ZINB-WaVE: dimension reduction and signal extraction for zero-inflated count data analysis

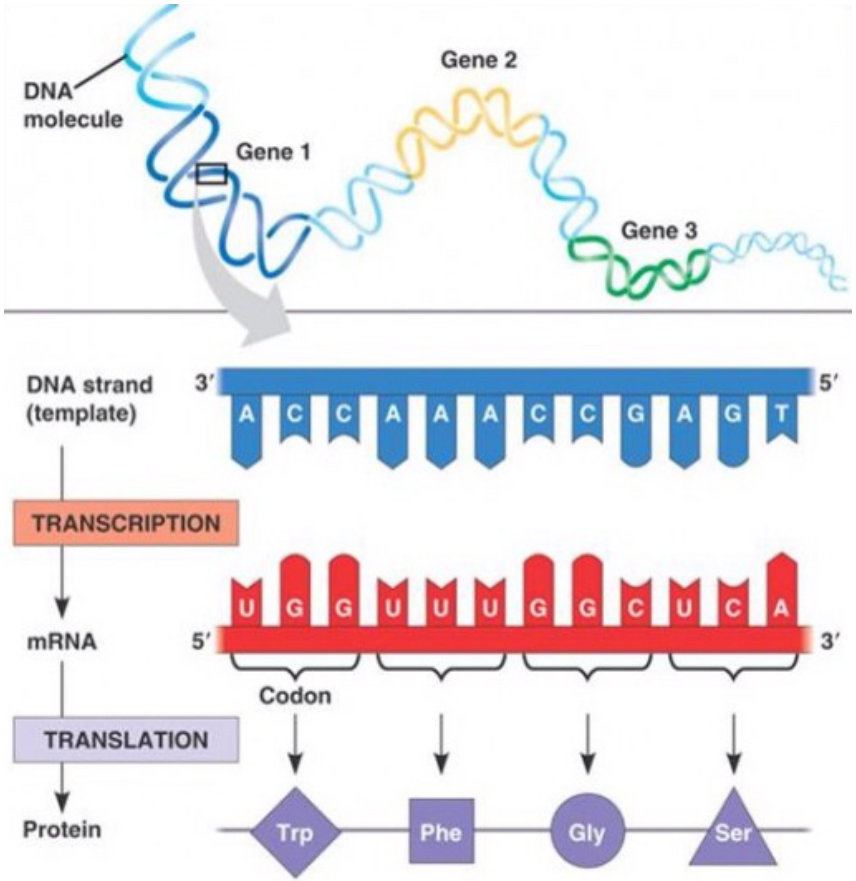
Joint work with Davide Risso, Fanny Perraudeau, Sandrine Dudoit and Jean-Philippe Vert

Svetlana Gribkova

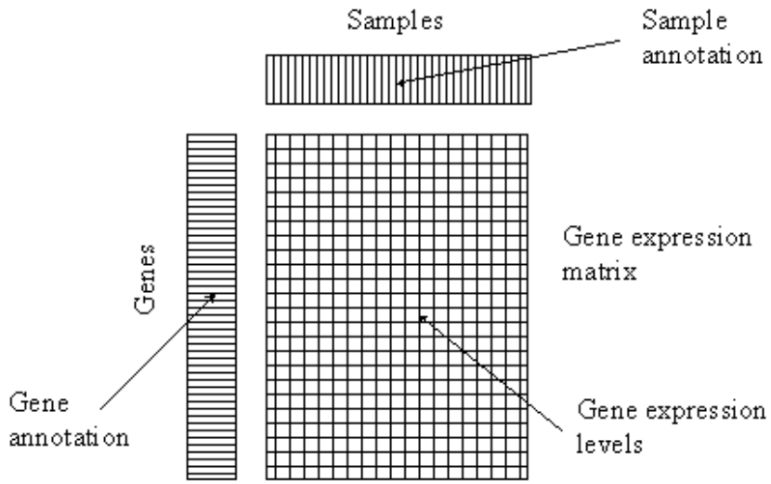
Laboratoire de Probabilités et Modèles Aléatoires, Université Paris Diderot

Mathematical Methods of Modern Statistics,
13 July 2017, CIRM Luminy

Gene expression and RNA-Seq data

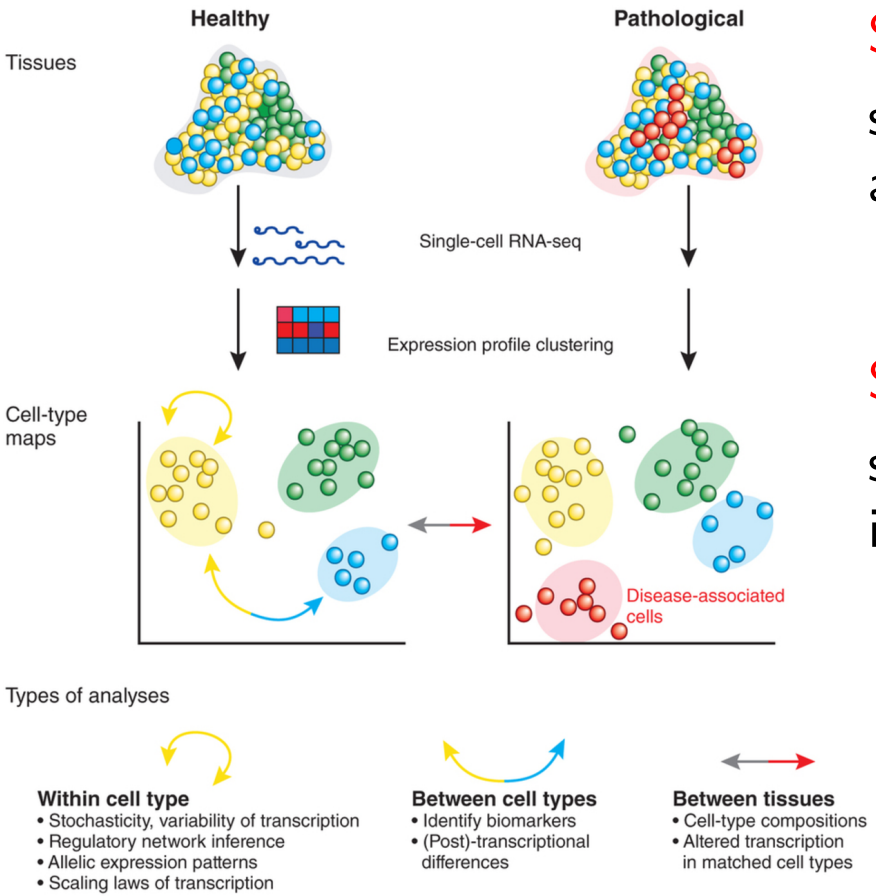


- Gene expression level = number of its RNA copies in the biological sample
- RNA-Seq = technology allowing to quantify RNA copies from each gene
- Gene expression matrix :



Single cell RNA-Seq : from tissue level to cell level

- Gene expressions varies across tissues (healthy vs illness)
- Gene expression also varies from cell to cell inside the same tissue !



Standard RNA-Seq :

sensitive enough to measure gene expressions averaged across single cells of the same tissue

Single-cell RNA-Seq (2009) :

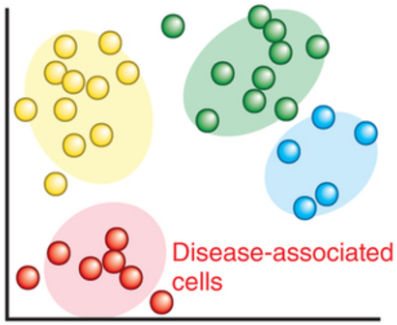
sensitive enough to measure gene expressions in individual single cells

- Compare gene expression distributions instead of their averages
- Study and compare structures of intercellular heterogeneity of expressions

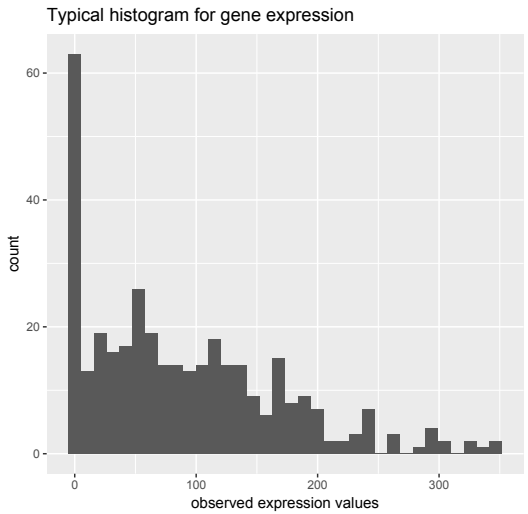
Dimension reduction for single-cell data analysis

Each single cell is described by $J \approx 10^4$ features $\rightarrow$ need the dimension reduction for visualization and clustering :

$$\begin{bmatrix} & \text{gene 1} & \text{gene 2} & \dots & \text{gene J} \\ \text{cell 1} & Y_{11} & Y_{12} & \dots & Y_{1J} \\ \text{cell 2} & Y_{21} & Y_{22} & \dots & Y_{2J} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ \text{cell n} & Y_{n1} & Y_{n2} & \dots & Y_{nJ} \end{bmatrix} \longrightarrow$$



Why do we need a new statistical model for the dimension reduction ?



Count data with inflation of zeros : many genes have positive RNA counts in some cells but zero counts in other cells ("dropout")

Over-dispersion : variance $>$ mean

Systematic noise : technical factors affecting measurements, normalization factors, etc.

Dimension reduction with zero-inflated count noise model and covariates

- Low-dimensional projection of data should summarize biological sources of expressional heterogeneity
- Low-dimensional projection by standard methods is determined by variation in number of $\neq 0$ genes and technical variation

Dimension reduction with zero-inflated count noise and covariates :

$$\begin{bmatrix} & \text{gene 1} & \text{gene 2} & \dots & \text{gene J} \\ \text{cell 1} & Y_{11} & Y_{12} & \dots & Y_{1J} \\ \text{cell 2} & Y_{21} & Y_{22} & \dots & Y_{2J} \\ \vdots & \vdots & \vdots & & \vdots \\ \text{cell n} & Y_{n1} & Y_{n2} & \dots & Y_{nJ} \end{bmatrix} \quad \text{noisy version of} \quad \begin{bmatrix} & \text{gene 1} & \text{gene 2} & \dots & \text{gene J} \\ \text{cell 1} & \mu_{11} & \mu_{12} & \dots & \mu_{1J} \\ \text{cell 2} & \mu_{21} & \mu_{22} & \dots & \mu_{2J} \\ \vdots & \vdots & \vdots & & \vdots \\ \text{cell n} & \mu_{n1} & \mu_{n2} & \dots & \mu_{nJ} \end{bmatrix}$$

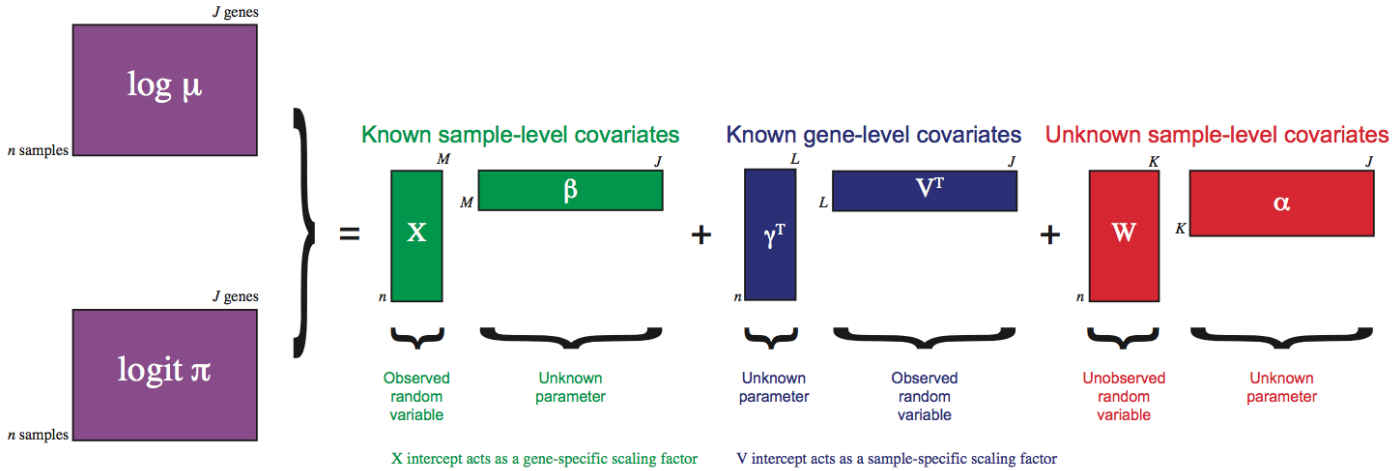
$$Y_{ij} \sim \pi_{ij}\delta_0(y) + (1-\pi_{ij})f_{NB}(y;\mu_{ij},\theta_{ij})$$

$$f_{NB}(y;\mu,\theta) = \frac{\Gamma(y+\theta)}{\Gamma(y+1)\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu}\right)^{\theta}\left(\frac{\mu}{\mu+\theta}\right)^{y}, \quad \forall y \in \mathbb{N}.$$

$\triangle$ $\pi := \{\pi_{ij}\}$ unknown latent matrix of zero inflation probabilities

ZINB-WaVE : matrix factorisation model

$\log \mu$ and $\text{logit}\, \pi$ may be explained by a small number of known and latent factors :



- X : $n \times M$ known matrix of cell level covariates (biol. or tech.)
- V : $J \times L$ known matrix of gene level covariates (e.g. gene length)
- W : $n \times K$ unknown matrix of K latent factors
- $\beta_\mu, \gamma_\mu, \alpha_\mu, \beta_\pi, \gamma_\pi, \alpha_\pi$ are unknown matrices of coefficients

Summary of the model and comments

Summary of the model : $Y_{ij} \sim \pi_{ij}\delta_0(y) + (1-\pi_{ij})f_{NB}(y;\mu_{ij},\theta_{ij})$

$$\begin{aligned} \log(\mu_{ij}) &= \left(X\beta_\mu + (V\gamma_\mu)^\top + W\alpha_\mu + O_\mu\right)_{ij} \\ \text{logit}(\pi_{ij}) &= \left(X\beta_\pi + (V\gamma_\pi)^\top + W\alpha_\pi + O_\pi\right)_{ij} \\ \ln(\theta_{ij}) &= \zeta_j \end{aligned}$$

Comments :

- Higher expression of gene $\Rightarrow$ smaller probability of non detection $\Rightarrow$ factors X, V, W are shared
- W is $n \times K$ matrix giving a low dimensional representation of n cells in K-dimensional space ($\approx$ PCA with the appropriate model for noise)
- X and V allows to account explicitly for known covariates and capture in W only unknown sources of heterogeneity

Estimation of the model

The observed data is a $n \times J$ matrix of counts $\{Y_{ij}\}$, the log-likelihood function is given by :

$$\ell(\beta, \gamma, W, \alpha, \zeta) = \sum_{i=1}^{n} \sum_{j=1}^{J} \ln f_{ZINB}(Y_{ij}; \mu_{ij}, \theta_{ij}, \pi_{ij})$$

Parameters are estimated via the max of the penalized log-likelihood :

$$\max_{\beta,\gamma,W,\alpha,\zeta} \{\ell(\beta, \gamma, W, \alpha, \zeta) - \text{Pen}(\beta, \gamma, W, \alpha, \zeta)\},$$

with

$$\text{Pen}(\beta, \gamma, W, \alpha, \zeta) = \frac{\epsilon_\beta}{2}||\beta^0||^2 + \frac{\epsilon_\gamma}{2}||\gamma^0||^2 + \frac{\epsilon_W}{2}||W||^2 + \frac{\epsilon_\alpha}{2}||\alpha||^2 + \frac{\epsilon_\zeta}{2}\text{var}(\zeta),$$

where $(\epsilon_\beta, \epsilon_\gamma, \epsilon_W, \epsilon_\alpha, \epsilon_\zeta)$ is the set of regularization parameters and β^0 means β without the intercept

Initialization. Approximate positive counts by log-normal distribution and alternate between the following steps :

1. Adjust for known covariates (β_μ and γ_μ) by ridge regression :

$$\min_{\beta_\mu,\gamma\mu} \sum_{(i,j)\in\mathcal{P}} (L_{ij} - (X\beta_\mu)_{ij} - (V\gamma_\mu)_{ji})^2 + \frac{\epsilon_\beta}{2}||\beta_\mu^0||^2 + \frac{\epsilon_\gamma}{2}||\gamma_\mu^0||^2 .$$

2. Regress out known effect and optimize in W and α_μ :

$$\min_{W,\alpha_\mu} \sum_{(i,j)\in\mathcal{P}} \left(L_{ij} - (V\hat{\gamma}_\mu)_{ji} - (X\hat{\beta}_\mu)_{ij} - (W\alpha_\mu)_{ij}\right)^2 + \frac{\epsilon_W}{2}||W||^2 + \frac{\epsilon_\alpha}{2}||\alpha_\mu||^2 .$$

3. Initialize $(\beta_\pi, \gamma_\pi, \alpha_\pi)$ as solutions of regularized logistic regression :

$$\min_{(\beta_\pi,\alpha_\pi,\gamma_\pi)} \sum_{(i,j)} \Big[- \hat{Z}_{ij}(X\beta_\pi + (V\gamma_\pi)^\top + \hat{W}\alpha_\pi)_{ij}$$
$$+ \ln\left(1 + e^{(X\beta_\pi + (V\gamma_\pi)^\top + \hat{W}\alpha_\pi)_{ij}}\right)\Big] + \frac{\epsilon_\beta}{2}||\beta_\pi||^2 + \frac{\epsilon_\gamma}{2}||\gamma_\pi||^2 + \frac{\epsilon_\alpha}{2}||\alpha_\pi||^2 .$$

Optimization. Alternate between the following steps :

1. Optimize in dispersion parameter :

$$\hat{\zeta} \leftarrow \arg\max_{\zeta} \left\{ \ell(\hat{\beta}, \hat{\gamma}, \hat{W}, \hat{\alpha}, \zeta) - \frac{\epsilon_\zeta}{2} \mathrm{var}(\zeta) \right\}$$

2. Optimize in cell level unknown coefficients :

$$\left(\hat{\gamma}, \hat{W}\right) \leftarrow \arg\max_{(\gamma, W)} \left\{ \ell(\hat{\beta}, \gamma, W, \hat{\alpha}, \hat{\zeta}) - \frac{\epsilon_\gamma}{2} ||\gamma^0||^2 - \frac{\epsilon_W}{2} ||W||^2 \right\}$$

3. Optimize in gene level unknown coefficients :

$$\left(\hat{\beta}, \hat{\alpha}\right) \leftarrow \arg\max_{(\beta, \alpha)} \left\{ \ell(\beta, \hat{\gamma}, \hat{W}, \alpha, \hat{\zeta}) - \frac{\epsilon_\beta}{2} ||\beta^0||^2 - \frac{\epsilon_\alpha}{2} ||\alpha||^2 \right\}$$

4. Orthogonalization (orthogonalize factors ; maximize locally)

$$\left(\hat{W}, \hat{\alpha}\right) \leftarrow \arg\min_{(W, \alpha) \,:\, W\alpha = \hat{W}\hat{\alpha}} \frac{1}{2} \left(\epsilon_W ||W||^2 + \epsilon_\alpha ||\alpha||^2 \right) .$$

Simulations

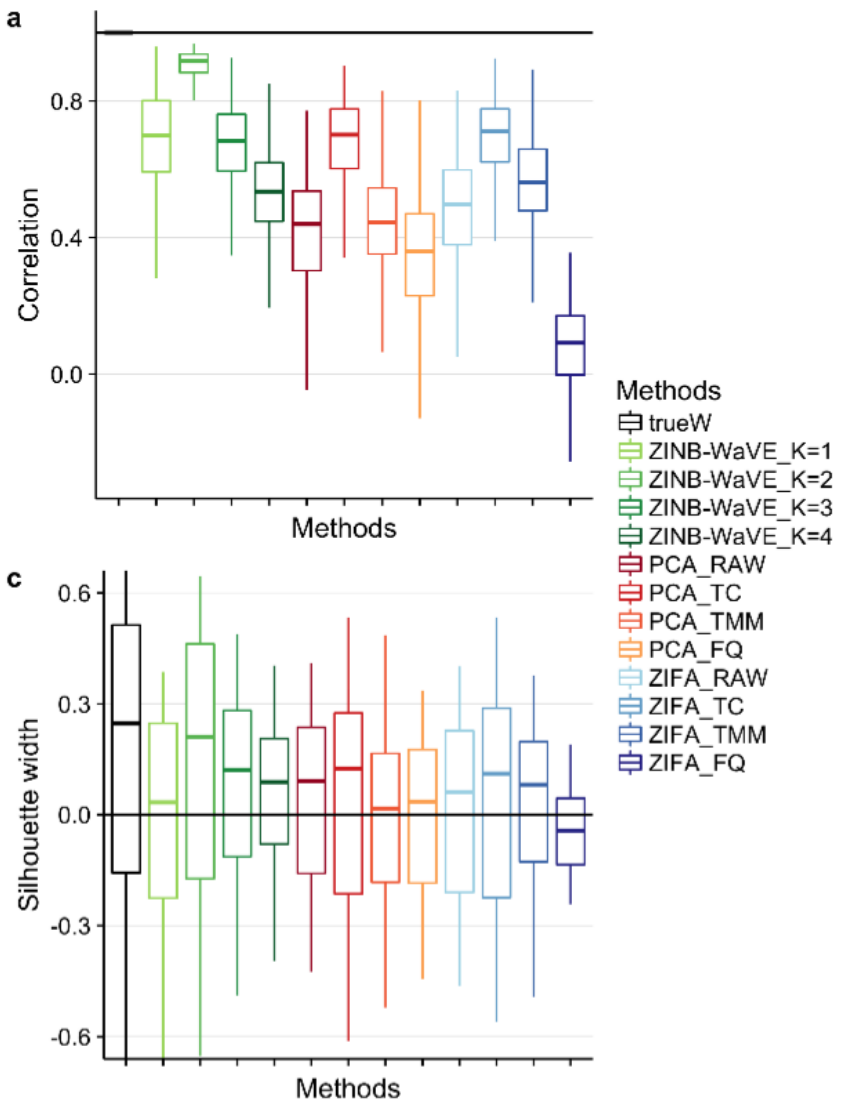
PCA, ZIFA, ZINB-WaVE were compared on simulated data :

- Data were simulated from ZINB-WaVE model, using W with $K = 2$
- Rows of W were simulated in a way to induce known clusters of single cells
- Different proportions of zeros

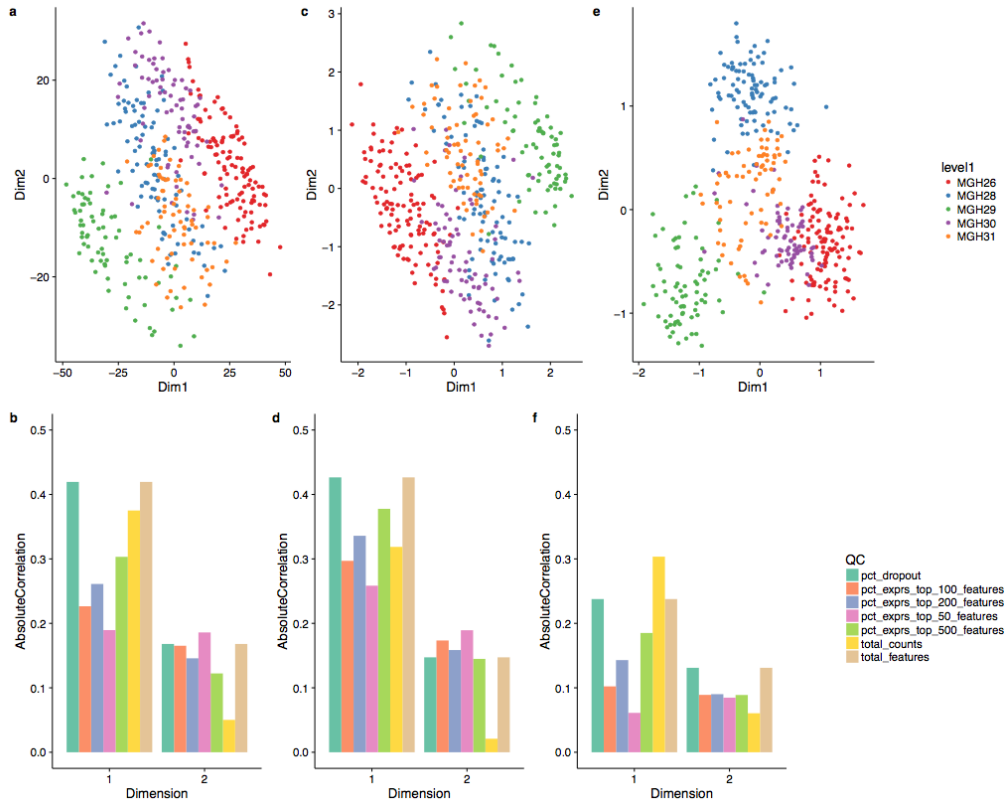
Criteria :

Quality of the low dimensional projection : correlation of distances between cells in true and estimated projection

Quality of cluster recovery : silhouette width of clustering based on estimated W



Glioblastoma dataset : real data with 430 cells from 5 patients suffering from glioblastoma :



- ZINB-WaVE leads to tighter clusters grouping cells by patients
- ZINB-WaVE axes less correlated to quality control measures of cells

Conclusion and references

Conclusions :

- Dimension reduction based on zero-inflated negative binomial model of noise allows for a better quality of low-dimensional representation of the data
- Clustering based on ZINB-based low-dimensional representation is higher quality compared to PCA or ZIFA.
- Covariates allow to include all known information and W captures only the unknown sources of heterogeneity

References :

- Preprint : http://biorxiv.org/content/early/2017/04/06/125112
- Package : http://github.com/drisso/zinbwave