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 très spécifiques.

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

3 934 single cells - gene regulatory network Data : niveau d'expresssion
Noeuds : gènes
Liens : interactions de régulation (activation ou inhibition)

1.1 Obtention du dataset

Le dataset proposé rassemble les données d'expression binarisées de différents gènes pouvant soit avoir un rôle de régulation transcriptionnel, soit d'autres gènes marqueurs (*housekeepers*). Les données étant binaires, le '1' représente un gène exprimé dans la condition correspondante, le '0' un gène non exprimé.

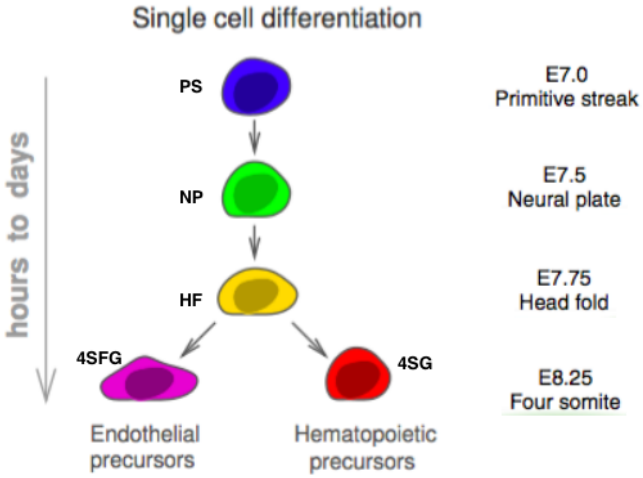

FIGURE 1. Processus de différenciation de la lignée primitive (PS) en 2 lignées distinctes : endothéliale (4SFG) et hématopoïétique (4SG)

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.

1.1.1 scRNA-Seq

Chaque type cellulaire se distingue par un profil d'expression de gènes bien spécifiques (bien que l'ADN soit le même). lorsque l'on veut étudier des phénomènes extrêmement précis ou que l'on est limité niveau quantité de matériel, les techniques traditionnelles révèlent leurs limites. Mais comme le disent les Shadoks, s'il n'y a pas de solution c'est qu'il n'y a pas de problème ! Et la solution ici tient en trois mots : Single Cell Sequencing. Cette technique apporte avec elle son lot de complications, néanmoins le single-cell sequencing permet une analyse génomique et transcriptomique à l'échelle d'une seule cellule et permet d'affiner encore et toujours plus nos connaissances. Avec ce procédé, il devient possible d'estimer la variabilité intra-tissulaire, d'étudier des stades embryonnaires précoces, de décortiquer la composition des tumeurs ou même encore de retracer les lignées cellulaires au cours du développement (source : bioinfo-fr) [1].

lmites : En effet, vu que l'on part avec une quantité réduite d'ARN (ou ADN), deux étapes d'amplification sont nécessaires (au lieu d'une seule normalement). L'amplification peut créer des biais dans vos données et il est important d'en avoir conscience, surtout quand on fait du single-cell. D'ailleurs, à cause de ces bruits générés par une grosse amplification, il va être difficile d'analyser de manière très fine l'expression de gènes faiblement exprimés.

Le second souci du single-cell, c'est la variabilité de l'expression entre les cellules, même de type identique. De prime abord, on pourrait s'attendre à avoir des profils d'expression assez similaires d'une cellule à une autre, surtout quand il s'agit du même type. Or, il faut savoir que ce n'est pas tout à fait le cas. Si vous savez que parmi vos cellules il y a plusieurs types cellulaires, mais vous ne savez pas qui est quoi, vous devrez passer par une étape de sélection d'un set de gènes que vous savez plus ou moins spécifiques à tel ou tel type cellulaire. Vous pourrez ensuite procéder à un clustering hiérarchique (et/ou ACP) de vos échantillons pour les classer en fonction de la manière dont elles expriment ces gènes. Si vous vous lancez dans ce type d'analyses sans sélectionner au préalable un set de gènes, il y a peu de chance que vous réussissiez à classer correctement vos cellules.

1.1.2 Binarisation des données

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 "housekeeper" a

1. http ://bioinfo-fr.net

été vérifiée via le site de la NCBI [2] / ou via mouse genom database :

— **Eif2b1**

— **Mrpl19**

— **Polr2a**

— **Ubc**

Dire qu'on a pensé à établir un seuil mais pas la peine.

1.3 Gènes d'intérêt

Expliquer tous les datasets formés : par lignée (fonctionnel) ou par type cellulaire (temporel).

1.4 Reconstruction de réseaux

Noeud = gène
Lien = relation entre les gènes (activation ou inhibition).

1.4.1 Algorithme PC

MIIC = Multivariate Information based Inductive Causation Interface cytoscape via Miic Web Server.
miic aims at reconstructing causal, non-causal or mixed networks between the variables of your dataset and is suitable for either categorical or quantitative discrete variables.
Default value : 1 means that all the samples are taken as independent observations given the common experimental conditions
Complexity : Complexity criterion for the network reconstruction (NML : Normalized Maximum Likelihood ; MDL / BIC : Minimum Description Length / Bayesian Information Criterion). Default value : NML
Orientation : Is the orientation of v-structures enabled ?
Propagation : Should orientation be propagated downstream of v-structures ? (see main text).
Latent : When enabled, this parameter allows to detect the effects of unobserved latent causes on the relationships between observed variables (i.e. bidirected edges, see main text).

Confidence plot : An Igraph-based plot of the network, where the color of the edges is scaled on their confidence. Correlation plot : An Igraph-based plot of the network, where edges' color is scaled on their partial correlation coefficient.

PC : Algorithme polynomial pour l'inférence de structure de graphe.

1.4.2 Réseau hierarchique

Code R.

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

- PC algorithm *(Spirtes, Glymour, Scheines, 1993)*

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

2. Résultats et discussion

Faire un choix entre précision (spécificité des données ?) et lisibilité.

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.

2.2 Similarité des réseaux obtenus avec ceux de la littérature

2.3 Vérifications expérimentales

Conclusion

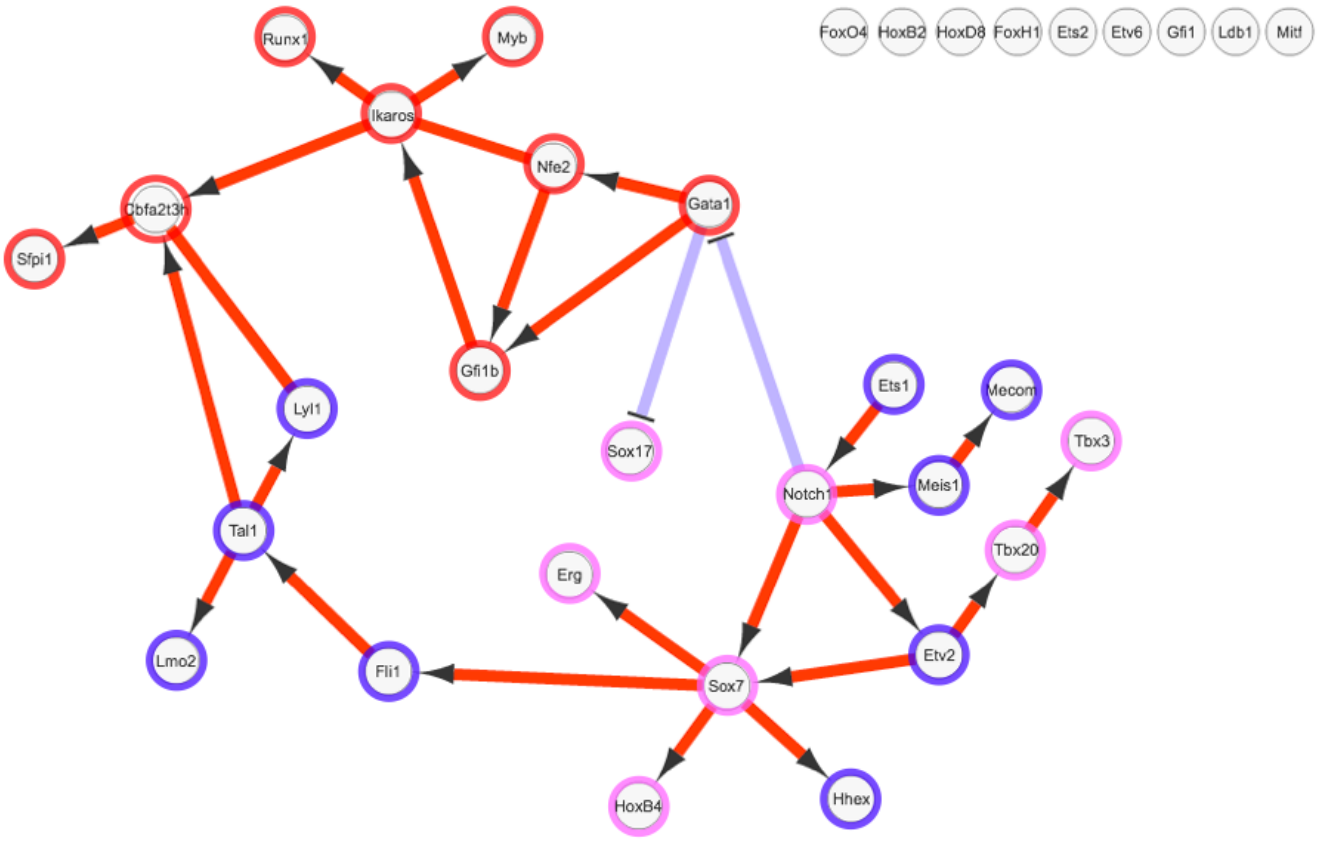

FIGURE 3. Résultats obtenus avec un seuil de confiance de ?

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.

Annexes

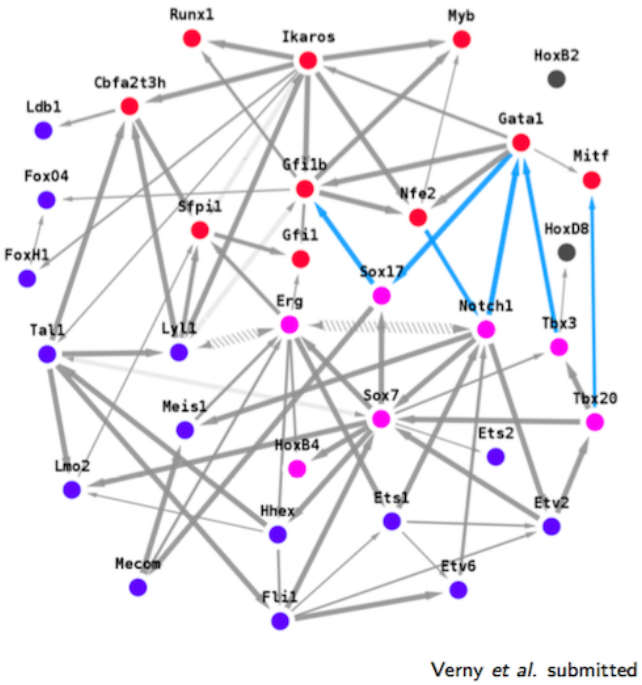

FIGURE 4. Résultats issus de *Verny et al. submitted* (dans le cours de RESYS)

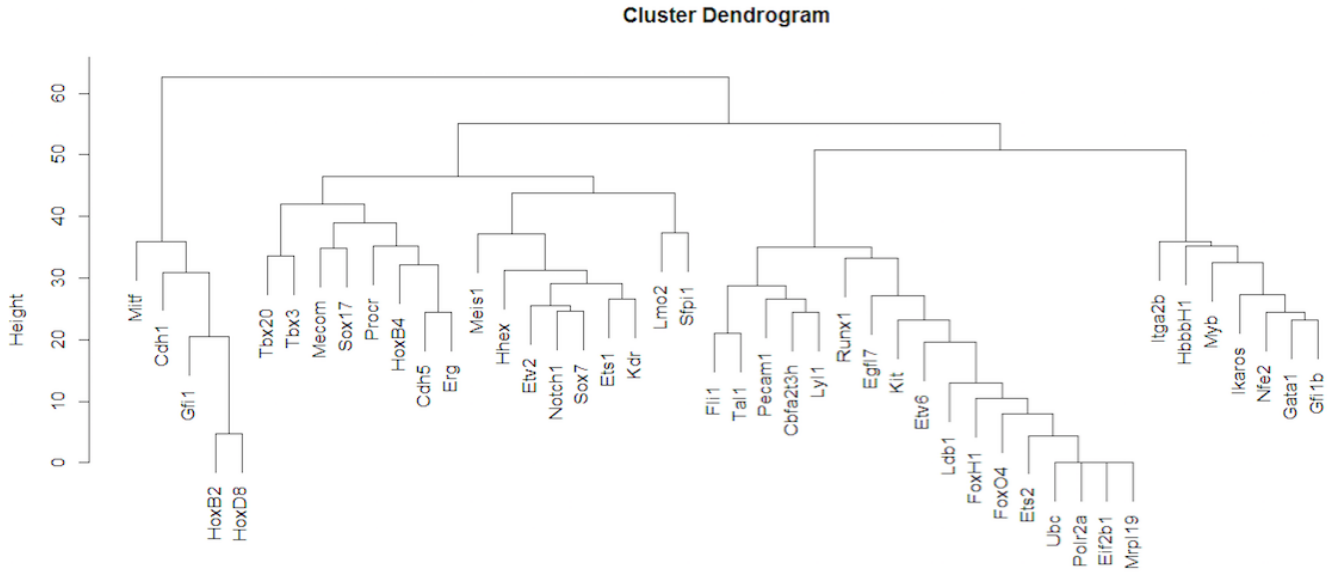

FIGURE 5. Arbre obtenu par clustering hierarchique non supervisé - 42 gènes

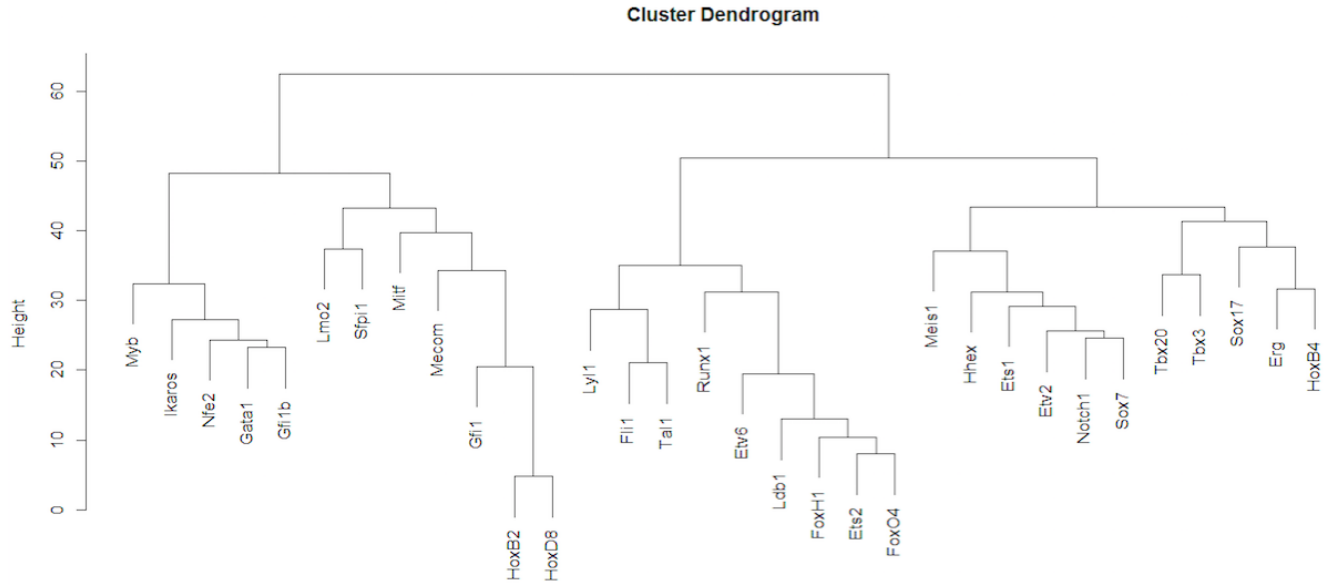

FIGURE 6. Arbre obtenu par clustering hierarchique non supervisé - 33 gènes

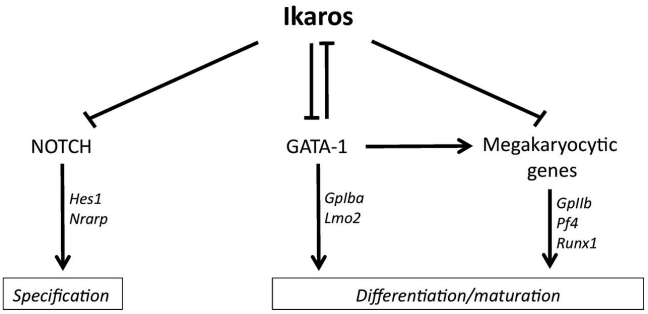

FIGURE 7. ikaros - Notch - Gata

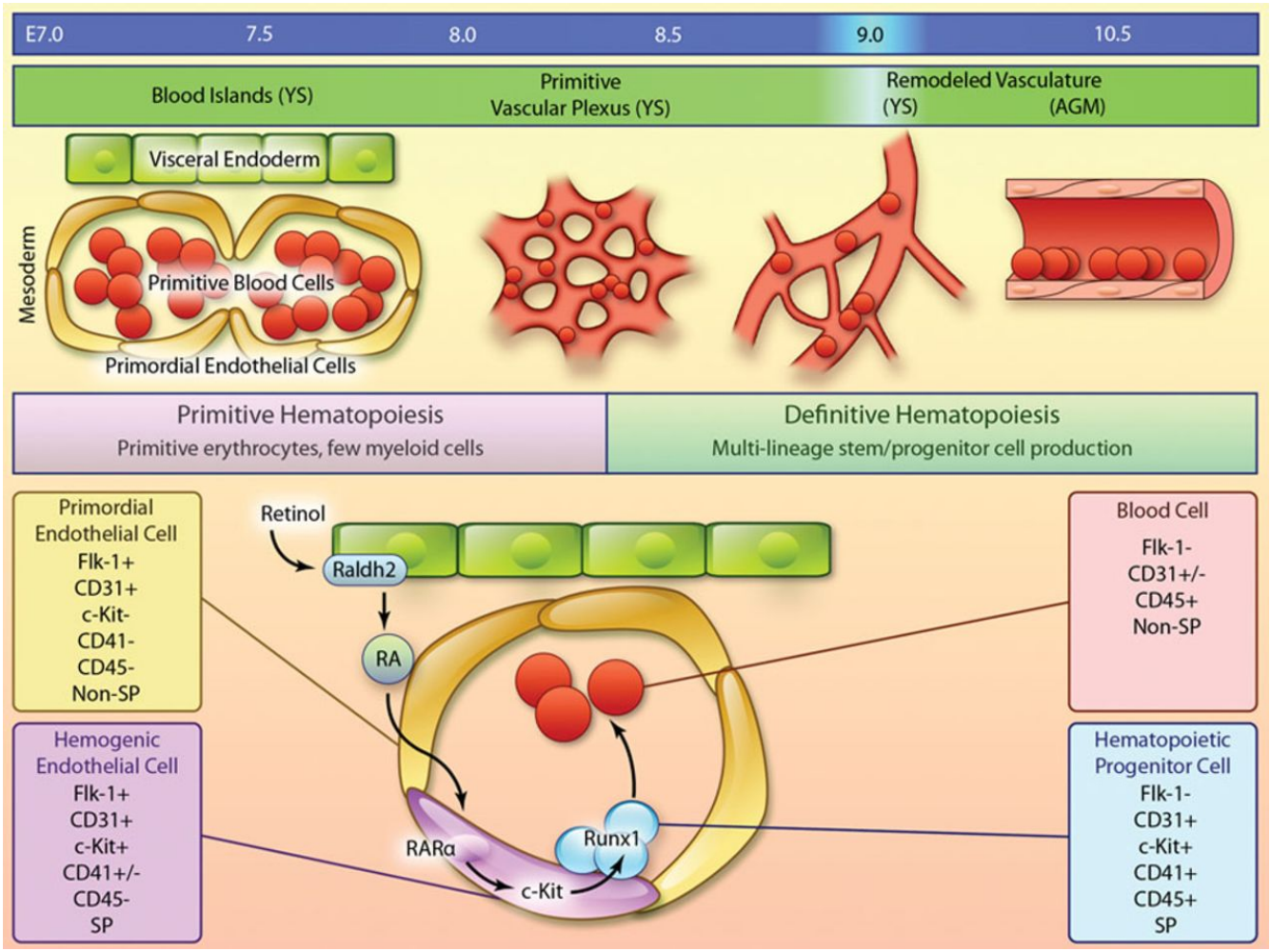

FIGURE 8. Hematopoiese

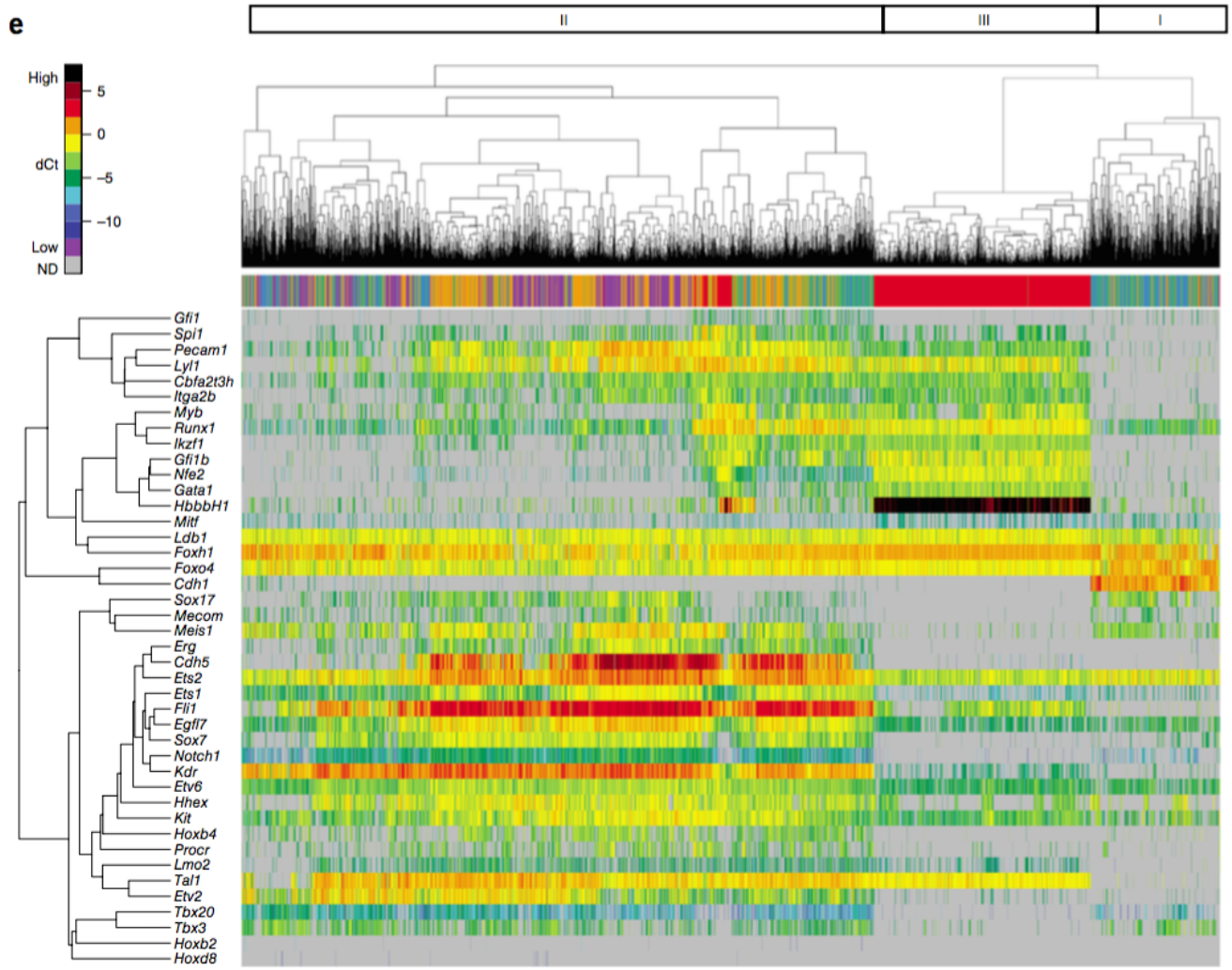

FIGURE 9. Résultats de... ? issus de *Moignard et al.*

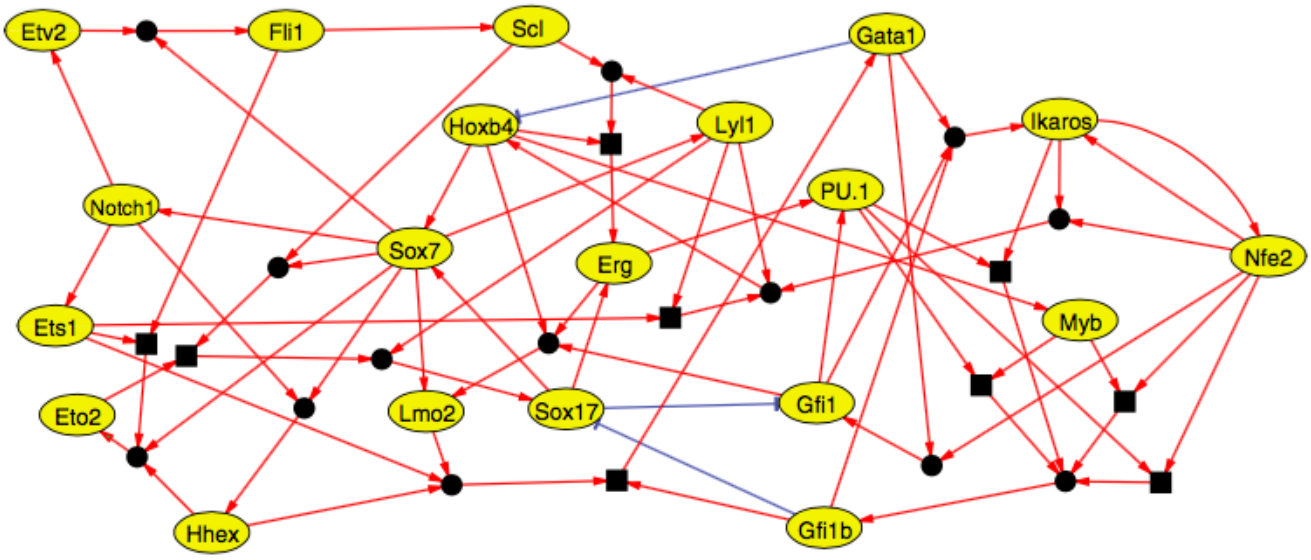

FIGURE 10. Graphe orienté issus de *Moignard et al.*

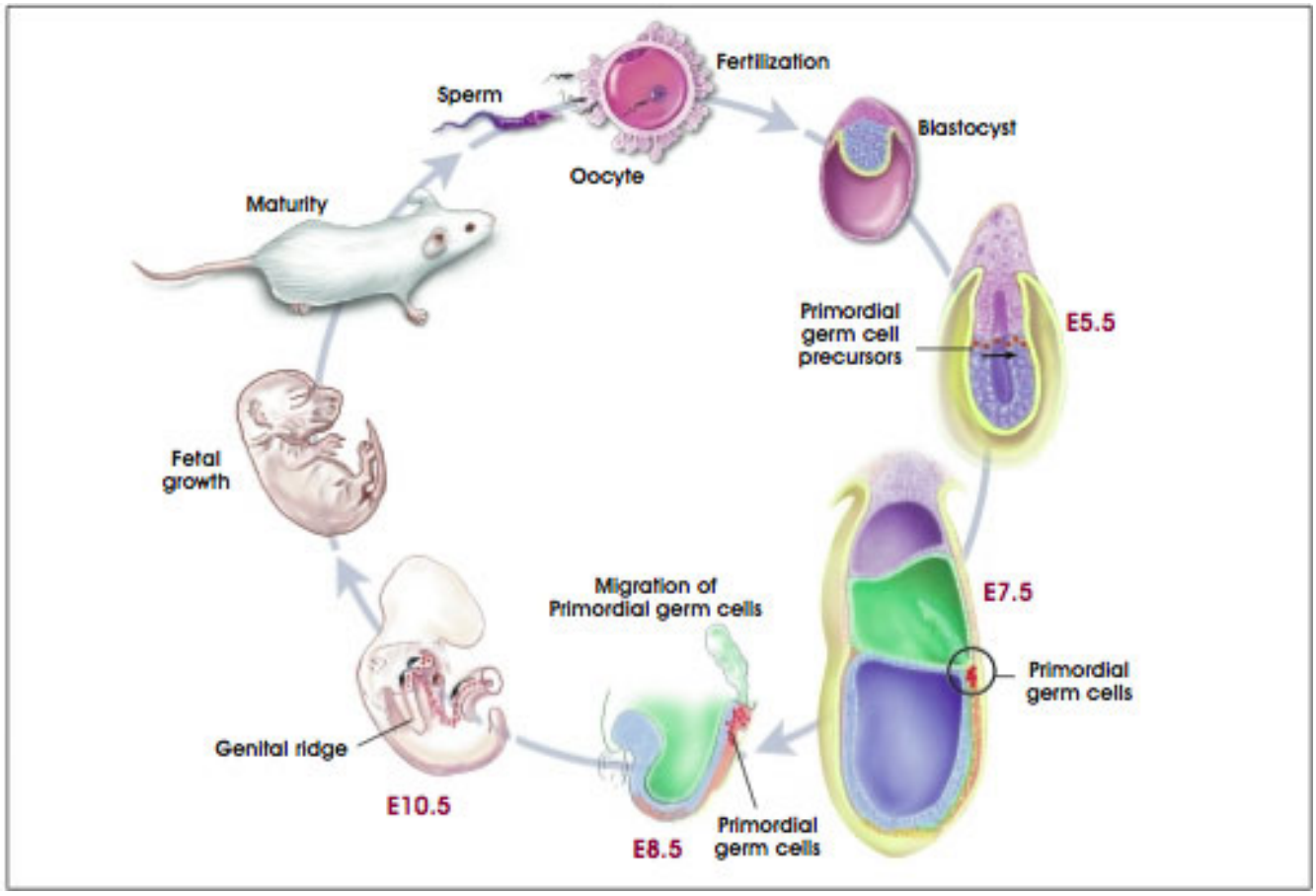

FIGURE 11. Cycle de développement de *Mus musculus*

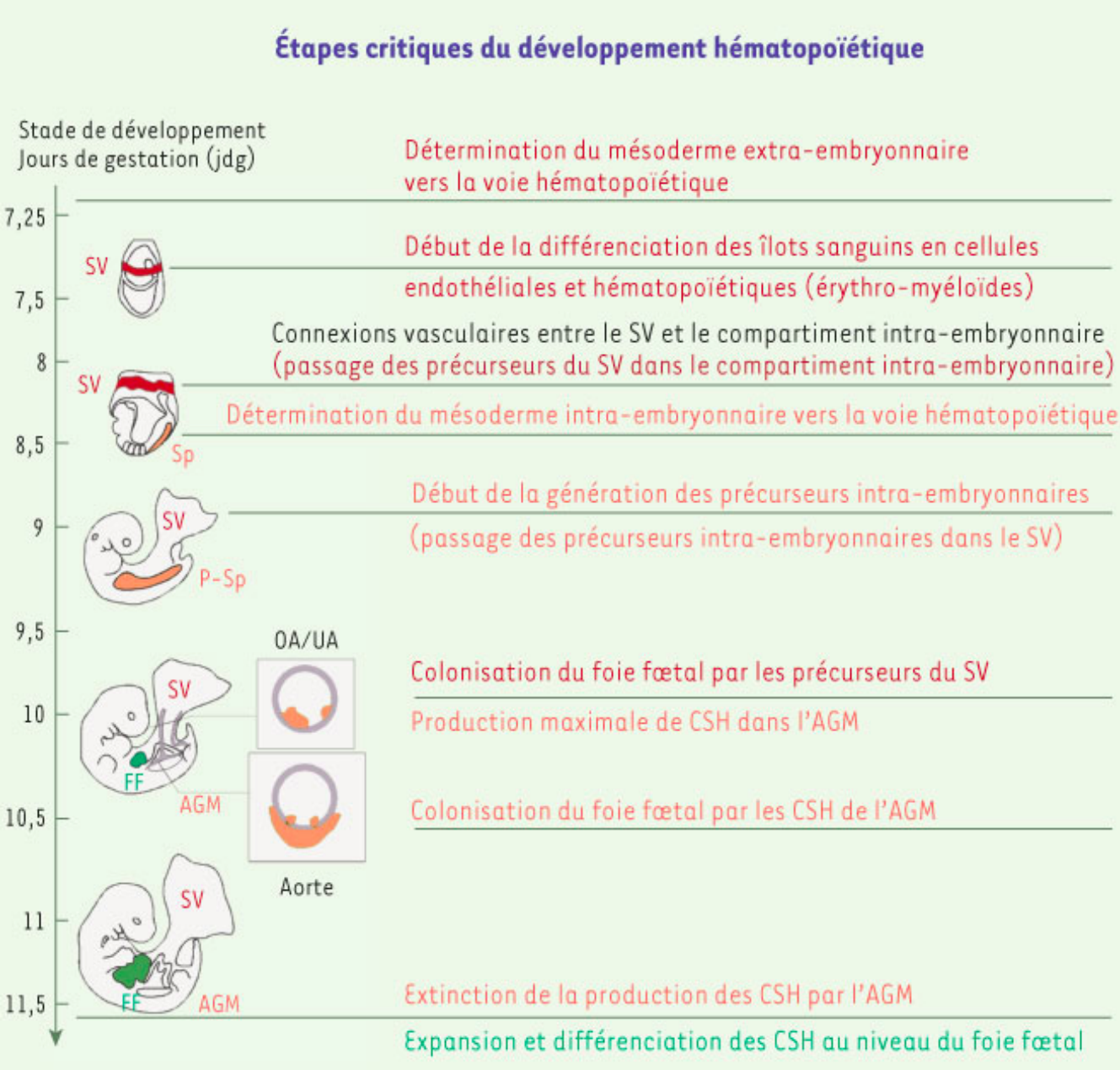

FIGURE 12. Détails du développement de *Mus musculus*[5] du stade E7 à E11.5, dans le compartiment extra-embryonnaire (en rouge) et intra-embryonnaire (en jaune). En encart figurent les sites impliqués dans la génération des CSH, c'est-à-dire l'aorte et sa partie ventrale (et les artères omphalomésentérique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