TP1 **Network reconstruction and analysis**

Durée: (voir date sur le site de cours)

Objectif: Reconstruction et analyse d'un réseau de gènes.

- Le TP doit être réalisé en binôme. Merci de bien préciser les noms et prénoms de chaque membre du binôme et de mettre en copie votre binôme lors de l'envoi des scripts et de votre rapport.
- Les résultats de vos analyses doivent être commentés et reprendre les différents notions vues en cours.
- Vous devez me faire parvenir votre projet (scripts et rapport) pour la date indiquée sur le site de cours à (*severine.affeldt@parisdescartes.fr*).
  Titre du message: [**MLDS-app. Graphes**] ou [**MLDS-fi. Graphes**]

Analysis of expression data in single cells

Problématique

We want to reconstruct a regulatory networks from single cell expression data at the time of endothelial and hematopoietic differentiations from the primitive streak cells of the mouse early embryo, Fig 2. This concerns the formation of primitive erythroid cells, a distinct and transient red blood cell lineage arising directly from mesodermal progenitors with restricted hematopoietic potential, by contrast to the highly studied definitive erythroid cells which arise from multipotent hematopoietic stem cells.

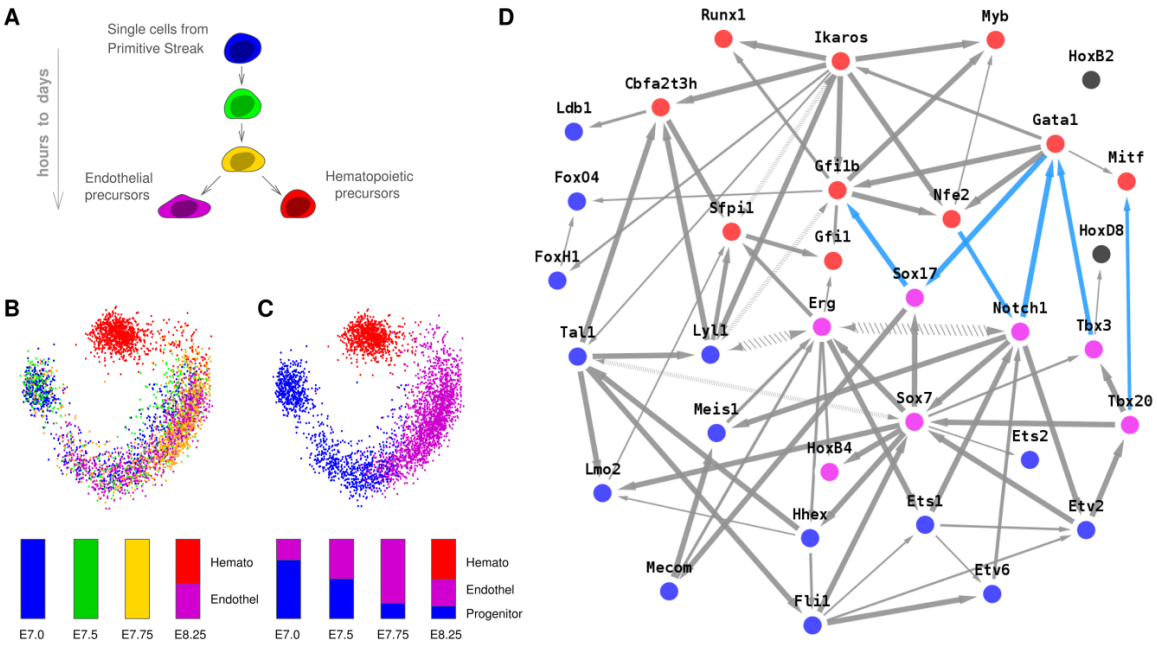

Fig 2. Network reconstruction at cellular level. (A) Hematopoietic / endothelial differentiation in single cells from mouse embryos [24]. (B) Principal component analysis and (C) K-means clustering of gene expression data [24] with histograms showing the relative proportions of cell populations at each data point (E7.0 to E8.25). (D) Hematopoietic / endothelial differentiation regulatory network between hematopoietic specific (red), endothelial (violet), common (blue) and unclassified (gray) TFs. Graph predicted with `miic` R-package and visualized using cytoscape (blue edges correspond to repressions).

The dataset for this application includes the expression of 33 transcription factors (TFs) along with 13 non-TF genes (markers) in 3,934 single cells extracted at 4 different times of the mouse embryo development (days E7.0, E7.5, E7.75 and E8.25), Fig 2. The cells extracted from E8.25 were also divided in two different pools: potential endothelial precursors and potential hematopoietic precursors based on the expression of the Runx1 hematopoietic marker. Gene expression was collected using single cell qRT-PCR and binarized, leading to two-state (on / off) expression levels in the available dataset.

Pooling all cells together regardless of their developmental timing (from day E7.0 to E8.25), we can analyze their population heterogeneity using principal component analysis (PCA), Fig 2B, and K-means clustering, Fig 2C. Three main cell populations are identified and can be interpreted, based on gene functional classification, as progenitor, endothelial precursor and hematopoietic precursor populations, whose relative proportions vary from E7.0 to E8.25, Fig 2C. See *Learning causal networks with latent variables from multivariate information in genomic data* for details[1].

1. **TF Correlation network**

   Install the R package `miic` and explore the *hematoData* dataset. Provide the characteristics of the dataset. Compute all paiwise correlations. Plot the correlation network with full customization (eg., nodes/edges - size/color/sign) and clear layout. Try different threshold to filter the edges.

   Does your correlation network make sense from a biological point of view? Where are the repressions? Compare to the network obtained by Verny *et al.*

2. **TF Partial Correlation network**

   For linear, undirected systems, the inverse of the covariance matrix Σ^{-1} is related the partial correlations. In particular, $\Sigma^{-1} \propto L = D - A$, where D is the diagonal degree matrix and A is the adjacency matrix. The elements of Σ^{-1} are defined as $\sigma_{ij}^{-}1 = -\sqrt{(\sigma_{ii}^{-}1\sigma_{jj}^{-}1)}\rho_{x_i x_j}.V \setminus x_i x_j$, where $\rho_{x_i x_j}.V \setminus x_i x_j$ is the partial correlation coefficient between x_i and x_j while fixing *all* other nodes, $V \setminus x_i x_j$[2].

   Compute and invert the covariance matrix of the *hematoData* dataset. Caution: you can invert your matrix if the determinant is not null. If necessary, you can perform a regularization by adding a small value λ to the covariance matrix diagonal. Try different λ values. Does *lambda* have a strong influence on your reconstructec network?

   Plot the partial correlation network with full customization (eg., nodes/edges - size/color/sign) and clear layout. Try different threshold to filter the edges.

   Does your correlation network make sense from a biological point of view? Where are the repressions? Compare to the correlation network and the network obtained by Verny *et al.*

[1] `https://doi.org/10.1371/journal.pcbi.1005662`

[2] $\rho_{XY.Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{ZY}^2}}$