Analysis of multi-factor RNA / ChIP-Seq experiments with respect to biological variation

Gordon Smyth

Bioconductor
28 July 2011

Walter+Eliza Hall
Institute of Medical Research

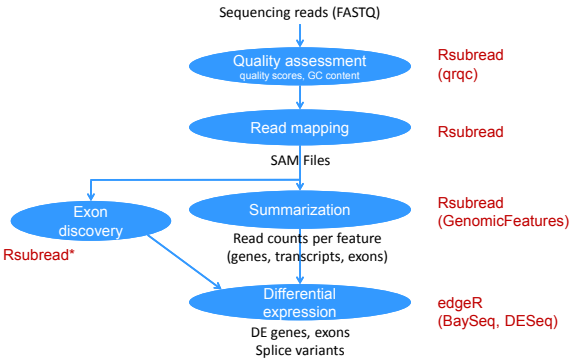
1

Packages

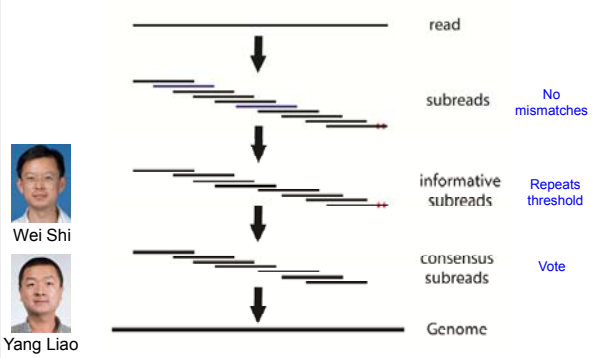
- Rsubread
  - Read alignment
  - Summarization by genomic features
  - Exon discovery
- limma
  - Normal-based DE analysis
  - Gene set analysis
- edgeR
  - Negative binomial-based DE analysis
  - Detection of splice-variants*
- goseq
  - Gene ontology analysis adjusted for gene length

2

Reads to genes: a Bioconductor pipeline



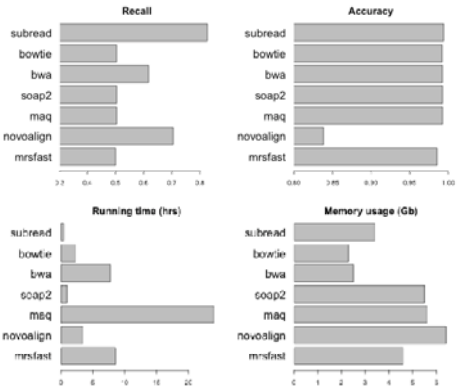
Rsubread maps reads to genome: multi-seed and vote



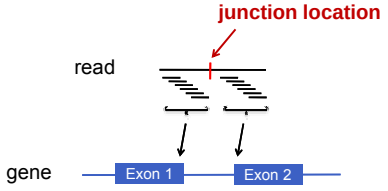
Wei Shi

Yang Liao

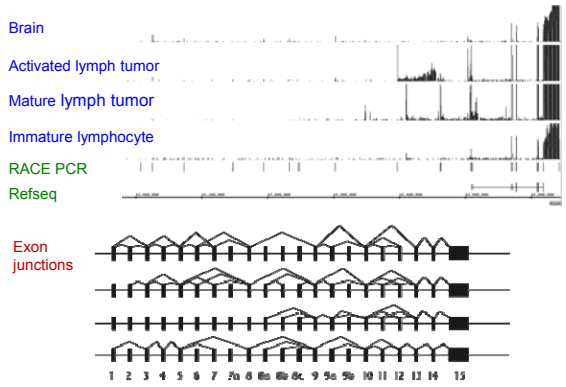
Simulation



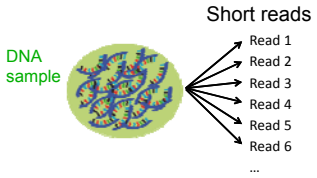
Discovery of exons and exon-junctions



Exon-junction discovery



Library of read counts



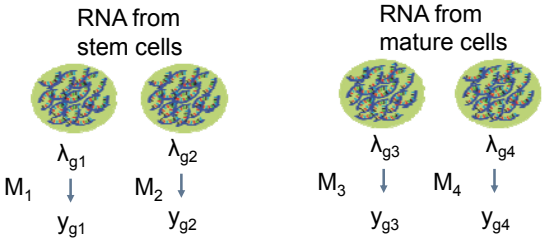
Count by exon, or gene, or …

M = total number of reads
λ_g = true proportion of gene g
y_g = number of reads for gene g

≈ Poisson with $E(y_g) = M \lambda_g$

8

A small RNA-seq experiment



Reads $M_i \approx 20$ million
Genes g = 1, …, 30k

$E(y_{gi}) = \mu_{gi} = M_i \lambda_{gi}$

9

log-linear models

$$\log \mu_{gi} = \log \lambda_{ig} + \log M_{gi}$$
$$= x_i^T \beta_g + \log M_{gi}$$

row of design matrix

vector of log fold changes

(normalized) library size

(unknown)

10

Normalization

- Scale normalization
  - "Effective" library size

Robinson and Oshlack, Genome Biol 2010

- Nonlinear normalization
  - Quantile normalization
  - Gene length
  - GC content (of reads, of fragments)

cqn
EDASeq

11

Counts show a quadratic mean-variance relationship

$$var(y_{gi}) = \mu_{gi} + \varphi_g \mu_{gi}^2$$

Poisson variation

CV^2 of the "true" expression levels λ_{gi} across replicates

CV = coefficient of variation = sd/mean

12

Biological coefficient of variation

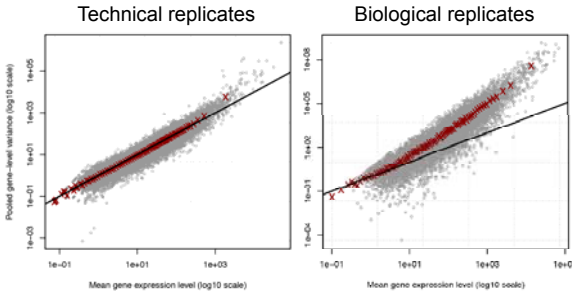
Total CV^2 = Technical CV^2 + Biological CV^2

From sequencing technology	CV of "true" expression level
↓ zero for large counts	≈ constant

$BCV = \sqrt{\varphi_g}$

13

Real data show quadratic variances



14

Statistical properties of read counts

- Properties
  - Integer values (discrete)
  - Mean-variance relationship
  - Distinguish technical from biological variation
- Approaches
  - log-counts as normal (limma)
  - counts as negative binomial (edgeR)

15

Limma approach

log-counts:

$$z_{gi} = \log_2\left(\frac{\text{count}_{gi}+0.5}{\text{libsize}_{gi}+0.5}\right) = \log_2\left(\frac{y_{gi}+0.5}{M_{gi}+0.5}\right)$$

normalize libsize in advance or normalize z_{gi} as for microarrays.

Linear modelling:

$$E(z_{gi}) = \mu_{gi} = x_i^T \beta_g$$

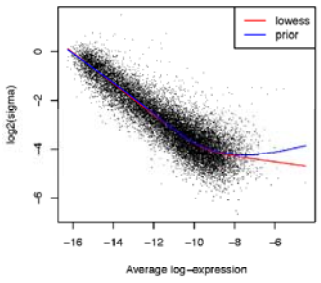
$$\text{var}(z_{gi}) = s(\mu_{gi})\,\sigma_g^2$$

Smooth function of mean


Charity Law

16

Empirical Bayes with abundance-dependent prior



eBayes(fit, trend=TRUE); plotSA(fit)

17

Negative binomial approach

If λ_{gi} are gamma distributed, then

$$y_{gi} \sim \text{NegBin}(\mu_{gi}, \varphi_g)$$

Once the dispersions are estimated, the log-linear models are generalized linear models


Mark Robinson


Davis McCarthy


Yunshun Chen

18

Ensuring glm convergence

- Iterative fitting of glms is computationally demanding, and standard glm code can diverge
- Pseudo Newton-Raphson strategy to reduce need for matrix decompositions
- Line searches to prevent divergence
- Highly vectorized code
- Fit genewise glms in a few seconds

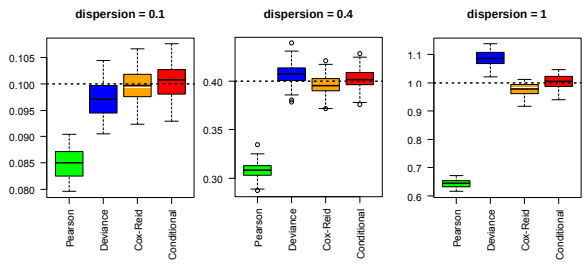
19

Conditional inference for the dispersions

- Need to adjust for estimation of β_g when forming likelihood for φ_g
- For two-group comparison, can compute conditional distributions given row totals and conduct exact inference
- For more general designs, use Cox-Reid adjusted profile likelihood to condition on estimator of β_g

20

Performance of conditional estimators of dispersion



21

Complexity of dispersion: sharing information between genes

- Separate gene-wise estimation of φ_g is impractical
- Common dispersion (Robinson & Smyth 2008)
- Trended dispersion (Anders & Huber 2010)
- Gene-wise by empirical Bayes shrinkage (Robinson & Smyth, 2007)

22

Common dispersion likelihood

Assume same dispersion for all genes
$\varphi_g = \varphi$

Genewise conditional log-likelihood
$\ell_g(\varphi; y_g)$

Common-dispersion log-likelihood
$\ell_c(\varphi) = (1/G) \sum_g \ell_g(\varphi; y_g)$

Maximized at φ_c

23

Empirical Bayes shrinkage for the dispersion

Estimate φ_g by empirical posterior mode:

Posterior = $\ell_g(\varphi_g) + \alpha\, \ell_c(\varphi_c)$

Genewise likelihood

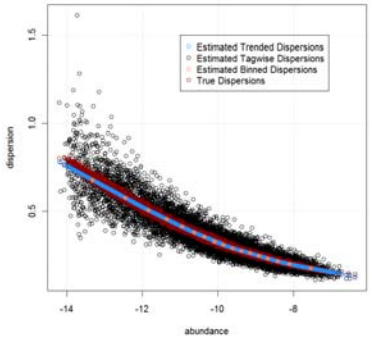
Precision of prior

Empirical prior distribution

Local weighting produces abundance dependent prior

24

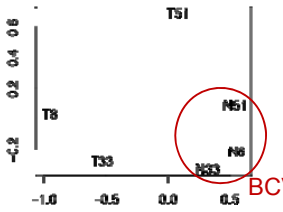
Estimated dispersions (simulation)



Oral squamous cancer

	Normal	Tumour
Patient 8	N8	T8
Patient 33	N33	T33
Patient 51	N51	T51

Tuch et al, *PloS ONE* 2010

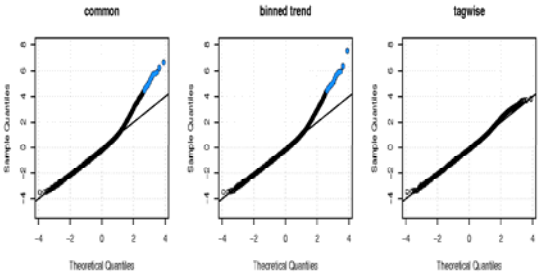


Multidimensional scaling plot using BCV distances

BCV=40%

26

Genewise goodness of fit tests



Tagwise dispersion gives the best fit

Differential expression

Fit models of increasing complexity:

Patient

LRTs — 1271 generally DE genes

Patient + Tissue Source

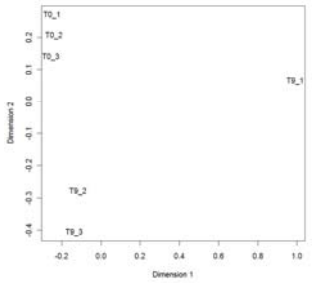
LRTs — 184 genes specific to individual tumours

*Patient * Tissue Source*

FDR < 0.05

28

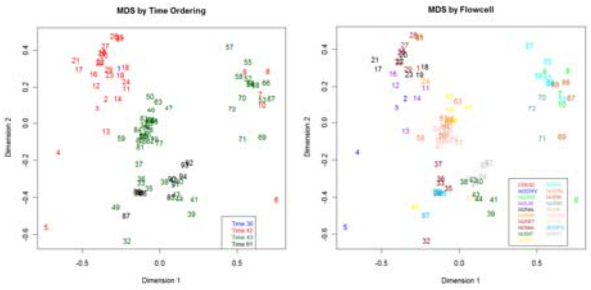
Multidimensional scaling plots with BCV as distance



Outlier

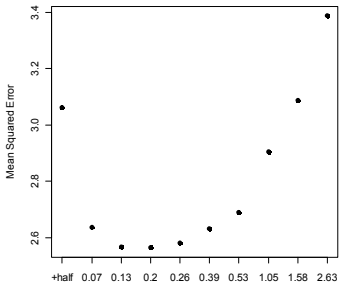
29

Finding technical effects



Empirical Bayes for the fold changes

Predicting PCR fold changes from SEQC RNA-Seq data



Accuracy

Shrinkage

Belinda Phipson

31

Screening for splice-variants

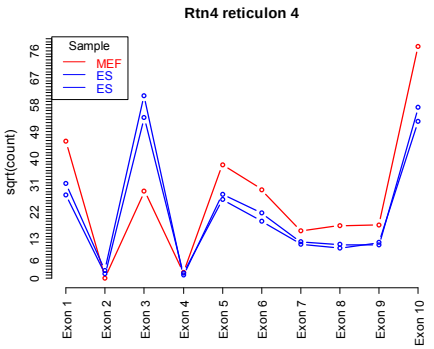
- Exon level summaries
- Estimate exon-wise dispersions
- Test exon x group interaction for each gene

Compare to:

Richard et al, NAR 2010
DEXSeq package

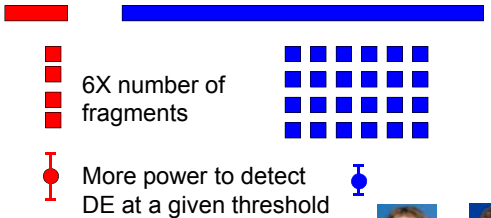
Davis McCarthy

Screening for splice-variants



GOseq

Genes vary in length ...



6X number of fragments

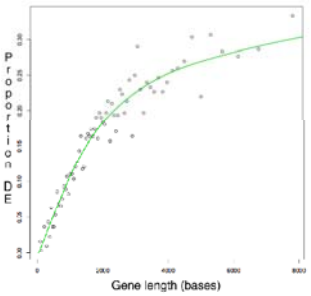
More power to detect DE at a given threshold

Alicia Oshlack

Matt Young

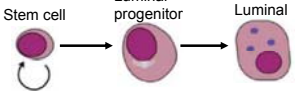
Correcting bias with GOseq

1) List of DE genes.
2) Quantify chance of being DE as a function of length.
3) Use genewise probabilities to compute enrichment probabilities for each GO category

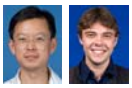


ChIP-Seq for epigenetic modifications

Mammary stem cells

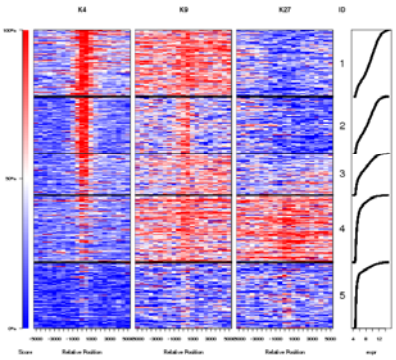


Surveying multiple epigenetic marks at each point



36

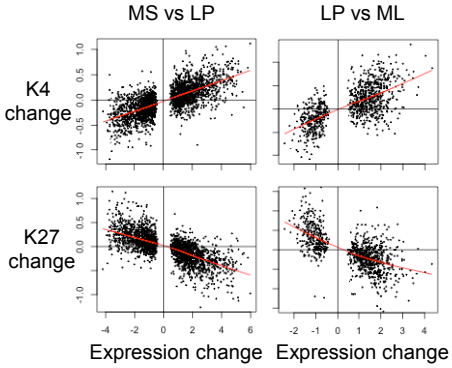
Epigenetic landscape (MS)



Repitools

37

Expression vs epigenetic changes



Conclusions

- Self-contained pipeline for RNA-Seq close at hand
- Methods of differential expression analysis of RNA-seq (etc) data based on mean-variance modelling of counts and conditional inference
- Shared-parameter likelihood priors provide a generally applicable paradigm for parameter shrinkage

39

Lab Members

- Matthew Ritchie
- Wei Shi
- Belinda Phipson
- Charity Law
- Yunshun Chen
- Joshy George
- Keith Satterley
- Yifang Hu
- Davis McCarthy
- Cynthia Liu
- Yang Liao
- Jenny Dai

Recent past:

- Alicia Oshlack
- Di Wu
- Matthew Young
- Luke Zappia
- Carolyn de Graaf
- Mark Robinson

40

Collaborators

- Lynn Corcoran
- Tim Thomas
- Anne Voss
- Bilal Sheikh
- Samir Taoudi
- Peter ‘t Hoen
- Jane Visvader
- Geoff Lindeman
- Bhupinder Pal
- Marie-Liesse Asselin-Labat

41