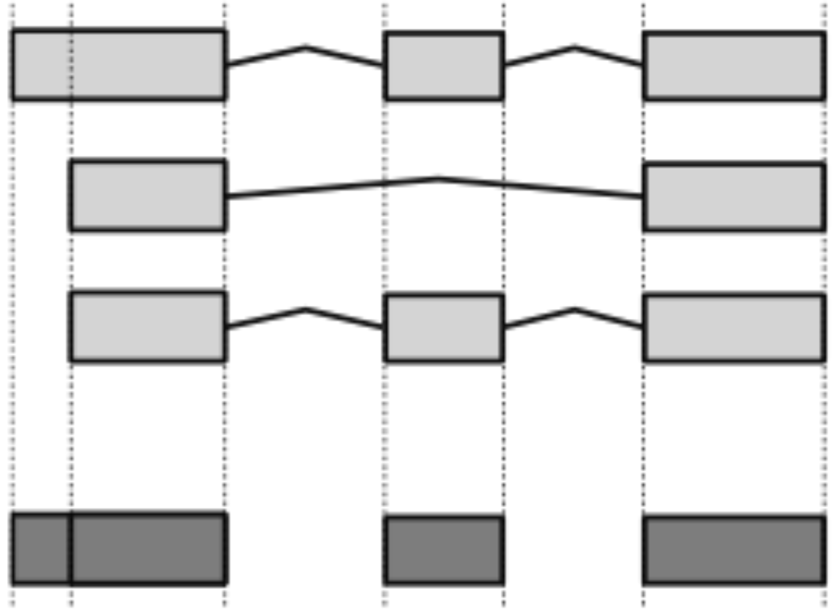


Figure 1. Flattening of gene models: This (fictional) gene has three annotated transcripts involving three exons (light shading), one of which has alternative boundaries. We form counting bins (dark shaded boxes) from the exons as depicted; the exon of variable length gets split into two bins.

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: exon-level models

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:

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

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

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

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)

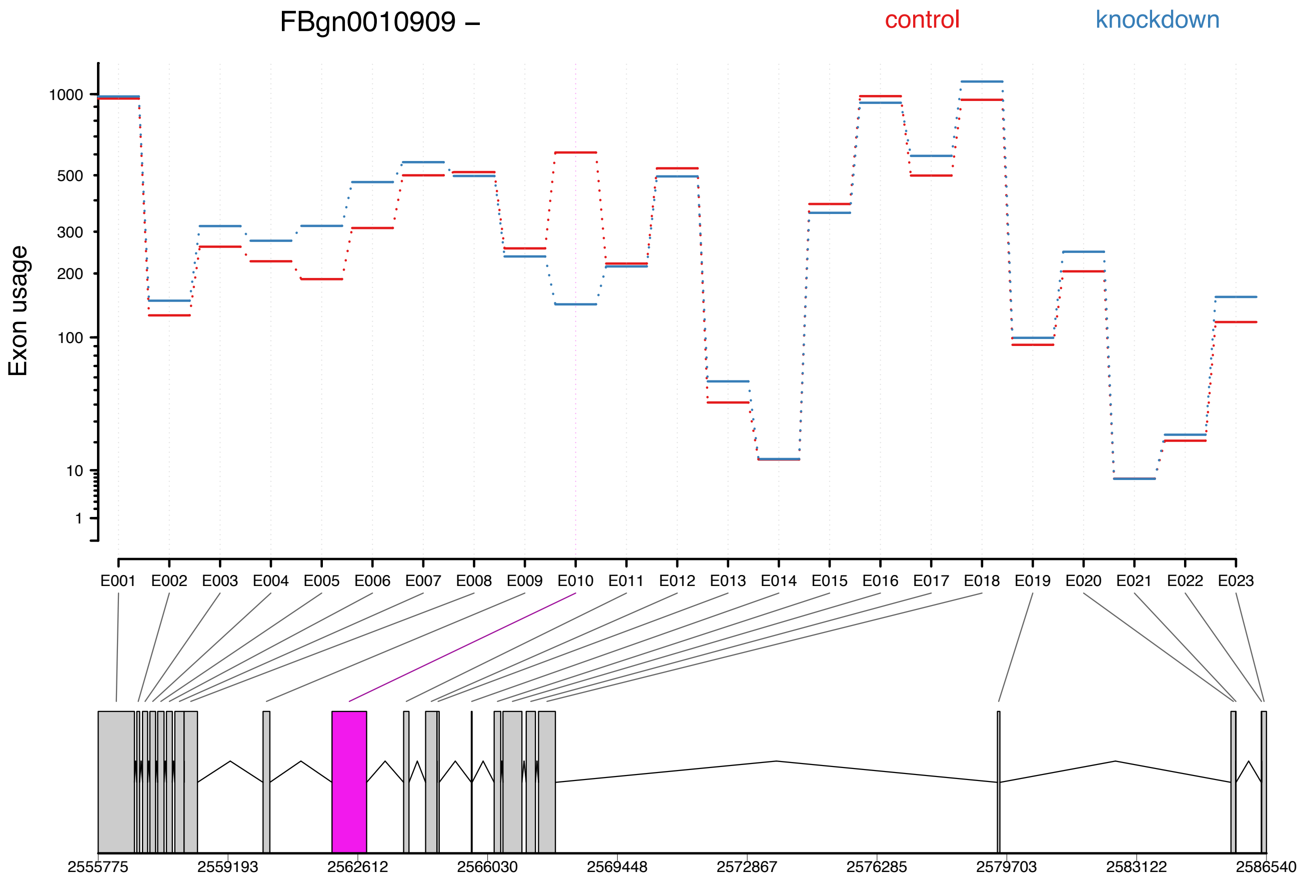
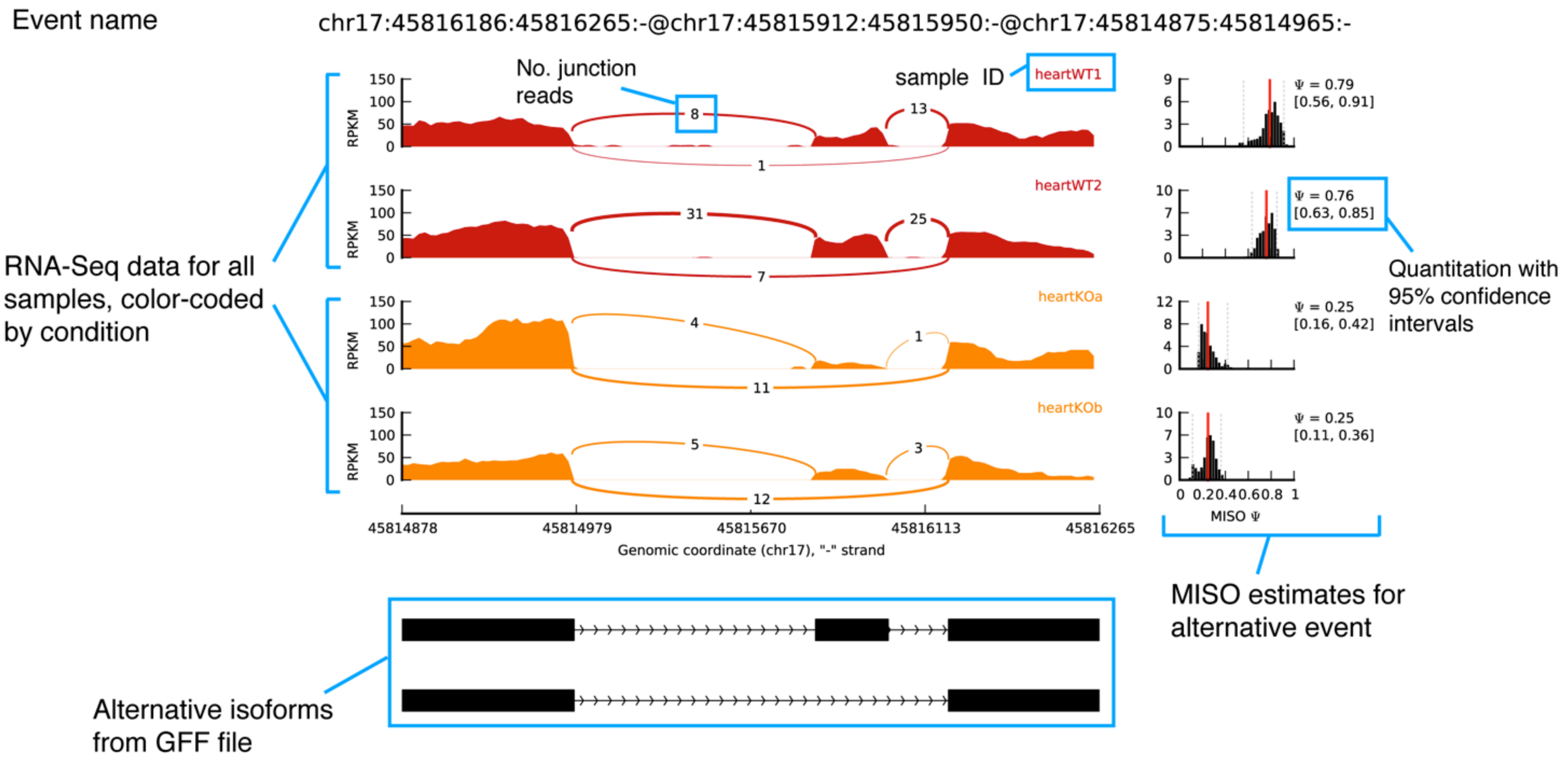


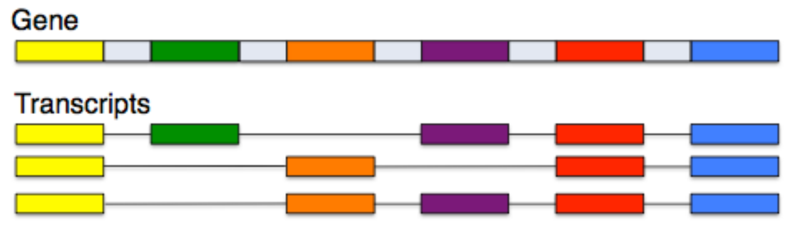
Figure 6: **Fitted splicing**
The plot represents the estimated effects, as in Figure 3, but after subtraction of overall changes in gene expression.

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)



DTU —> dirichlet-multinomial distribution

Estimated:

- transcript ratios

$$\Pi = (\pi_1, \pi_2, \pi_3)$$

Observed:

- transcript counts

$$Y = (y_1, y_2, y_3)$$

- gene expression

$$n = \sum_{j=1}^{k} y_j$$

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:

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

DRIMSeq

Nowicka and Robinson 2016, F1000 Research

Many more details here (and in the 200 papers)

RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis

Annual Review of Biomedical Data Science

Vol. 2:139-173 (Volume publication date July 2019)
First published as a Review in Advance on April 30, 2019
https://doi.org/10.1146/annurev-biodatasci-072018-021255

Koen Van den Berge,[1,*] Katharina M. Hembach,[2,*] Charlotte Soneson,[2,3,*] Simone Tiberi,[2,*] Lieven Clement,[1,†] Michael I. Love,[4,†] Rob Patro,[5,†] and Mark D. Robinson[2,†]

[1]Bioinformatics Institute Ghent and Department of Applied Mathematics, Computer Science and Statistics, Ghent University, 9000 Ghent, Belgium
[2]Institute of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland; email: mark.robinson@imls.uzh.ch
[3]Current Affiliation: Friedrich Miescher Institute for Biomedical Research and SIB Swiss Institute of Bioinformatics, 4058 Basel, Switzerland
[4]Department of Biostatistics and Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27514, USA
[5]Department of Computer Science, Stony Brook University, Stony Brook, New York 11794, USA

Extra slides

Some interesting generalizations of NB modeling for RNA-seq (1)

If Y_{ijk} has a Poisson distribution, then $\mathrm{Var}(Y_{ijk}) = \mu_{ik}$.
If Y_{ijk} has an NB2 distribution, then $\mathrm{Var}(Y_{ijk}) = \mu_{ik}(1 + \phi\mu_{ik})$.

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

(generalization of the model:
mean-variance relationship)

Di et al., SAGMB 2011 10(1): 24

Some interesting generalizations of NB modeling for RNA-seq (2)

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

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

Higher order asymptotics

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

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

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.

Di et al., SAGMB 2013; 12(1): 49–70