Différenciation des précurseurs hématopoïétiques chez l'embryon

Nathalie Lehmann[1], Mariam Sissoko[1]
Enseignants référents: **Hervé Isambert[2], Louis Verny[2], Nadir Sella[2]**

Résumé
Le but de ce projet de Master 2 de bioinformatique est de reconstruire le réseau de régulation régissant les expressions des facteurs de transcription clés pour la différenciation des précurseurs hématopoïétiques chez l'embryon. Dans cet objectif, nous avons d'abord procédé à un filtrage des données afin de garder les gènes d'intérêt, puis reconstruit des réseaux selon deux méthodes différentes : par clustering hierarchique et via l'algorithme polynomial PC[1] (*Peter-Clark*). Enfin, en comparant nos résultats avec ceux présents dans la littérature scientifique, nous ferons état d'un modèle graphique simplifié expliquant les mécanismes impliquant la différenciation des cellules primitives en deux lignées distinctes hématopoïétique et endothéliale. Cette reconstruction de réseau a été effectuée à partir de données analysées par *single cell RNA-seq* puis binarisées.

Mots-clés
Réseaux de régulation – Facteurs de transcription – Hématopoïèse

[1] *Master 2 Bioinformatique et Modélisation, Université Paris 6, France*
[2] *Institut Curie, France*

Table des matières

Introduction

Au cours du développement de l'embryon des Vertébrés, tous les tissus hématopoïétiques successivement actifs (foie, thymus, rate et moelle osseuse) sont colonisés par des cellules souches hématopoïétiques (CSH) d'origine extrinsèque. Le sac vitellin (SV) constitue l'unique exception à cette règle, puisque des CSH s'y développent in situ. Il a été observé que le SV constitue le premier site d'hématopoïèse de l'embryon[2] : c'est le lieu d'apparition des premières cellules sanguines propres à l'embryon. Cependant, de la lignée primitive à l'origine de celles-ci, émerge aussi les premières cellules endothéliales (constituant la paroi interne des vaisseaux sanguins). Dans ces conditions, quels sont les facteurs de transcription suffisants et/ou nécéssaires pour induire cette différenciation de la lignée primitive ?

Reconstruire le réseau de régulation contôlant cette différenciation pourrait permet de mieux appréhender les mécanismes de l'hématopoïèse primitive et de la formation des tissus sanguins. Or l'origine de certaines leucémies (ie l'anémie de Fanconi[3]) reste encore difficile à déterminer, et l'établissement de tels réseaux pourrait alors favoriser la compréhension et l'établissement de protocoles expérimentaux mieux ciblés.

Il est important de spécifier que ce projet s'appuie largement sur l'article de Moignard et al., *Decoding the regulatory network of early blood development from single-cell expression measurements*[4]. En effet, les données utilisées pour réaliser ce projet sont similaires

à celles utilisées par les auteurs de l'article sus-nommé, et la démarche globale de reconstruction de réseau est relativement semblable, bien qu'allégrement simplifiée.

1. Données et méthodes

La reconstruction des voies moléculaires contrôlant le développement embryonique des organes est entrâvé par le manque de méthodes adaptées pour l'étude de phénomènes extrêmement précis, d'autant plus si le matériel disponible est limitant. Les techniques traditionnelles telles que le RNA-seq se révèlent alors insuffisantes. La stratégie que les auteurs de l'article de référence ont choisi est particulièrement pertinente car il s'agit d'une approche combinant le séquençage *single cell* et les analyses computationnelles de reconstruction de réseaux à partir des graphes des états de transition. En effet, le séquençage *single cell* (sc) permet une analyse transcriptomique à l'échelle d'une seule cellule. Grâce au sc-RNAseq, il devient possible d'estimer l'hétérogénéité intra-tumorale, mais aussi d'étudier des stades embryonnaires précoces, ou encore de retracer les lignées cellulaires au cours du développement (source : bioinfo-fr) [1].

1.1 Dataset

Le dataset proposé rassemble les données d'expression binarisées de différents gènes pouvant soit être des facteurs de transcription (33 gènes), soit d'autres gènes marqueurs (spatiaux ou temporels - 9 gènes), ou encore des gènes servant de contrôle (*housekeepers* - 4 gènes). Chaque ligne du dataset correspond au profil d'expression d'une cellule, analysée par sc-RNAseq. Les colonnes, quant à elles, correspondent aux gènes dont l'expression a été quantifiée. Les données étant binaires, le '1' représente un gène exprimé dans la condition correspondante, le '0' un gène non exprimé. Au total, cela fait donc 46 gènes analysés dans 3934 cellules issues d'embryons de souris, prelevées à quatre stades différents du développement embryonnaire précoce. Celles-ci deviendront éventuellement des cellules sanguines (en rouge sur la figure 1) ou endothéliales (en violet). Comme indiqué sur la figure 1, cinq populations sont analysées :

— E7.0 (*primitive streak*, PS),

— E7.5 (*neural plate*, NP)

— E7.75 (*head fold*, HF).

— E8.25, cellules GFP+ (*four somite*, 4SG) cellules sanguines potentielles

— E8.25, cellules Flk1+GFP (4SFG) cellules endothéliales potentielles.

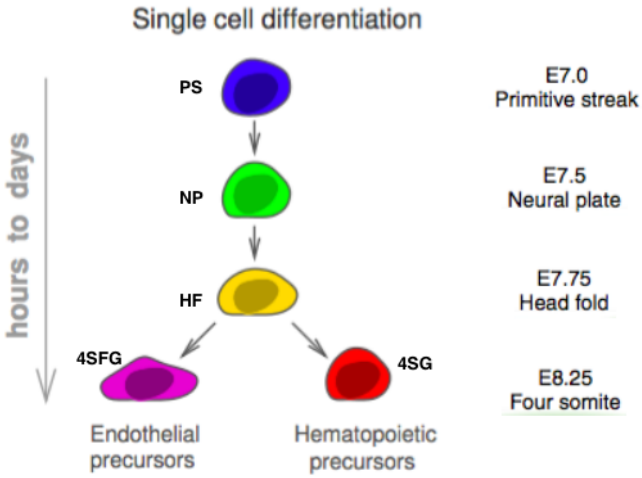

FIGURE 1. Processus de différenciation de la lignée primitive (PS) en 2 lignées distinctes : endothéliale (4SFG) et hématopoïétique (4SG)

Pour ceux qui souhaiteraient visualiser le cycle de développement embryonnaire murin de manière globale ainsi que les étapes critiques de l'hématopoïèse, les figures récapitulatives 11 et 12 sont accessibles dans la partie annexe.

1.2 Filtre des données

Afin d'obtenir un set de données non biaisées, nous avons choisi d'appliquer un filtre afin d'éliminer les gènes exprimés dans 100% des cas (codé en Python, on ôte du dataset les colonnes où il n'exsite que des '1'). Les gènes qui disparaissent alors sont référencés ci-dessous, et leur fonction en tant que *housekeepers* a été vérifiée via le site de la NCBI [2] : **Eif2b1**, **Mrpl19**, **Polr2a**, **Ubc**.

1.3 Gènes d'intérêt

Une fois le filtre appliqué, il nous a fallu procéder à différents tris des données.

1.3.1 Tri par facteur de transcription

Les 42 gènes restant dans notre dataset n'étant pas tous impliqués directement dans la différenciation cellulaire, nous avons effectué un premier tri où ne sont conservés que les 33 facteurs de transcription. Les gènes

1. http ://bioinfo-fr.net

2. https ://www.ncbi.nlm.nih.gov/gene

marqueurs qui ne se trouvent plus dans le dataset sont les suivants : **Cdh1**, **Cdh5**, **EgfI7**, **Hbb-bH1**, **Itga2b** (ou CD41), **Kdr** (ou Flk1), **Kit**, **Pecam1**, **Procr**. Dans chacun des tris décrits ci-dessous, a été conservé un dataset avec les 42 gènes, et un autre avec les 33 facteurs de transciption afin de permettre d'analyser de façon spécifique toutes les interactions potentielles.

1.3.2 Tri par stade embryonnaire

Afin d'étudier les relations au niveau temporel, les données ont été séparées par type cellulaire présent (PS, NP, HF, 4SFG, 4SG).

1.3.3 Tri par lignée

Enfin, une séparation des données au niveau fonctionnel a été effectué. Nous avons alors un dataset pour les gènes préférentiellement exprimés dans la lignée primitive (en bleu sur la figure 6), un autre dont les gènes sont davantage associés à l'hématopoïèse (en rouge) et le dernier pour les gènes impliqués dans la formation de l'endothélium (en rose/violet).

1.4 Reconstruction de réseaux

Les réseaux ont été établis en définissant les gènes pour noeuds et les interactions de régulation (activation ou inhibition) comme liens (les données étant des niveaux d'expression binarisés comme décrit ci-dessus). Nous avons choisi d'utiliser deux types d'algorithmes différents.

1.4.1 Algorithme PC

Une première démarche pour reconstruire les réseaux est d'appliquer l'algorithme PC[1] (Peter-Clark). Il s'agit d'un algorithme polynomial pour l'inférence de la structure de graphe. Pour cela, deux choix s'offraient à nous : utiliser le package *pcalg* disponible sur R, ou bien le Miic Web Server [3] (*Multivariate Information based Inductive Causation*), outil développé par l'équipe enseignante. La maîtrise facile de l'interface et la diversité des paramètres modifiables pour la manipulation des données ont vite orienté notre choix pour l'utilisation de ce dernier. Nous avons notamment fait usage de l'interface Cytoscape [4] accessible via le Miic Web Server. Miic a pour but de reconstruire des réseaux de causalité, non-causalité, ou bien mixte, entre les variables du dataset qui lui est soumis. Parmi les paramètres par défaut (ceux que nous avons utilisés), on peut noter que les échantillons sont considérés comme indépendants, même pour des conditions expérimentales identiques. Aussi, le réseau est reconstruit par maximum de vraisemblance normalisé (pour des analyses futures, on pourrait faire varier ce critère de complexité, notamment en utilisant la reconstruction basée sur les informations bayésiennes). Enfin, les effets de causes latentes sur les relations entre les noeuds ne sont pas mesurées par défaut mais pourrait également être un critére intéressant pour une analyse plus fine du réseau.

1.4.2 Réseau hierarchique

Mariam. Code R. Bioconductor.

3. https ://miic.curie.fr
4. http ://cytoscape.org

A 3-steps algorithm

REQUIRE Conditional independence information among all variables in V, and an ordering order(V) on the variables

1. Find a **skeleton** and **separation sets**
2. **Orient unshielded triples** in the skeleton based on the separation sets
3. **Orient** as many of the **remaining undirected edges** as possible by repeated application of rules $R1 - R3$

RETURN Best case $\rightarrow$ *DAG* (usually a *PDAG*) and separation sets

FIGURE 2. Algorithme PC (Spirtes, Glymour, Scheines (1993)

2. Résultats et discussion

2.1 Réseaux obtenus

Face à la diversité des paramètres qui peuvent être modifiés via MIIC (sur l'interface Cytoscape), nous avons choisi de ne nous focaliser que sur un unique paramètre : le seuil de confiance. Ainsi nous avons construit, pour chaque set de données généré, de 2 à 10 réseaux différents, les premiers réseaux ayant un seuil de confiance élevé (par rapport à l'étendue de celui-ci) et les derniers ayant un seuil de confiance plus bas. Le nombre de réseau obtenu est fonction du nombre de gènes compris dans le dataset. Tous les autres paramètres par défaut sont restés inchangés par souci de compréhension.

Il faut prendre en compte que plus le seuil de confiance est élevé, moins les relations sont nombreuses, donc des gènes disparaissent du réseau ainsi formé. Cependant, ces relations restantes sont d'autant plus fiables.

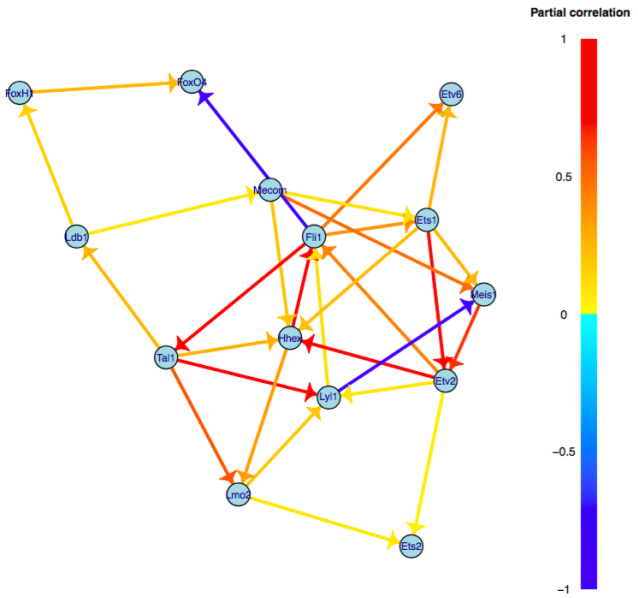

FIGURE 3. Lignée primitive

2.1.1 Tri par stade embryonnaire

Asynchrone, difficulté de recréer un réseau causal. Importance des gènes marqueurs en tant qu'inducteur d'un stade embryonnaire... transition assurée par ces gènes ? Ou si gènes présents, suffisants pour passer dans l'état embryonnaire suivant ?

2.1.2 Tri par lignée

2.2 Similarité des réseaux obtenus avec ceux de la littérature

2.3 Vérifications expérimentales

D'après les données de l'article de référence[4], on peut observer grâce à la figure 9 présente en annexe, que

.....

hub ! ! !

rôle central de Sox7 (violet).

rôle central (de controle) des cdh1 - cdh5 =¿ nécéssaires pour le changement de transition ?

différences 4SG et 4SG33genes.

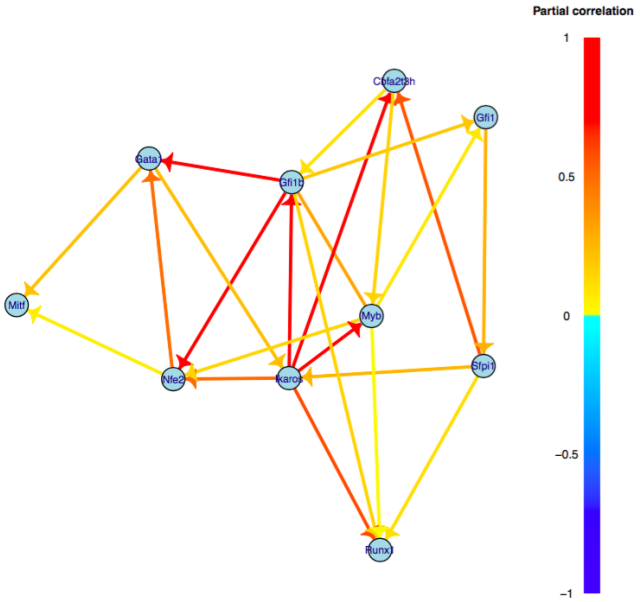

FIGURE 4. Lignée hématopïétique

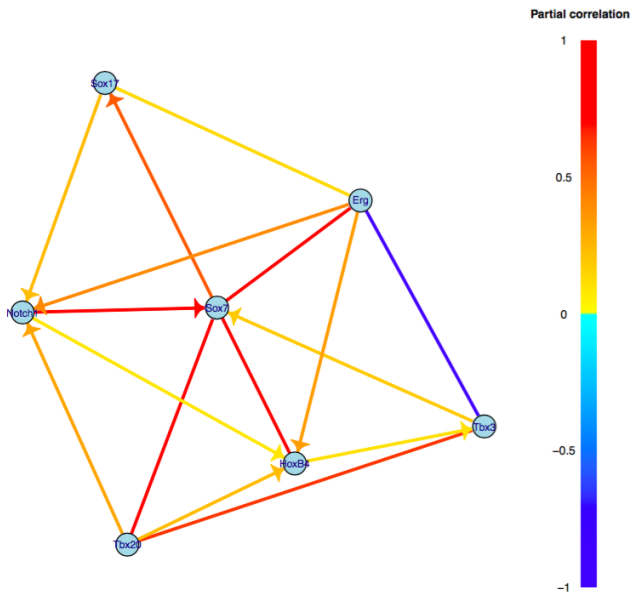

FIGURE 5. Lignée endothéliale

Conclusion

Références

[1] Spirtes and al. 5.4.1 :82, (2000).

[2] Cumano and al. Hématopoïèse intra-embryonnaire chez la souris : Emergence et caractérisation de cellules souches hématopoïétiques pendant le développement : aspects fondamentaux et cliniques. *Comptes rendus des séances de la Société de biologie et de ses filiales*, 189(4) :617–627, 1995.

[3] Sahar Messouadi Anass Es-Seddiki, Anass Ayyad and Rim Amrani. Fanconi anemia : report of a new case. *Pan Afr Med J.*, 20(92), 2015.

[4] Moignard and al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33 :269–276, 2015.

[5] Gaudin and Cumano. Les cellules souches hématopoïétiques : une double origine embryonnaire ? *Med Sci (Paris)*, 23(8-9) :681–684, 2007.

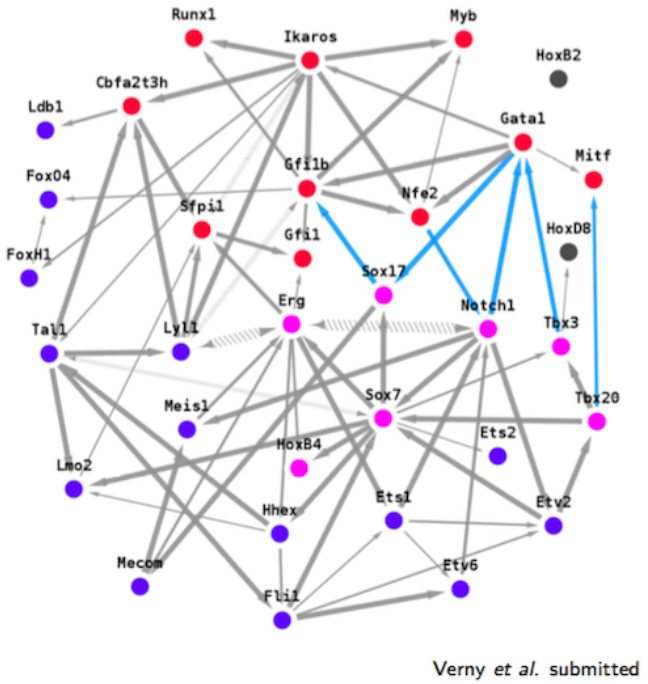

FIGURE 6. Résultats issus de *Verny et al. submitted* (les couleurs correspondent aux lignées décrites dans la figure 1).

Annexes

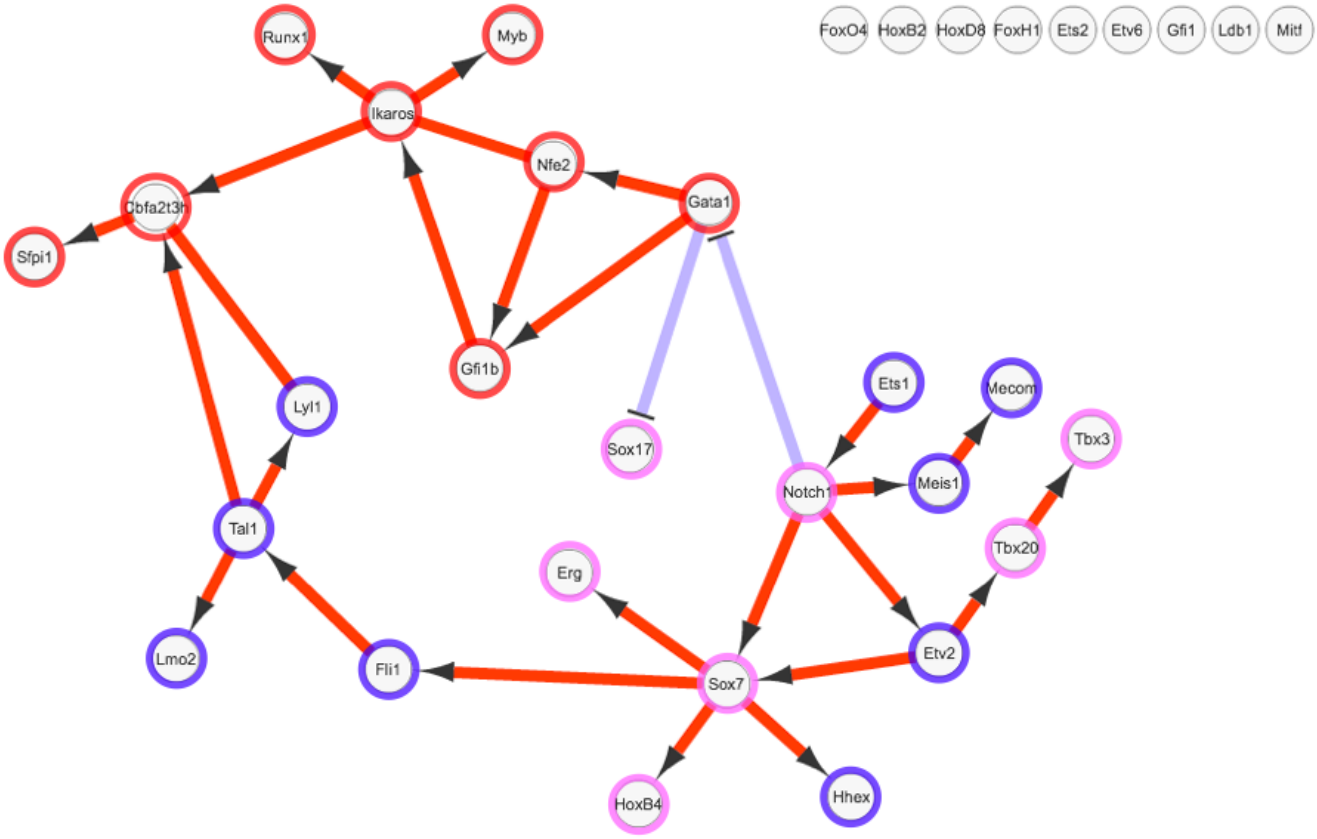

FIGURE 7. Compromis obtenu après de nombreux tests sur les facteurs de transcription uniquement (les couleurs correspondent aux lignées décrites dans la figure 1).

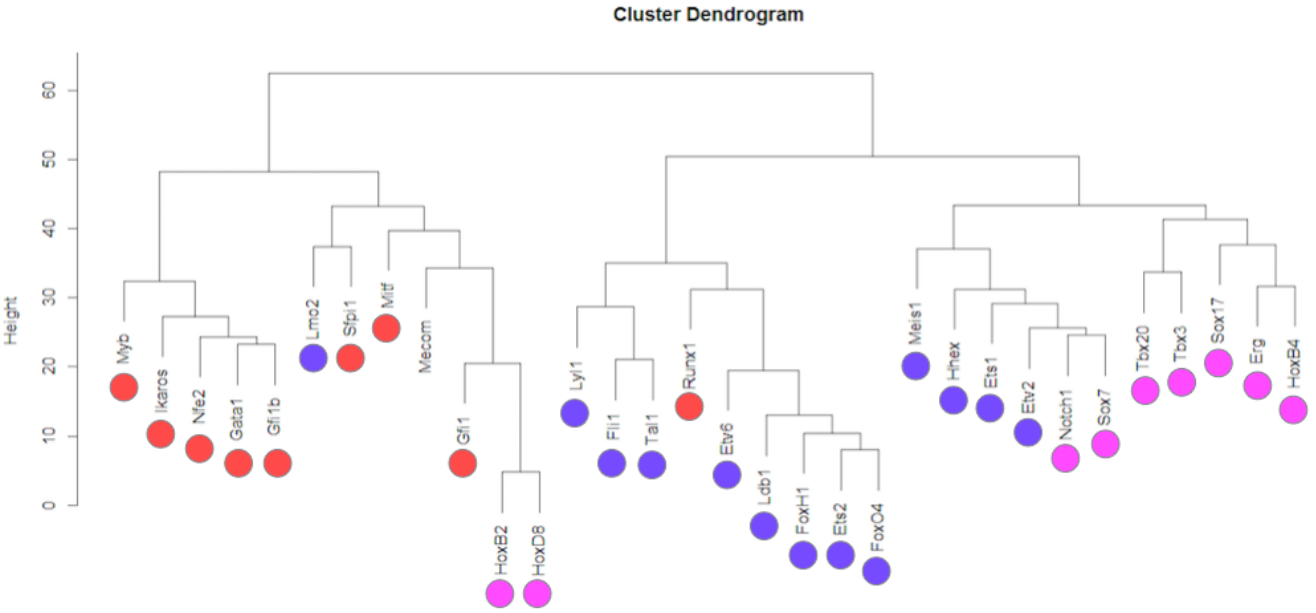

FIGURE 8. Arbre obtenu par clustering hierarchique non supervisé - 33 gènes

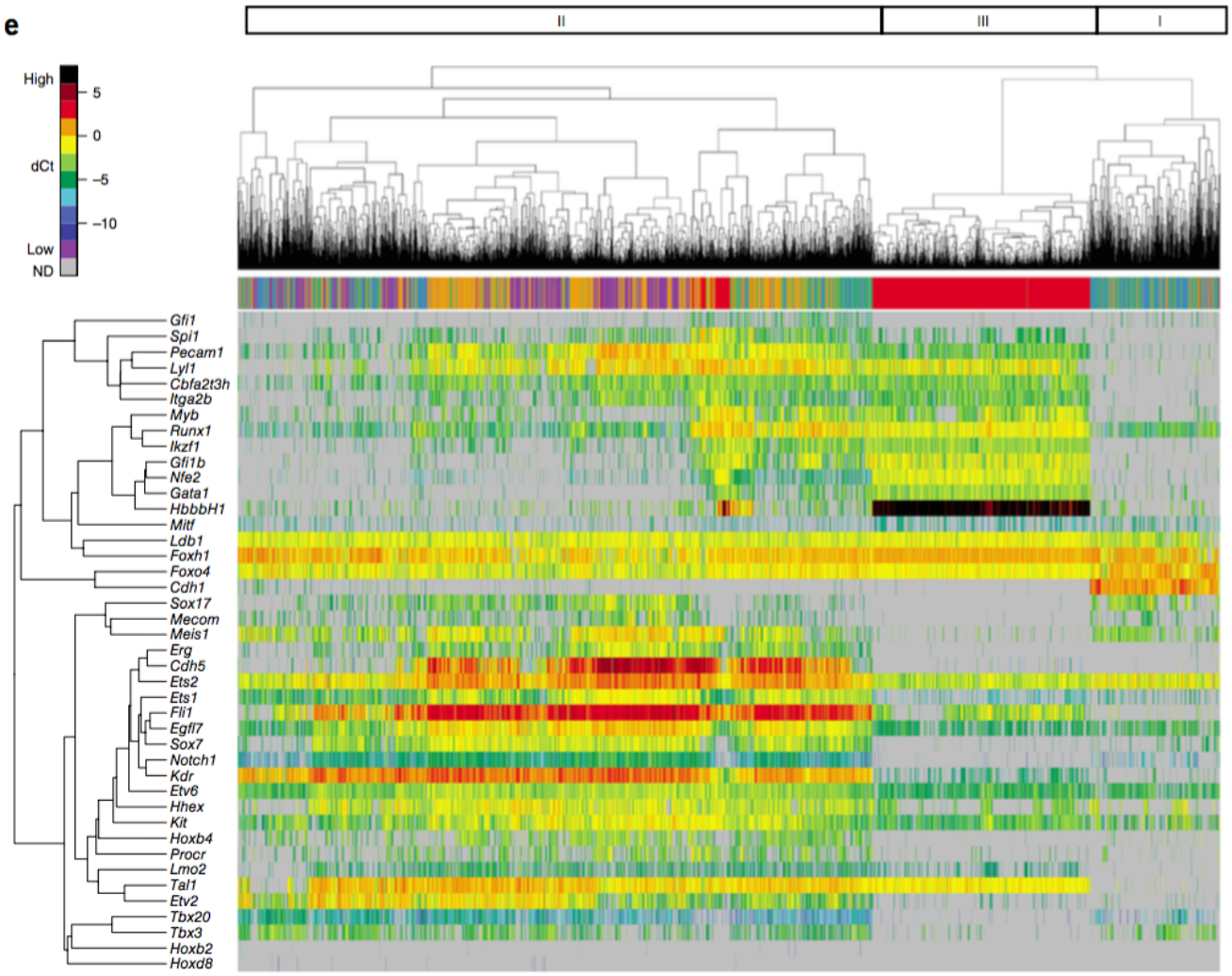

FIGURE 9. Résultats du clustering hierarchique non supervisé issus de *Moignard et al.*. On peut observer le niveau d'expression pour chaque gène dans toutes les cellules. Les colonnes représentent les cellules et les lignes les gènes. Les couleurs correspondent au stade embryonnaire d'où chaque cellule a été extraite.

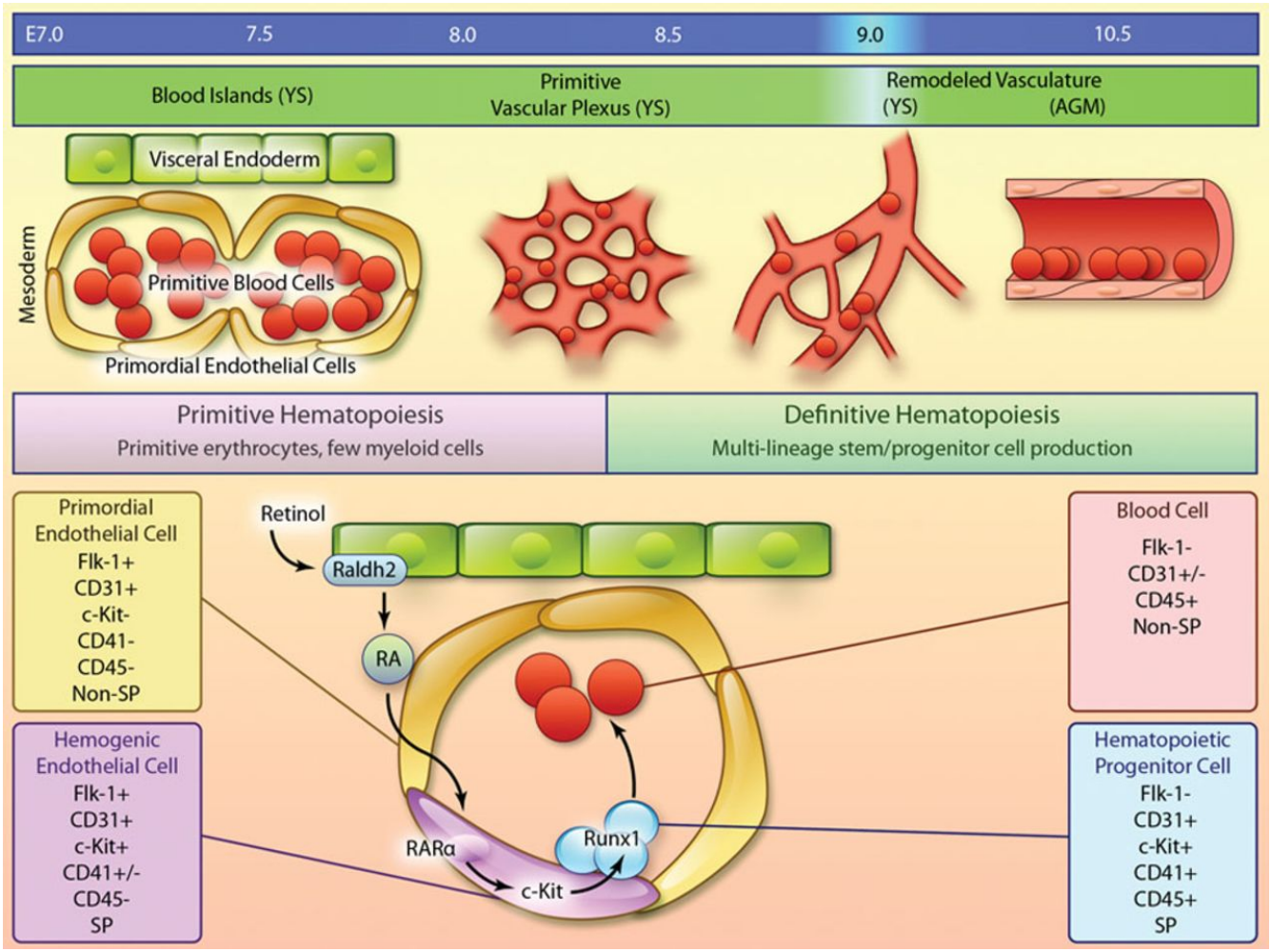

FIGURE 10. Hematopoiese

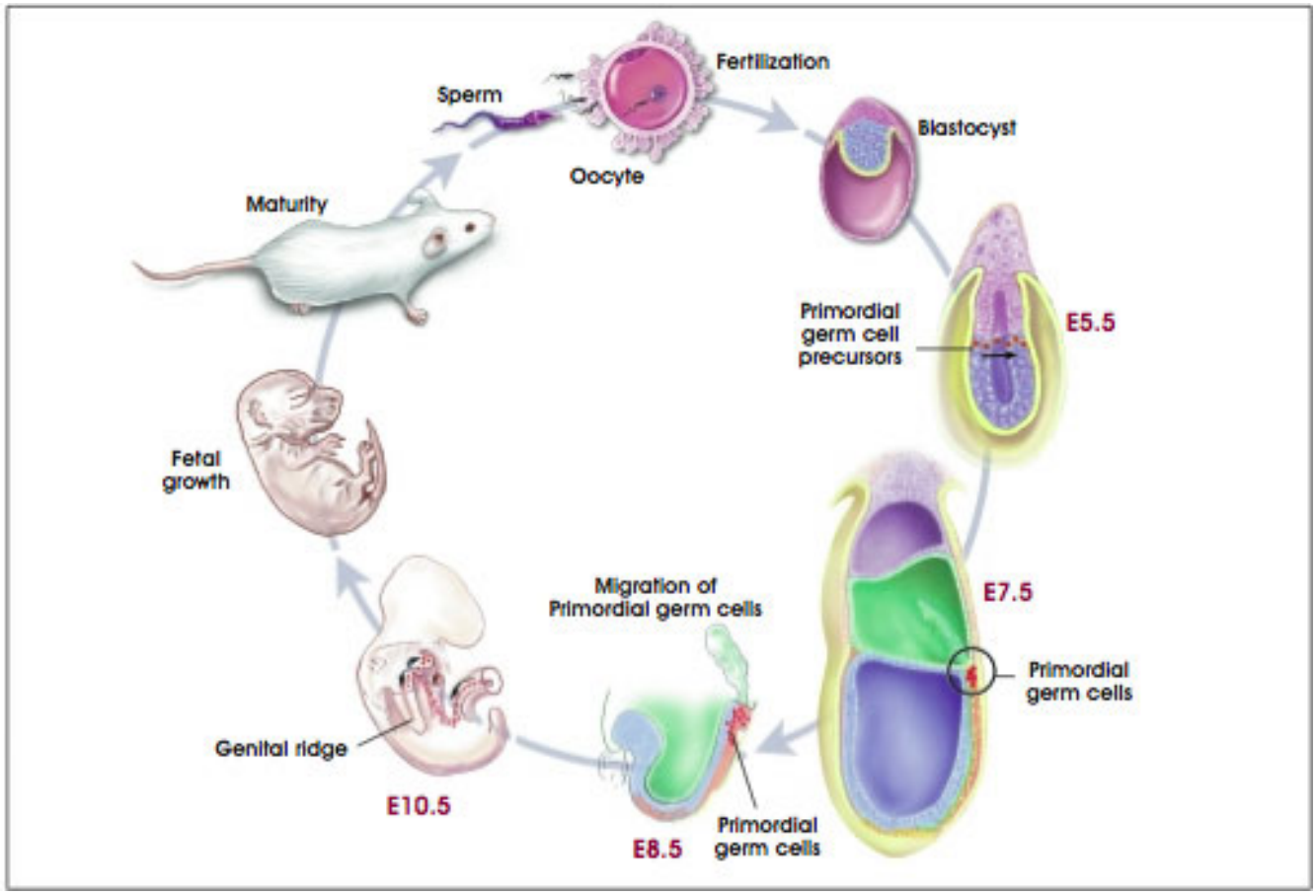

FIGURE 11. Cycle de développement de *Mus musculus*

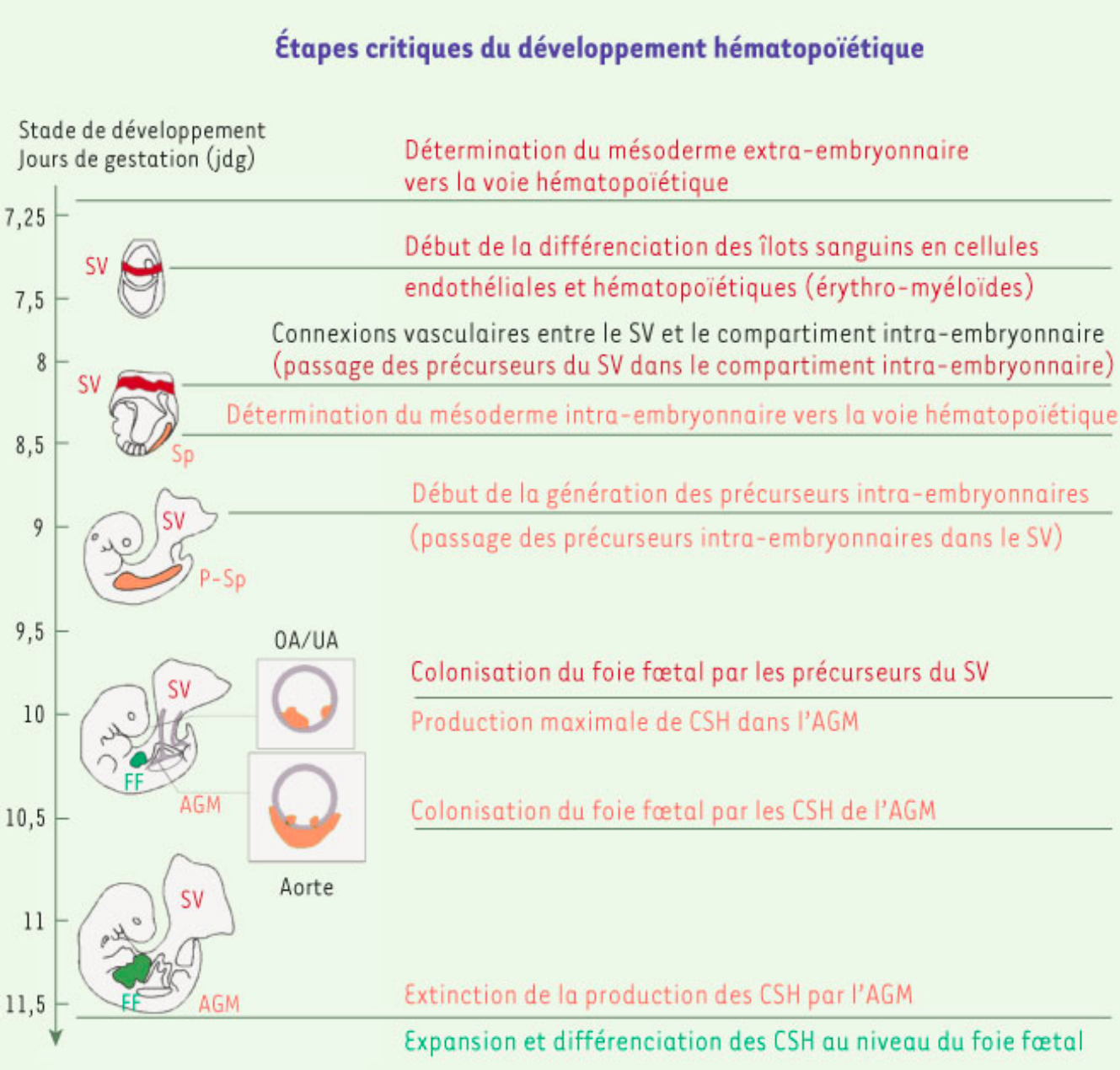

FIGURE 12. Détails du développement de *Mus musculus*[5] du stade E7 à E11.5, dans le compartiment extra-embryonnaire (en rouge) et intra-embryonnaire (en jaune). En encart figurent les sites impliqués dans la génération des CSH, c'est-à-dire l'aorte et sa partie ventrale (et les artères omphalomésentérique (OA) et ombilicale (UA)). AGM : aorte-gonades-mésonephros ; FF : foie fœtal ; P-Sp : splanchnopleure para-aortique ; Sp : splanchnopleure ; SV : sac vitellin