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 entravé 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.

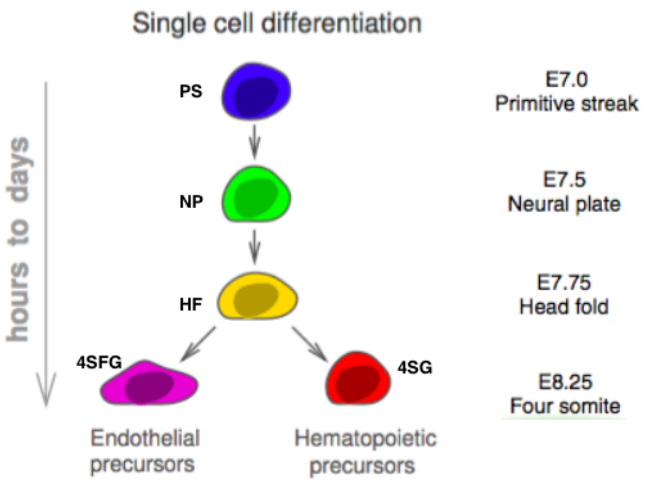

FIGURE 1. Processus de différenciation de la lignée primitive (PS) en 2 lignées distinctes : endothéliale (4SFG) et hématopoïétique (4SG)

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 **??** 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 marqueurs qui ne se trouvent plus dans le dataset sont les suivants : **Cdh1**, **Cdh5**, **Egfl7**, **Hbb-bH1**, **Itga2b**

1. http ://bioinfo-fr.net

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

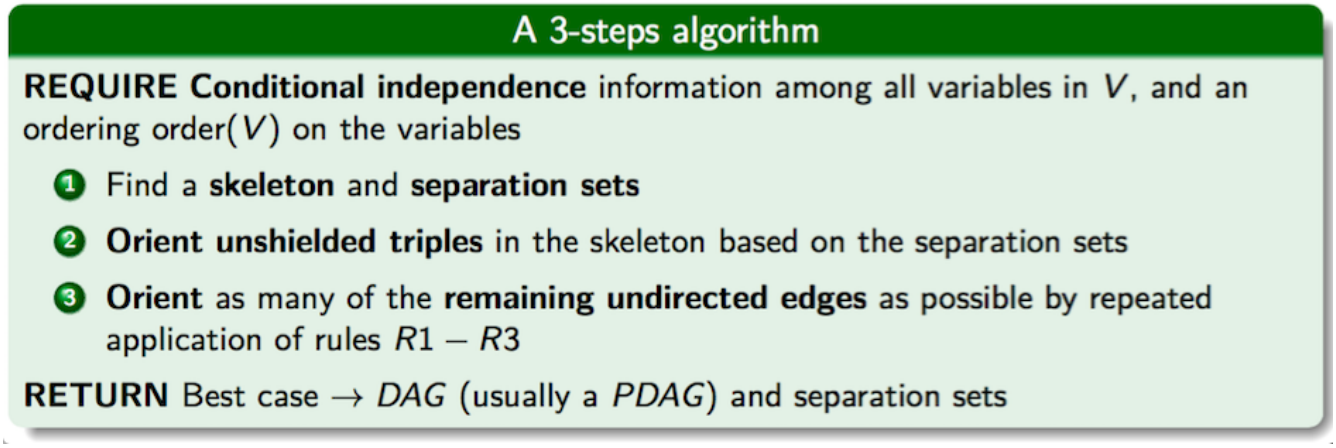

FIGURE 2. Algorithme PC (Spirtes, Glymour, Scheines (1993)

(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 dans le dataset (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

Un reseau est defini comme un ensemble de points appeles noeuds connectes entre-eux par des liens, ces derniers pouvant etre orientes ou non. On appelle degre d'un noeud le nombre de liens que celui-ci établi avec ses voisins. Dans le cas d'un reseau oriente, le degre regroupe les liens entrants et sortants. Dans notre étude, 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 l'architecture des réseaux. 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 robustesse de l'outil, 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 mixte, entre les variables du dataset qui lui est soumi. Il permet de reconstruire des graphes acycliques dirigés (DAG).

Parmi les paramètres par défaut (ceux que nous avons utilisés), on peut noter que les variables du réseaux sont considérées comme indépendantes, 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.

Les différentes étapes de l'algorithme sont détaillées figure 2. Brièvement, la première etape consiste en la reconstruction de l'architecture du réseau. Les directions des arcs appartenant aux V-structures sont ensuite déterminées. Enfin, quand cela est possible, la direction des arcs restants est calculée en tenant compte du prin-

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

cipe d'acyclicité.

1.4.2 Réseau hiérarchique

L'objectif principal des methodes de classification automatique est de repartir les elements d'un ensemble en groupes, c'est-a-dire d'etablir une partition de cet ensemble. A cette partition vient s'ajouter un critère de hierarchie de parties, qui permettent alors de former un arbre binaire, appele le dendrogramme [5]. L'algorithme de clustering hiérarchique non supervisé est disponible au travers de la fonction *hclust* [6] de R. Celle-ci prend pour *input* une matrice de distance. Nous avons donc préalablement constitué cette matrice à partir des données à analyser (distance euclidienne calculée entre chaque couple de données via la fonction *dist* de R). Le dendogramme permet de visualiser simplement les regroupements de gènes par profil d'expression, et ainsi d'établir d'éventuelles catégories fonctionnelles.

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 et de clareté. Ne sont représentés ici que les réseaux qui nous ont semblé les plus pertinents et intelligibles. Notons tout de même que pour certains réseaux, il nous a fallu faire un choix entre précision et lisibilité du réseau. C'est sur ce dernier critère que nous nous sommes concentrées afin de pouvoir obtenir des données interprétables. 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.

2.1.1 Vue d'ensemble : clustering hiérarchique

Le dendogramme obtenu avec *hclust*, à partir des 33 facteurs de transcription, est visualisable sur la figure 8. On peut remarquer que deux branches principales apparaissent, dont l'une présente une grande majorité (89%) des gènes impliqués préférentiellement dans la lignée hématopoïétique (en rouge). La seconde branche principale diverge elle-même en deux embranchements principaux. L'un forme un cluster constitué à 90% des gènes de la lignée primitive (en bleu). L'autre forme un cluster qui se réparti en deux plus petits clusters dont l'un contient uniquement des gènes précurseurs de l'endothélium (en rose/violet).

Cette répartition est donc relativement bien définie, bien que l'on retrouve des gènes de la lignée primitive dans presque tous les clusters. Le déroulement asynchrone de l'hématopoïèse dont l'article de *Moignard et al.* fait mention (à partir des données d'analyse *single cell* - figure 9) pourrait expliquer cette répartition : *"le fait d'obtenir des cellules issues de différents stades embryonnaires dans les mêmes clusters suggère que la maturation des cellules du mésoderme précoce est asynchrone, avec notamment des cellules issues de différents stades qui présentent un profil d'expression identique"*.

2.1.2 Vue d'ensemble : réseau obtenu par PC

Les réseaux obtenus sur l'ensemble des 42 gènes ou des 33 facteurs de transcription (FT) sont difficilement lisible. Le réseau obtenu a donc été choisi en faisant des compromis (via la modification du seuil de confiance), à partir du jeu de données des FT, afin d'être comparable au réseau présenté lors du cours de RESYS, figure 6. Nos résultats sont présentés figure 7. On y retrouve les mêmes *hubs* (noeuds dont le degré est élevé) : **Ikaros**, **Sox7**, **Tal1** et **Notch1**. Cependant, le *hub* formé par **Erg** a disparu dans notre réseau. Les autres noeuds présentent des degrés plus faible, ce qui tend à proposer l'hypothèse d'une répartition non aléatoire. En effet, le nombre moyen de connexions entre les noeuds suit une répartition trop étendue pour être aléatoire. La distribution de degre pourrait donc suivre une loi de puissance. Le reseau representant les relations regissant la differenciation des precurseurs hematopoietiques chez l'embryon est donc un reseau de type *scale free*[5].

2.1.3 Tri par lignée - fonctionnel

2.1.4 Tri par stade embryonnaire - temporel

2.2 Similarité des réseaux obtenus avec ceux de la littérature

5. http ://pbil.univ-lyon1.fr/R/pdf/stage7.pdf
6. https ://cran.r-project.org/web/packages/cluster/cluster.pdf

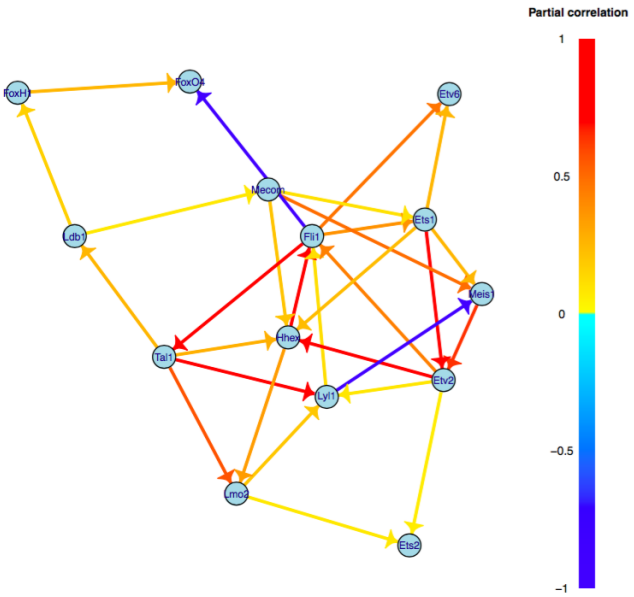

FIGURE 3. Lignée primitive

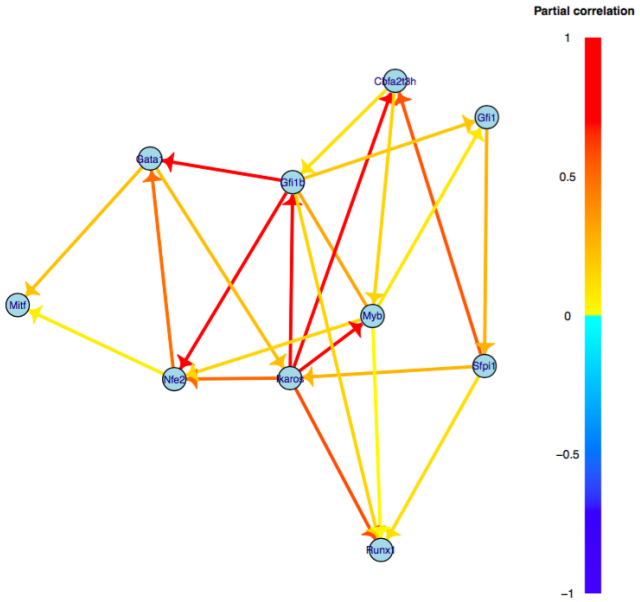

FIGURE 4. Lignée hématopoïétique

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] Oliver Hein Michael Schwind Wolfgang Konig. Scale-free networks - the impact of fat tailed degree distribution on diffusion and communication processes. *Wirtschaftsinformatik*, 48(4) :267–275, 2006.

[6] Karen K. Hirschi Kathrina L. Marcelo, Lauren C. Goldie. Regulation of endothelial cell differentiation and specification. *Circulation Research*, 112 :1272–1287, 2013.

[7] Shu Huang Ya Zhou Kohichiro Tsuji Bo Chen, Bin Mao and Feng Ma. Human embryonic stem cell-derived primitive and definitive hematopoiesis. *INTECH*, chap. 4, 2014.

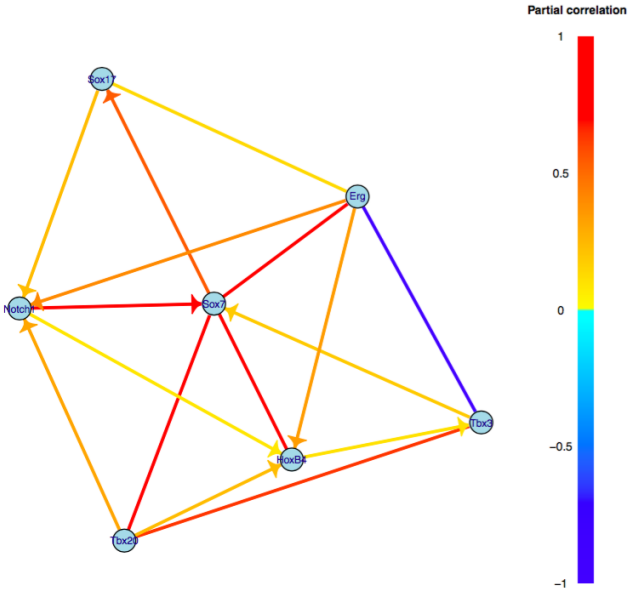

FIGURE 5. Lignée endothéliale

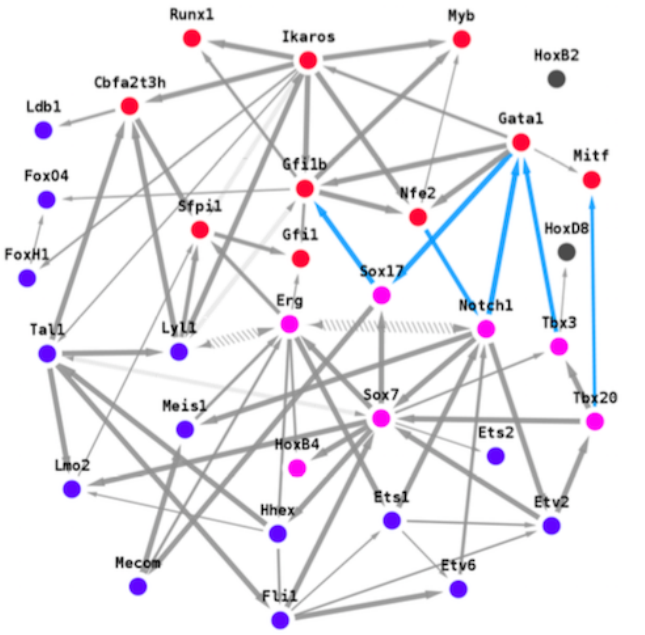

FIGURE 6. Résultats issus de *Verny et al. submitted* (les couleurs correspondent aux lignées décrites dans la figure 1).

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

Annexes

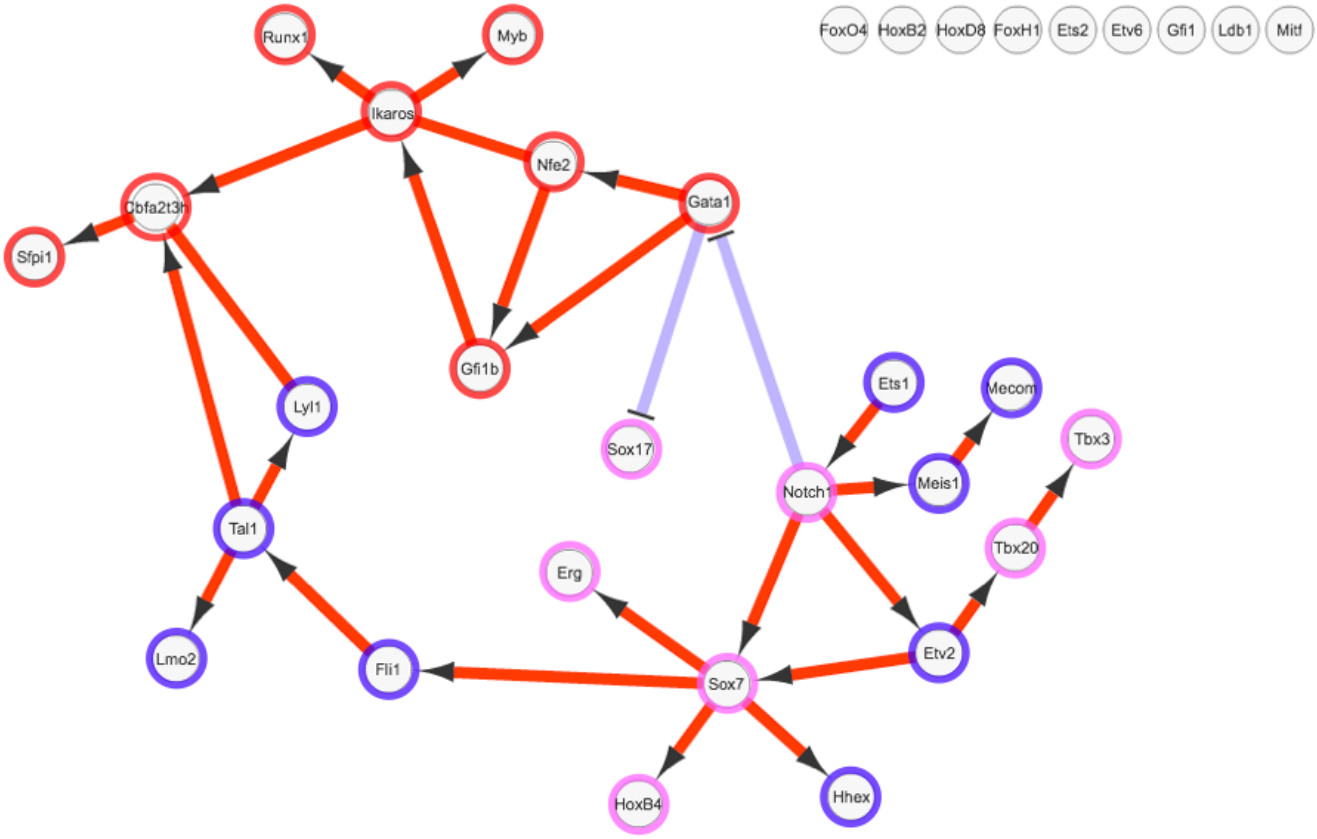

FIGURE 7. Compromis obtenu après de nombreux tests sur les facteurs de transcription uniquement (les couleurs correspondent aux lignées décrites dans la figure 1).

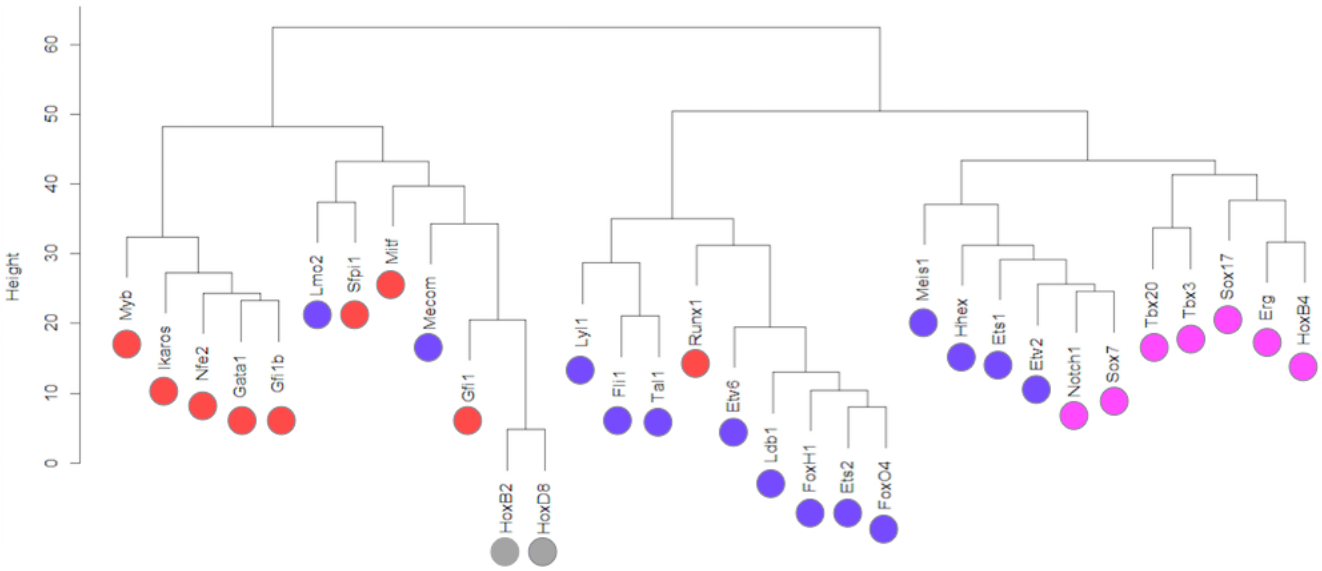

FIGURE 8. Arbre obtenu par clustering hierarchique non supervisé - 33 gènes

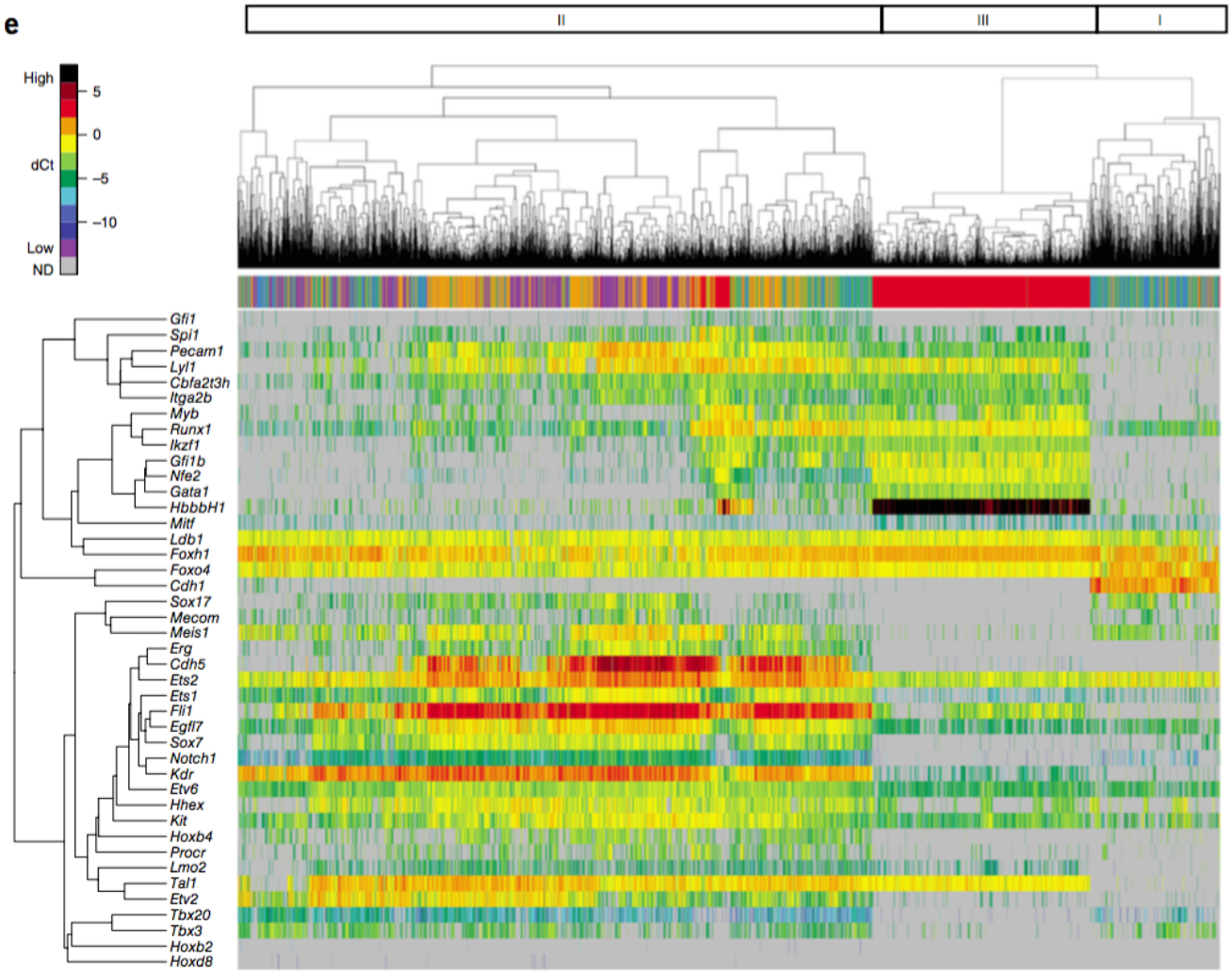

FIGURE 9. Résultats du clustering hierarchique non supervisé issus de *Moignard et al.* à partir de l'analyse par *single cell*. On peut observer le niveau d'expression pour chaque gène dans toutes les cellules. Les colonnes représentent les cellules et les lignes les gènes. Les couleurs correspondent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

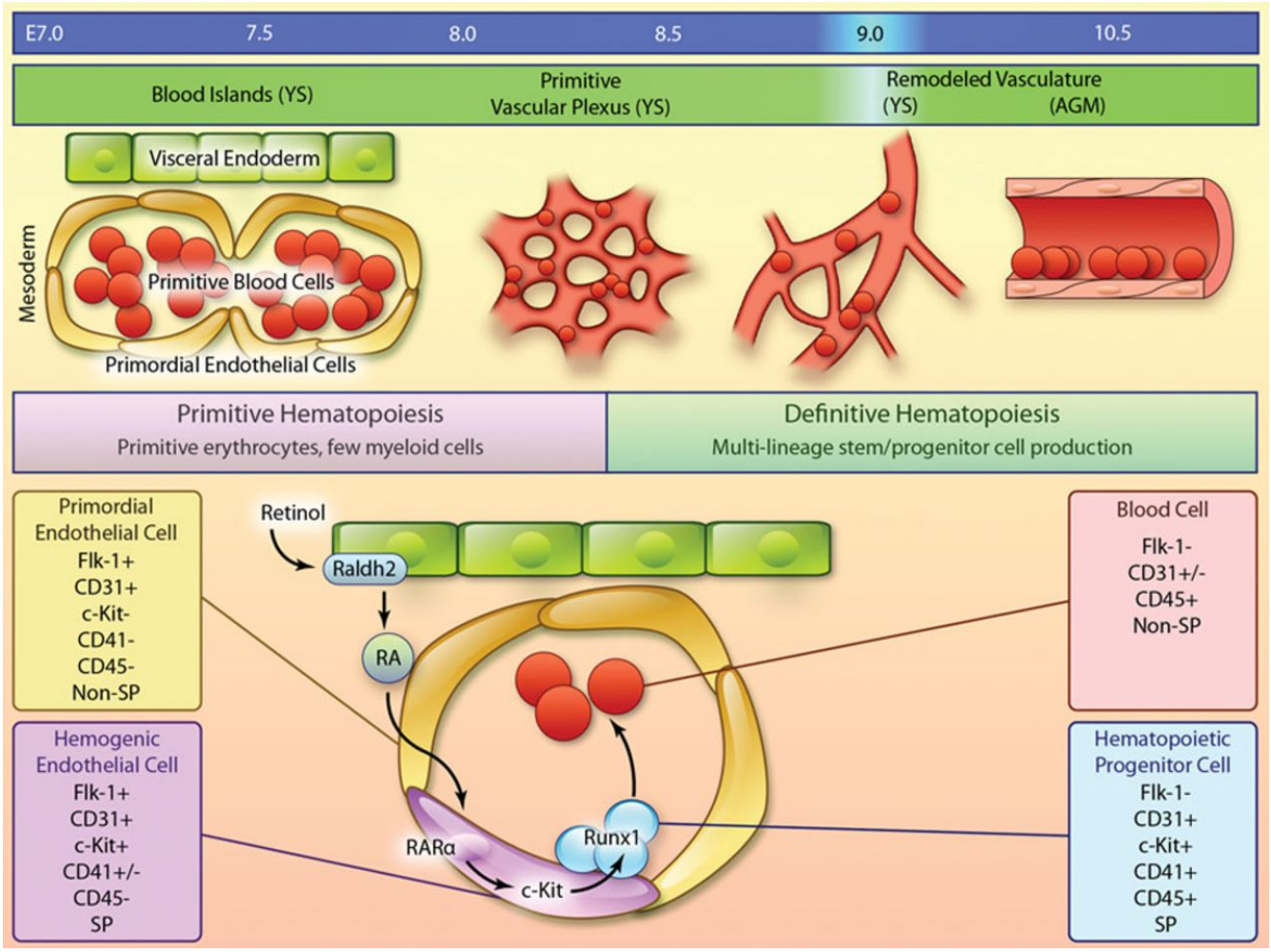

FIGURE 10. Parallèle entre l'hématopoïèse et la formation des vaisseaux sanguins à des stades du développement précoce embryonnaire (vertébrés)[6]

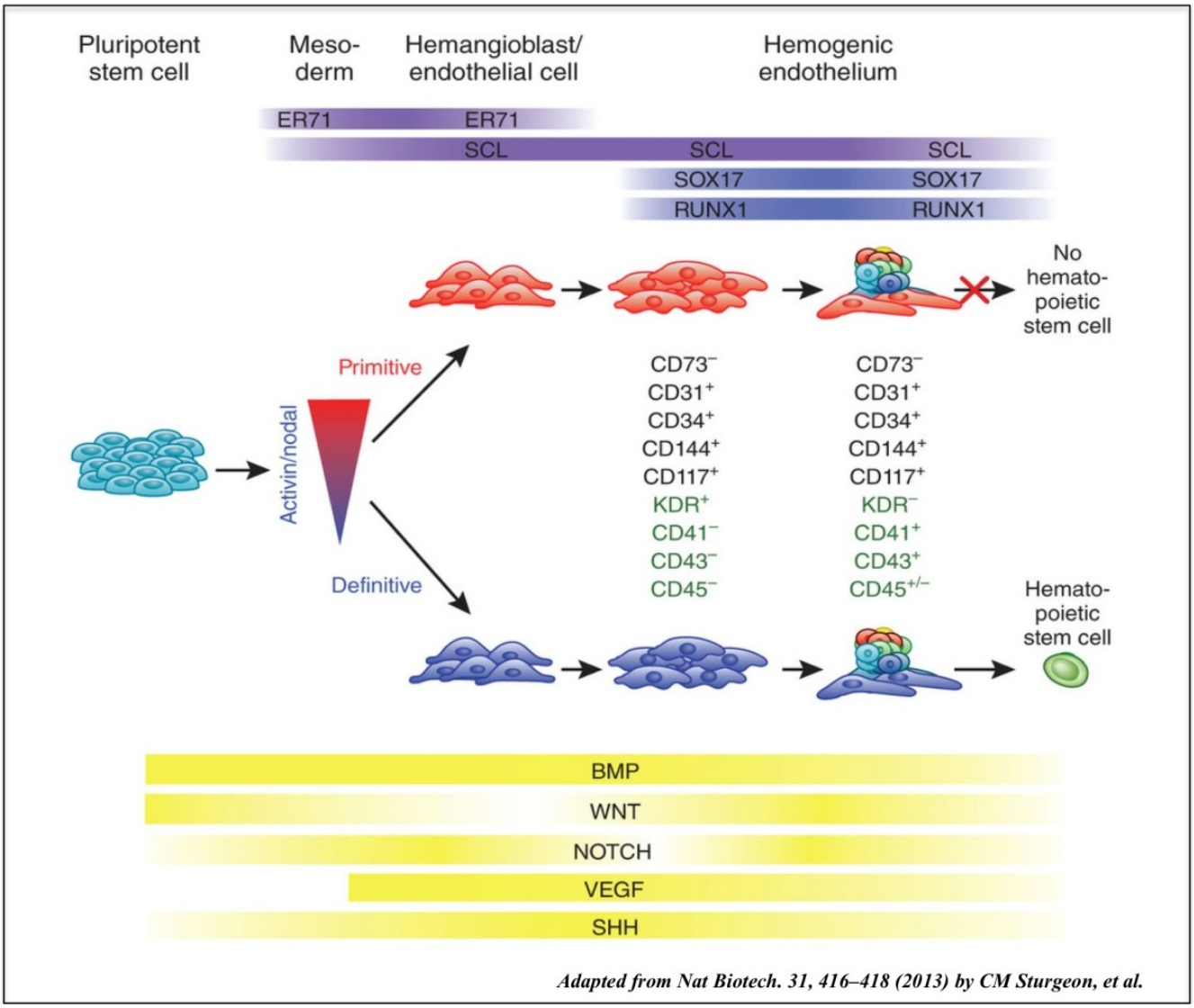

FIGURE 11. Développement de l'hémangioblaste (précurseur commun des cellules endothéliales et hématopoïétiques) et de l'épithélium hémogénique à partir de cellules pluripotentes[7]

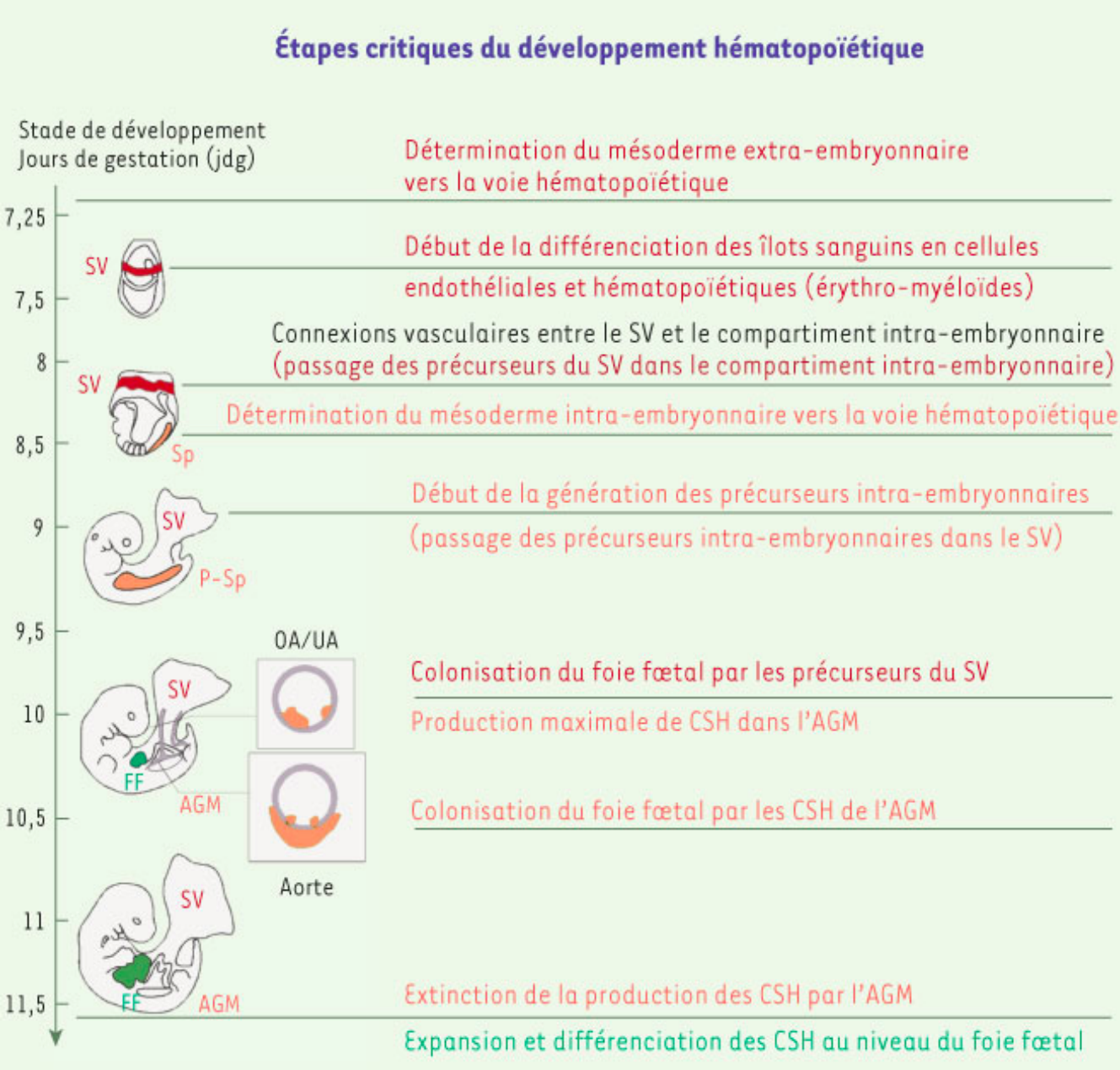

FIGURE 12. Détails du développement précoce de *Mus musculus*[8] du stade E7 à E11.5, dans le compartiment extra-embryonnaire (en rouge) et intra-embryonnaire (en jaune). En encart figurent les sites impliqués dans la génération des CSH, c'est-à-dire l'aorte et sa partie ventrale (et les artères omphalomésentérique (OA) et ombilicale (UA)). AGM : aorte-gonades-mésonephros ; FF : foie fœtal ; P-Sp : splanchnopleure para-aortique ; Sp : splanchnopleure ; SV : sac vitellin