Comparative analysis of DE gene analysis tools for RNA sequencing time course data

Introduction

- Comparison between the existing TC RNA-Seq data analysis tools.
- EBSeqHMM, edgeR, DEseq2 and Next maSigPro

Next maSigPro

A method to identify significantly differential expression profiles in time-course RNA-seq experiments

Introduction

- Originally developed for Microarray data
- Updates to support count data by introducing generalized linear models
- Statistical procedure to identify differentially expressed genes in time-course data
- Two-step regression strategy
    1. Selects genes with non-flat profiles
    2. Selects the best regression model for each gene with time or series-associated changes

Methodology

- Data has to be normalized at the beginning
- Model
- Time - Continuous variable
- Experimental conditions – can be categorical or quantitative
- Example: with t=1,…, T time points and Z- experimental condition with two levels
- The gene expression value y_i at condition i at time t_i
- The polynomial model is,
- $\mu_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \beta_3\, z_{1i} + \beta_4 t_i z_{1i} + \beta_5\, t_i^2 z_{1i}$
- $y_i \sim NB(\mu_i, \theta)$ and $E(y_i) = \mu_i$ and $V(y_i) = \mu_i + \frac{\mu_i^2}{\theta}$ for i=1,2

- All the parameters are estimated using MLE including overdispersion
- Then we calculate the goodness of fit for each model using deviance statistic
- Select the DE gene list by using 0.05 cutoff for FDR.

EBSeqHMM

- It applies an empirical Bayes autoregressive hidden Markov model (AR-HMM).
- First, parameters are estimated using a negative binomial (NB) model.
- Then categorize genes at each time point by a Markov-switching autoregressive model and classify genes into expression paths.
- Requires a minimum of 3 time points.

- Expression (X) for gene (g) at time (t) for sample (n) is NB distributed with dependencies on the mean (r) and variance (q) of the previous time point and the regression change status (S) being either up, down or stable.

- $$\left(X_{gnt} \middle| r_{g,t-1}, q_{g,t-1}, S_g^{\Delta t} = s\right) \sim NB\left(r_{g,t-1}\xi_g^s, q_{g,t-1}\right)$$

  - with $$\xi_g^s = \begin{cases} c, & s = up \\ \frac{1}{c}, & s = down \\ 1, & s = stable \end{cases}$$

- Fluctuations of the mean are modeled by defining a prior distribution, thus the marginal predictive conditional distribution becomes Beta-Negative Binomial.

- $$\begin{cases} \left(q_{gt} \middle| \alpha, \beta, X_{g,t-1} = x_{g,t-1}\right) \sim Beta\left(\alpha + N_t r_{g,t-1}, \beta + \sum_j x_{g,t-1,j}\right) & , \mathrm{g} > 1, \mathrm{t} > 1 \\ \left(q_{gt} | \alpha, \beta,\right) \sim Beta(\alpha, \beta) & , t = 1 \end{cases}$$

- $$\begin{cases} \left(X_{gtn} \middle| X_{g,t-1} = x_{g,t-1}, S_g^{\Delta t}, \Theta\right) \sim BetaNB\left(\alpha + N_{t-1} r_{g,t-1}, \beta + \sum_j x_{g,t-1,j}, \xi_g^s r_{g,t-1}\right) & , t > 1 \\ \left(X_{g1n} | \Theta\right) \sim BetaNB\left(\alpha, \beta, r_{g,1}\right) & , t = 1 \end{cases}$$

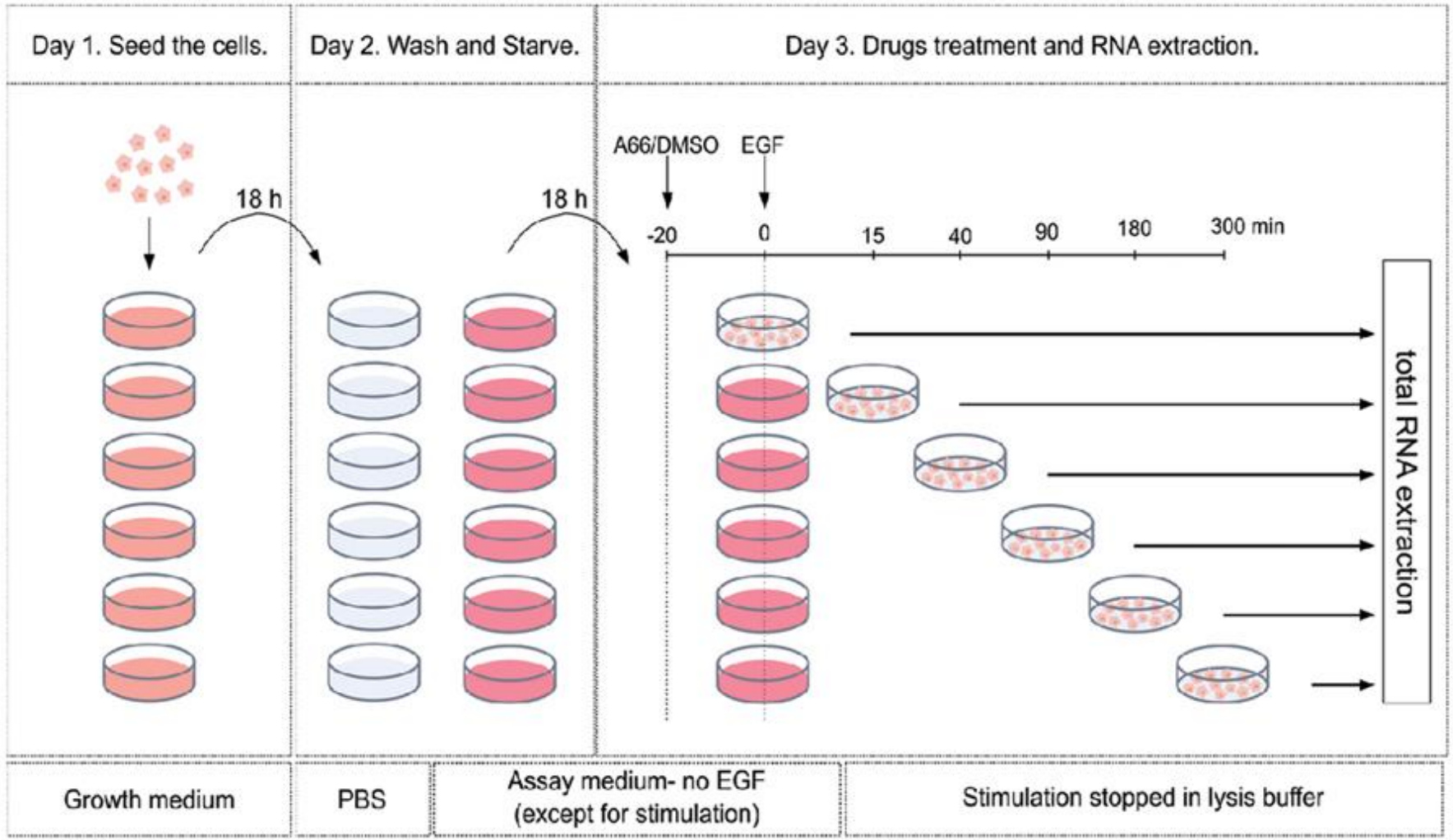
- with $\Theta = \left[\alpha, \beta, r_{g,t-1}, \xi_g^s\right]$

edgeR / DEsqe2

- While having different methods for estimating the dispersion,
- Based on a NB model and are considered as gold standards in the DE analysis field.
- Generalized linear models (GLM) as well as a likelihood ratio tests.
- Can do pair-wise

- Count matrix *K* with one row for each gene *i* and one column for each sample *j*.
- $K_{ij} \sim \text{NB}(\mu_{ij}, \alpha_i)$
- The mean is taken as a quantity q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, scaled by a normalization factor s_{ij}, i.e., $\mu_{ij} = s_{ij} q_{ij}$
- $log(q_{ij}) = \sum_r x_{jr} \beta_{ir}$

Example



Adapted from "Perturbations of PIP3 signalling trigger a global remodelling of mRNA landscape and reveal a transcriptional feedback loop" by Kiselev et al., NAR 2015