Supplementary materials

Title: Gene expression variability – the other dimension in transcriptome analysis
Authors: Tristan V de Jong, Yuri M Moshkin & Victor Guryev

1. Supplementary methods

1.1. Statistical inference of non-Poisson noise in RNA-sequencing counts using GAMLSS

The Distribution of a gene's RNA-sequencing counts (X) is commonly defined by a negative binomial distribution as ($X \overset{\text{ind}}{\sim} NB(\mu, \alpha)$) with two orthogonal parameters: mean (μ) and overdispersion (α):

$$(eq.\ 1)\ P(X = x|\ \mu, \alpha) = \frac{\Gamma(1/\alpha + x)}{\Gamma(1/\alpha)\Gamma(x+1)} \left(\frac{1}{1+\alpha\mu}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{x}.$$

The orthogonality of μ and α stems from the Fisher information matrix as its element $I_{\mu\alpha} = -E \frac{\partial^2}{\partial\mu\partial\alpha} \log\big(P(X|\ \mu, \alpha)\big) = 0$, which implies an asymptotic independence between μ and α. From a biological prospective, this means that average expression (μ) and overdispersion (α) are likely to be controlled independently.

The NB distribution belongs to a family of mixed-Poisson distributions, for which the expectation (mean) of mRNA counts: $E(X) = \mu$ and the variance depends quadratically on the μ: $Var(X) = \mu + \alpha\mu^2$ (Rigby, Stasinopoulos, & Akantziliotou, 2008). Based on this, the total gene noise, expressed as a squared coefficient of variation, partitions into Poisson (1st summand) and non-Poisson (2nd summand) or "intrinsic" and "extrinsic":

$$(eq.\ 2)\ cv^2(X) = Var(X)/E(X)^2 = (\mu + \alpha\mu^2)/\mu^2 = \mu^{-1} + \alpha.$$

Thus, non-Poisson noise is represented by the overdispersion parameter of $NB(\mu, \alpha)$ and it has been coined as a squared biological coefficient of variation - bcv^2 (McCarthy, Chen, & Smyth, 2012).

Statistical inference of the Poisson noise from mRNA counts is straightforward as it is a reciprocal of the mean of the mRNA copy number. This could be evaluated through generalized linear modeling (GLM) of differential gene expression (and thus, Poisson noise), which is implemented in a number of software packages including *edgeR* (McCarthy et al., 2012). However, GLM allows for the analysis of factor effects on only one parameter of the $NB(\underline{\mu}, \alpha)$, namely the mean, but not for the overdispersion $NB(\mu, \underline{\alpha})$.

The Bioconductor package *MDseq* attempts to extend the GLM to model both parameters of the NB (Ran & Daye, 2017), but it has several limitations and drawbacks. The major one is that *MDseq* parametrization of the NB implies a linear mean-variance relationship: $Var(X) = \mu\varphi$, where $\varphi = 1 + \alpha\mu$. Consequently, *MDseq* does not allow to directly model Poisson and non-Poisson noise.

Generalized Additive Model for Location, Scale and Shape (GAMLSS) offers a unique opportunity to model fixed and/or random factor effects on both parameters of the $NB(\mu, \alpha)$ distribution as defined in the *eq. 1* (Stasinopoulos et al., 2017). Here, we showcase the GAMLSS approach to model the age-dependent effects on both the mean counts and the non-Poisson noise of genes expressed in the liver of C57BL/6J mouse strain (Müller et al., 2018). To this end, the GAMLSS model was specified as following:

(*eq. 3a*) $\log(X_i) \sim age_j\beta_{\mu_j} + log(N_i) \Rightarrow \log(X_i/N_i) \sim age_j\beta_{\mu_j}$,

(*eq. 3b*) $\log(\alpha) \sim age_j\beta_{\alpha_j}$,

where $i = 1, \dots, n$ is i^{th} observation of gene's mRNA counts (X_i) and $j = 1, \dots, p$ - j^{th} factor level (young – 5 weeks, old – 20 weeks).

The first equation of GAMLSS (*eq. 3a*) specifies a model of a factor effect, namely age_j, on library size (N_i) normalized mean mRNA counts ($\mu_j = e^{\beta_{\mu_j}}$, $cpm_j = 10^6\mu_j$). Essentially, this part of the model corresponds to a GLM model of differential gene expression (McCarthy et al., 2012), however, GAMLSS allows for more flexibility as random effects and smoothing terms can be also included (Stasinopoulos et al., 2017). The second equation of GAMLSS (*eq. 3b*) models a factor effect on non-Poisson noise (α), where β_{α_j} is an MLE estimation of overdispersion parameter ($\alpha_j = e^{\beta_{\alpha_j}}$).

The significance of factor effects on both the mean and overdispersion (non-Poisson noise) could then be evaluated by calculating the log-likelihood ratio test statistics (D_μ and D_α respectively), which are asymptotically χ^2-distributed, as following:

$$(\textit{eq. 4a})\ D_{\mu} = -2log\frac{\text{likelihood for reduced model}}{\text{likelihood for GAMLSS model}} = -2log\frac{\mathcal{L}(\mu_0,\alpha_j \mid X_{ij})}{\mathcal{L}(\mu_j,\alpha_j \mid X_{ij})},$$

$$(\textit{eq. 4b})\ D_{\alpha} = -2log\frac{\text{likelihood for GLM model}}{\text{likelihood for GAMLSS model}} = -2log\frac{\mathcal{L}(\mu_j,\alpha_0 \mid X_{ij})}{\mathcal{L}(\mu_j,\alpha_j \mid X_{ij})},$$

The null model for D_{μ} excludes factor effects on the mean mRNA counts: $\log(X_i/N_i)\sim\beta_{\mu_0}$ and the D_{α} null model omits factor effect on the overdispersion as in a GLM model.

The estimation of differential gene expression by GAMLSS might differ from one performed by a GLM as the latter estimates only the shared overdispersion. For GLM, the log-likelihood ratio test statistic is calculated as:

$$(\textit{eq. 4c})\ D_{\mu_{GLM}} = -2log\frac{\text{likelihood for null model}}{\text{likelihood for GLM model}} = -2log\frac{\mathcal{L}(\mu_0,\alpha_0 \mid X_{ij})}{\mathcal{L}(\mu_j,\alpha_0 \mid X_{ij})}.$$

In practice, given small sample sizes (usually, from our experience, about 3) for many RNA-sequencing experiments a GLM model suffices to evaluate differential gene expression. However, if larger sample sizes are available, then the statistical inference of non-Poisson noise by GAMLSS might yield novel insights into the other dimension of gene regulation.

Below, we provide a detailed R script illustrating a GAMLSS analysis of genes expressed in the liver of young (5 weeks) and old (20 weeks) C57BL/6J mice (Müller et al., 2018). To execute the script, the R package *gamlss* and Bioconductor package *edgeR* should be installed and the supplementary data file (gamlss_age.Rdat) should be loaded. RNA-sequencing counts are contained within the *Counts_edgeR* variable (*edgeR* object) along with a phenotypic description (Age), see *Counts_edgeR$samples*. The script returns the *gamlss_NB_clean* data frame variable, which is also included in the supplementary data file (gamlss_age.Rdat). *gamlss_NB_clean* contains GAMLSS estimates of the expression level (counts per million reads, CPM) and non-Poisson noise ($bcv = cv(\mu) = \sqrt{\alpha}$) for genes expressed in young and old mice along with a statistical evaluation of the impact of age on both parameters.

1.2. Computer code for typical analysis of gene expression variability

See computer code in file “ExpVarQuant.R”

1.3. Downstream bioinformatics analysis of gene’s non-Poisson variation

A downstream analysis of the non-Poisson noise presented in the main text. Figures 2-5 have been plotted with standard R tools, except for the KEGG-pathway gene annotation. The annotation of each mouse gene’s promoter elements: TATA-box, Initiator motif, CCAAT-box and GC-box was extracted from the Eukaryotic Promoter Database (Dreos, Ambrosini, Cavin Périer, & Bucher, 2013). A classical pathway enrichment analysis requires a selection of genes significantly affected by a factor of interest followed by their mapping to biological pathways and a subsequent calculation of the over-representation statistic. However, the selection of genes relies on a) the statistical testing procedure, b) the statistical power of a test and c) the threshold for rejection of null hypothesis. All of these might interfere with pathway enrichment analysis. Thus, we turned to a ridge linear regression model to annotate the KEGG-pathways (Kanehisa, Sato, Kawashima, Furumichi, & Tanabe, 2016) associated with age-mediated changes in non-Poisson noise (Hastie, Tibshirani, & Friedman, 2009). In brief, model coefficients $\hat{\beta}^{\mathrm{ridge}}$ were estimated by the minimizing of penalized residual sum of squares (RSS):

$$(\textit{eq. 5})\ \hat{\beta}^{\mathrm{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \underbrace{\textstyle\sum_i^n \left(y_{ij} - \sum_j^p \beta_j \mathrm{KEGG}_{ij} \right)}_{=RSS} + \Lambda \textstyle\sum_j^p \beta_j^2 \right\},$$

where y – response variable, KEGG – *n x p* model matrix (*n* – number of genes, *p* – number of KEGG pathways), β – ordinary least squares coefficients, Λ – ridge penalty factor. Genes mapped to a given KEGG-pathway were assigned 1 or 0 in the KEGG model matrix. Λ was chosen to minimize the mean squared error. Ridge regression has multiple advantages over an enrichment analysis: a) it does not rely on gene selection, b) it allows for a general estimation, whether and to what a response variable extends depends on the annotated gene’s functions, c) the values of the model coefficients can be interpreted directly in terms of the variable importance and the coefficient’s sign indicates the direction of an effect, d) ridge regression copes with ill-formulated problems, i.e. over-parametrization *p* > *n* and multicollinearity. Below, we demonstrate an implementation of a KEGG ridge regression model using the R package *h2o*.

1.4. Computer code for functional annotation of genes detected in gene expression variability analysis.

See computer code in file "KEGG_Annotation.R"

1.5. Supplementary references

Dreos, R., Ambrosini, G., Cavin Périer, R., & Bucher, P. (2013). EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era. *Nucleic Acids Research*, *41*(D1), D157–D164. http://doi.org/10.1093/nar/gks1233

Hastie, T., Tibshirani, R., & Friedman, J. (2009). Basis Expansions and Regularization (pp. 139–189). http://doi.org/10.1007/978-0-387-84858-7_5

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., & Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, *44*(D1), D457–D462. http://doi.org/10.1093/nar/gkv1070

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297. http://doi.org/10.1093/nar/gks042

Müller, C., Zidek, L. M., Ackermann, T., de Jong, T., Liu, P., Kliche, V., … Calkhoven, C. F. (2018). Reduced expression of C/EBPβ-LIP extends health and lifespan in mice. *ELife*, *7*. http://doi.org/10.7554/eLife.34985

Ran, D., & Daye, Z. J. (2017). Gene expression variability and the analysis of large-scale RNA-seq studies with the MDSeq. *Nucleic Acids Research*, *45*(13), e127–e127. http://doi.org/10.1093/nar/gkx456

Rigby, R. A., Stasinopoulos, D. M., & Akantziliotou, C. (2008). A framework for modelling overdispersed count data, including the Poisson-shifted generalized inverse Gaussian distribution. *Computational Statistics & Data Analysis*, *53*(2), 381–393. http://doi.org/10.1016/J.CSDA.2008.07.043

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., Bastiani, F. De, Rigby, R. A., … Bastiani, F. De. (2017). *Flexible Regression and Smoothing*. Chapman and Hall/CRC. http://doi.org/10.1201/b21973