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

Gordon Smyth
Bioconductor
28 July 2011

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

Reads to genes: a Bioconductor pipeline

Rsubread maps reads to genome: multi-seed and vote

Simulation

Discovery of exons and exon-junctons
Analysis of multifactor RNA/ChIP-Seq experiments with respect to biological variation

Gordon Smyth
Bioconductor, Seattle 28 July 2011

Exon-junction discovery

Library of read counts

A small RNA-seq experiment

log-linear models

Normalization

Counts show a quadratic mean-variance relationship

- Scale normalization
  - "Effective" library size
  - Robinson and Oshlack, Genome Biol 2010

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

Counts show a quadratic mean-variance relationship

\[ \text{var}(y_{gi}) = \mu_{gi} + \varphi_{gi} \mu_{gi}^2 \]

Poisson variation

CV of the "true" expression levels \( \lambda_{gi} \) across replicates

\[ CV = \text{coefficient of variation} = \frac{sd}{mean} \]
Biological coefficient of variation

Total $CV^2 = \text{Technical } CV^2 + \text{Biological } CV^2$

From sequencing technology

CV of "true" expression level

↓ zero for large counts

≈ constant

$BCV = \sqrt{\phi_g}$

Real data show quadratic variances

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)

Limma approach

log-counts:

$z_g = \log\left(\frac{\text{count}_g + 0.5}{\text{libsize}_g + 0.5}\right) = \log\left(\frac{y_g + 0.5}{M_g + 0.5}\right)$

normalize libsize in advance or normalize $z_g$ as for microarrays.

Linear modelling:

$E(z_g) = \mu_g = x_g^T \beta_g$

$\text{var}(z_g) = s(\mu_g) \sigma_g^2$

Smooth function of mean

Negative binomial approach

If $\lambda_g$ are gamma distributed, then

$y_g \sim \text{NegBin}(\mu_g, \phi_g)$

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

Conditional inference for the dispersions
- Need to adjust for estimation of $\beta_g$ when forming likelihood for $\phi_g$
- 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 $\beta_g$

Performance of conditional estimators of dispersion

Common dispersion likelihood
Assume same dispersion for all genes $\phi_g = \phi$
Genewise conditional log-likelihood $\ell_g(\phi; y_g)$
Common-dispersion log-likelihood $\ell_c(\phi) = \frac{1}{G} \sum_g \ell_g(\phi; y_g)$
Maximized at $\phi_c$

Complexity of dispersion: sharing information between genes
- Separate gene-wise estimation of $\phi_g$ is impractical
- Common dispersion (Robinson & Smyth 2008)
- Trended dispersion (Anders & Huber 2010)
- Gene-wise by empirical Bayes shrinkage (Robinson & Smyth, 2007)

Empirical Bayes shrinkage for the dispersion
Estimate $\phi_g$ by empirical posterior mode:
$$\text{Posterior} = \ell_g(\phi_g) + \alpha \ell_c(\phi_c)$$
- Genewise likelihood
- Precision of prior
- Empirical prior distribution
- Local weighting produces abundance dependent prior
Analysis of multifactor RNA/ChIP-Seq experiments with respect to biological variation

Gordon Smyth
Bioconductor, Seattle 28 July 2011

Estimated dispersions (simulation)

Oral squamous cancer

Tuch et al,
PloS ONE 2010

Multidimensional scaling plot using BCV distances

BCV=40%

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

Multidimensional scaling plots with BCV as distance

Finding technical effects
Analysis of multifactor RNA/ChIP-Seq experiments with respect to biological variation

Gordon Smyth
Bioconductor, Seattle 28 July 2011

Empirical Bayes for the fold changes

Predicting PCR fold changes from SEQC RNA-Seq data

Accuracy

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

Screening for splice-variants

Rtn4 reticulon 4

Genes vary in length ...

6X number of fragments
More power to detect DE at a given threshold

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

Gordon Smyth
Bioconductor, Seattle 28 July 2011

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

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

Collaborators
- Lynn Corcoran
- Tim Thomas
- Anne Voss
- Bilal Sheikh
- Samir Taoudi
- Peter ’t Hoen
- Jane Visvader
- Geoff Lindeman
- Bhupinder Pal
- Marie-Liesse Asselin-Labat

Epigenetic landscape (MS)

Expression vs epigenetic changes

MS vs LP
LP vs ML
K4 change
K27 change
Expression change
Expression change

Repitools