



HAL
open science

Mémoire transcriptionnelle et plasticité moléculaire au cours de la différenciation érythrocytaire aviaire

Camille Fourneaux

► To cite this version:

Camille Fourneaux. Mémoire transcriptionnelle et plasticité moléculaire au cours de la différenciation érythrocytaire aviaire. Biologie cellulaire. Ecole normale supérieure de lyon - ENS LYON, 2022. Français. NNT : 2022ENSL0036 . tel-03920505

HAL Id: tel-03920505

<https://theses.hal.science/tel-03920505>

Submitted on 3 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Numéro National de Thèse : 2022ENSL0036

THESE

en vue de l'obtention du grade de Docteur, délivré par
l'ECOLE NORMALE SUPERIEURE DE LYON

Ecole Doctorale N° 340
Biologie Moléculaire, Intégrative et Cellulaire (BMIC)

Discipline : Sciences de la vie et de la santé

Soutenue publiquement le 25/11/2022, par :

Camille FOURNEAUX

Mémoire transcriptionnelle et plasticité moléculaire au cours de la différenciation érythrocytaire aviaire

Devant le jury composé de :

PERONNET, Frédérique	DR	CNRS	Rapporteuse
KOSMIDER, Olivier	PU-PH	Institut Cochin	Rapporteur
YVERT, Gaël	DR	CNRS	Membre
CANTINI, Laura	CR	CNRS	Membre
GANDRILLON, Olivier	DR	CNRS	Membre
GONIN-GIRAUD, Sandrine	MCF-HDR	UCBL	Directrice de thèse

REMERCIEMENTS

Il est d'usage de commencer son manuscrit de thèse par les remerciements. Beaucoup de thésards le font de bon coeur et je fais partie de cette catégorie. Merci à ma directrice de thèse Sandrine pour son accompagnement à tous les niveaux pendant ces 3+ années de thèse. Merci pour sa bienveillance, sa patience, tous ce que qu'elle m'a appris et apporté scientifiquement et aussi humainement. Merci à mon directeur d'équipe Olivier pour sa disponibilité, sa bonne humeur et cet environnement si enrichissement qui défini son équipe. Merci à toute l'équipe SBDM pour l'ambiance, l'entraide, les échanges et tous les à côtés, Catherine, Elodie, Franck, Christophe, Maxime, Elias, Hugues, Margaux, Souad et toutes les personnes de passage particulièrement Fanny, Gérard, Piwi et Blob Marley. J'ai plusieurs mentions spéciales pour chacun d'entre eux mais ça va être aussi long que cette thèse... Un énorme merci à Laurent, j'ai tellement appris à ses côtés et merci pour son indéfectible patience et disponibilité. Un très gros merci à Hélène pour tous les coups de pouces et aussi pour tous les échanges qu'on a eu. Je remercie également Sébastien pour tout son investissement sur mes projets loufocs et aussi pour son coaching pendant nos longues heures de tri et aussi à Strasbourg. Merci aux différentes équipes du LBMC pour les interactions, les échanges et la super ambiance. Et merci à l'équipe de gestion pour leur bienveillance et leur soutien! Merci aussi à Christophe Place, j'espère sincèrement que nos routes scientifiques et amicales continueront de se croiser. Merci aux copains du labo qui rendent le quotidien encore plus agréable, et surtout les copines Léa et Laura. Merci à Thibault, je n'en dirais pas plus parce que tu m'as oublié dans tes remerciements! Merci au club jeu de rôle (maître Mangeot et les compatriotes) j'espère qu'un jour on finira notre quête. Et aussi les copains pas du labo, le witches' club, Coline, Paul, Mélanie, Marjo, Cynthia, Elsa... Ainsi que Pouxpoux et Nono. Et je souhaite remercier ma grande famille pour tout leur soutien et leur enthousiasme, pour les week-ends et les vacances. Et mon grand-père Jean qui m'a donné envie de faire de la science. Last but not least, Clémence et Jérémy parce que vous êtes fiers de moi alors je le suis aussi.

Table des matières

Table des matières	3
Table des figures	7
1 Introduction	9
1.1 Les processus de décision cellulaire	9
1.1.1 Les types de décision cellulaire	9
1.1.2 Les processus de décision cellulaire d'un point de vue déterministe et d'un point de vue stochastique	11
1.2 La variabilité de l'expression génique	15
1.2.1 Les premières investigations sur la variabilité de l'expression génique	15
1.2.2 Les causes de la variabilité de l'expression génique	16
1.2.2.1 Les sources extrinsèques	16
1.2.2.2 Les sources intrinsèques	18
1.2.3 Les différents niveaux de variabilité	19
1.2.3.1 La variabilité au niveau de l'ADN	19
1.2.3.2 La variabilité au niveau de l'ARN	20
1.2.3.3 La variabilité au niveau des protéines	21
1.2.4 La Stochasticité de l'Expression Génique : nuisance ou rôle biologique?	22
1.2.4.1 Les mécanismes de réduction du bruit	22
1.2.4.2 Le variabilité comme moteur des processus de décision cellulaire	24
1.3 Les outils pour étudier la variabilité de l'expression génique au niveau transcriptionnel	28
1.3.1 Les outils expérimentaux pour étudier la variabilité de l'expression génique en cellule unique	28

1.3.1.1	La détection et la localisation des ARNm issus d'un gène	28
1.3.1.2	La détection et la localisation des ARNm issus de quelques gènes	29
1.3.1.3	Les approches single cell RT-qPCR	30
1.3.1.4	Le single-cell RNA sequencing (sc-RNA-seq)	33
1.3.2	Les outils computationnels pour étudier les données d'expression obtenues en cellules uniques	39
1.3.2.1	Les pipelines bio-informatiques	40
1.3.2.2	Les filtres qualités	41
1.3.2.3	La normalisation	42
1.3.2.4	Les outils d'analyse	44
1.3.2.5	Les métriques utilisées pour mesurer la variabilité de l'expression génique	47
1.3.3	Les outils d'inférence à partir de données d'expression issues de cellules uniques	48
1.3.3.1	Les inférences de trajectoires	48
1.3.3.2	Les inférences de réseaux de gènes	49
1.3.4	Les apports des études en sc-RNA-seq pour étudier la variabilité de l'ex- pression génique	49
1.4	L'hématopoïèse et l'érythropoïèse comme modèles de décision cellulaire pour étu- dier la variabilité de l'expression génique	51
1.4.1	L'hématopoïèse chez les mammifères	51
1.4.1.1	L'origine embryonnaire des cellules hématopoïétiques	51
1.4.1.2	La description de l'hématopoïèse adulte chez les mammifères	53
1.4.2	L'hématopoïèse des mammifères comme modèle de décision cellulaire	54
1.4.2.1	La pertinence du modèle	54
1.4.2.2	Les modèles actuels de la différenciation hématopoïétique	55
1.4.3	L'érythropoïèse comme modèle de décisions cellulaires	56
1.4.3.1	L'érythropoïèse humaine	56
1.4.3.2	Le poulet un organisme modèle	57
1.4.3.3	L'érythropoïèse chez le poulet	57
1.4.3.4	Le modèle cellulaire aviaire T2EC	58

1.4.4	La variabilité de l'expression génique lors de la différenciation érythrocytaire aviaire et la différenciation hématopoïétique humaine	59
1.5	La mémoire non génétique	62
1.5.1	La description de la mémoire non génétique	62
1.5.1.1	La mémoire au niveau de la chromatine	63
1.5.1.2	La mémoire au niveau des ARNm	65
1.5.1.3	La mémoire au niveau des protéines	66
1.5.1.4	La compréhension de la dynamique de l'expression génique des cellules dans le contexte de la multicellularité	67
1.5.2	La mémoire transcriptionnelle et mémoire de l'état cellulaire	68
1.5.2.1	Le priming	68
1.5.2.2	La plasticité cellulaire et dé-différenciation	69
1.6	Les outils pour étudier la mémoire transcriptionnelle	71
1.6.1	Le suivi des cellules apparentées	71
1.6.1.1	Les méthodes manuelles	71
1.6.1.2	Les méthodes de « tagging » génétique	71
1.6.1.3	Les approches de microfluidique	74
1.6.1.4	Les approches par FACS	76
1.6.2	Les métriques qui permettent de mesurer la proximité des transcriptomes de cellules	78
1.6.2.1	Les distances géométriques	78
1.6.2.2	Les modèles à effet mixte	80
1.7	Aperçu général de la thèse	81
2	Mémoire transcriptionnelle au cours des générations cellulaires	84
2.1	Introduction	84
2.2	Optimisations	84
2.2.1	Isolement des cellules soeurs	84
2.2.2	Isolement des cellules cousines	85
2.2.3	Bio-informatique	86
2.3	Résumé des résultats	86
2.4	Principales conclusions	88

2.5	Publication - article 1	89
3	Isolement de cellules apparentées par puce microfluidique pour analyse trans-	
	criptomique	139
3.1	Introduction	139
3.2	Optimisations	140
3.3	Résumé des résultats	142
3.4	Principales conclusions	143
3.5	Publication - article 2	143
4	Caractérisation moléculaire de la réversibilité phénotypique de progéniteurs	
	érythrocytaires induits à se différencier	167
4.1	Introduction	167
4.2	Résumé des résultats	168
4.3	Principales conclusions	169
4.4	Publication - article 3	170
5	Discussion	187
5.1	Mémoire transcriptionnelle : transmission des niveaux d'ARNm lors de la diffé-	
	renciation érythrocytaire	188
5.2	Mémoire transcriptionnelle : plasticité moléculaire lors de la différenciation éry-	
	throcytaire	193

Table des figures

1.1	<i>Schéma de la théorie de probabilités.</i>	13
1.2	<i>Schéma du paysage de Waddington.</i>	14
1.3	<i>Schéma des sources extrinsèques et intrinsèques de SEG.</i>	17
1.4	<i>Schéma des effets des taux de transcription et traduction sur la variabilité de l'expression génique.</i>	23
1.5	<i>Schéma des effets des boucles de rétro-contrôle négatif sur la variabilité de l'expression génique.</i>	24
1.6	<i>Schéma des effets de la dynamique de l'export nucléaire des ARNm sur la variabilité de l'expression génique.</i>	25
1.7	<i>Schéma du Tagging MS2.</i>	29
1.8	<i>Schéma du Single-molecule FISH.</i>	30
1.9	<i>Schéma très simplifié de la PCR digitale.</i>	31
1.10	<i>Schéma de la Single-cell RT-qPCR.</i>	32
1.11	<i>Schéma général des méthodes de single cell RNA-seq.</i>	34
1.12	<i>Schéma du barcoding des ARNm.</i>	35
1.13	<i>Schéma du séquençage Illumina.</i>	37
1.14	<i>Schéma des applications possibles en fonction des méthodes de single cell RNA-seq.</i>	38
1.15	<i>Schéma général d'un pipeline bio-informatique de pré-processing des données de sc-RNA-seq.</i>	41
1.16	<i>Schéma général et simplifié des principaux outils d'analyse des données sc-RNA-seq.</i>	45
1.17	<i>Schéma du paysage de Waddington mis à jour.</i>	50
1.18	<i>Schéma de l'origine embryonnaire des cellules souches hématopoïétiques.</i>	52
1.19	<i>Schéma simplifié de l'hématopoïèse adulte.</i>	53
1.20	<i>Schéma simplifié de l'érythropoïèse.</i>	58

1.21	<i>Mesure de l'hétérogénéité inter-cellulaire des progéniteurs érythrocytaires aviaires à l'aide de l'entropie de Shannon.</i>	60
1.22	<i>Schéma du suivi « manuel » de cellules en division.</i>	72
1.23	<i>Schéma du tagging génétique.</i>	73
1.24	<i>Schéma du suivi de cellules par puce microfluidique.</i>	75
1.25	<i>Schéma du suivi des divisions de cellules marquées par des Cell traceurs, par FACS.</i>	77
1.26	<i>Schéma des différentes distances géométriques.</i>	79
1.27	<i>Schéma de la problématique de la thèse.</i>	82
3.1	<i>Schéma de la puce microfluidique.</i>	140
5.1	<i>Schéma représentant l'effacement progressif de la mémoire transcriptionnelle au cours de la différenciation érythrocytaire.</i>	189

Chapitre 1

Introduction

1.1 Les processus de décision cellulaire

Lors du développement d'un organisme, les cellules doivent intégrer différents signaux qui leur permettent de répondre de manière adaptée aux changements de leur environnement. L'intégration des signaux et la réponse qu'elle engendre sont des processus de décision cellulaire. La coordination précise des différents processus de décision cellulaire permet le développement et l'homéostasie tissulaire, c'est-à-dire l'état d'équilibre interne de n'importe quel organisme. Il s'agit donc d'assurer le bon nombre de cellules, leur bonne position spatiale et au bon moment, ainsi que la génération de tous les types cellulaires nécessaires au bon fonctionnement de l'organisme [1].

1.1.1 Les types de décision cellulaire

Les trois grands types de décision cellulaire sont la mort-survie, la division et la différenciation. Ils sont nécessaires à la plupart des systèmes biologiques et peuvent se produire simultanément. Comprendre leur régulation est un enjeu clé de la biologie cellulaire [2].

La division cellulaire est le mode de multiplication de toutes les cellules. Chez les eucaryotes, elle se définit par la division d'une cellule mère en deux cellules filles possédant le même matériel génétique (nous ne parlerons que de la mitose et pas de la méiose où la partition du matériel génétique est différente).

Très brièvement, cette division consiste en la duplication du matériel génétique dans la cellule

mère, puis en sa ségrégation en deux qui s'accompagne de la partition du cytoplasme, des protéines qu'il contient et des membranes lipidiques et conduit à la formation de deux nouvelles cellules.

Lors du développement embryonnaire, la division cellulaire est finement régulée temporellement et spatialement pour aboutir à la formation de tous les tissus à partir d'une unique cellule [1, 3]. La division cellulaire est aussi impliquée dans l'homéostasie tissulaire chez les individus adultes en permettant le maintien d'un pool de cellules souches, par exemple au niveau de l'épithélium intestinal [4], ou encore de la niche hématopoïétique [5].

D'autre part, la division cellulaire peut être symétrique ou asymétrique en fonction de la position du fuseau mitotique lors de la mitose. Une division symétrique conduit à la génération de deux cellules filles ayant le même potentiel de choix de destin. Lors d'une division asymétrique, l'une des deux cellules filles va recevoir une plus grosse partie du cytoplasme que l'autre conduisant à la génération de cellules filles ayant des potentiels différents bien que partageant le même fond génétique. Il a d'ailleurs été proposé que les divisions asymétriques puissent être à l'origine du choix de se diviser ou de se différencier [5, 6].

La différenciation est l'engagement d'une cellule dans un processus de spécialisation par l'acquisition de caractéristiques moléculaires et phénotypiques spécifiques. Cette spécialisation donne à la cellule son identité associée à un type cellulaire.

La différenciation est évidemment cruciale pendant le développement embryonnaire, puisqu'elle permet l'apparition de tous les types cellulaires [1]. La différenciation est impliquée dans le développement des différents tissus comme par exemple le développement du cerveau [7], ou encore le développement du système hématopoïétique [8]. C'est également un type de décision cruciale pour la maintien de l'homéostasie tissulaire *via* le renouvellement cellulaire, comme dans le cas du repeuplement de l'épiderme lors du remplacement cutané ou en cas de blessure [9].

Initialement décrit comme un phénomène passif qui rendait compte de la fin inévitable de n'importe quel système biologique, la mort cellulaire programmée est aujourd'hui considérée comme un programme de décision cellulaire à part entière [10].

La mort cellulaire programmée est étroitement régulée *via* des programmes génétiques, pour certains, très bien identifiés [11]. Bien qu'il existe plusieurs types de mort cellulaire programmée,

tels que la pyroptose, la nécroptose ou encore l'autophagie, c'est majoritairement la mort par apoptose qui est au centre des études [12].

L'apoptose, en tant que décision cellulaire, a d'abord été mise en évidence chez le nématode *C. elegans*, où le nombre de cellules chez les individus adultes est invariant. En effet, pendant le développement d'individus hermaphrodites, 131 cellules somatiques sur un total de 1090 meurent systématiquement, principalement pendant l'embryogenèse [13].

La mort cellulaire programmée est également impliquée dans le développement des mammifères, par exemple pour l'élimination des neurones surnuméraires [14] ou des cellules T auto-réactives [15]. Elle est aussi importante pour le maintien de l'homéostasie tissulaire. Par exemple, lors de la régulation du nombre de cellules, en contre-balançant la division rapide des cellules de l'épithélium intestinal [16], ainsi que pour l'équilibre entre la production des cellules érythropoïétiques et l'élimination des globules rouges, *via* la voie Epo (érythropoïétine) et son récepteur EpoR dont l'activation promeut la survie des progéniteurs érythrocytaires [17].

1.1.2 Les processus de décision cellulaire d'un point de vue déterministe et d'un point de vue stochastique

Les processus de décision cellulaire ont longtemps été vus comme des processus déterministes, principalement décrits à partir de données provenant de populations de cellules [18]. De manière générale, un processus est dit déterministe si la réponse à ce processus est toujours la même : pour des conditions initiales identiques, le processus donne des résultats identiques. Appliqué à la biologie, les processus de décision cellulaire seraient alors caractérisés par la succession d'évènements discrets dont l'enchaînement serait stéréotypé. En d'autres termes, l'identité et les choix d'une cellule seraient déterminés par les niveaux d'expression et d'activité de différentes protéines, notamment les facteurs de transcription, selon une séquence prédéterminée [19, 20]. Cette définition répond parfaitement à l'observation selon laquelle rien n'est plus semblable au développement d'une grenouille que le développement d'une autre grenouille.

Cependant, très tôt des observations sont venues contredire ce modèle et démontrent le caractère probabiliste des processus de décision cellulaire chez les procaryotes et les eucaryotes.

En 1957, Novick et Weiner ont mis en évidence que dans une population homogène de bactéries *E. coli* possédant le même fond génétique et cultivées dans le même environnement, la production de la β -galactosidase était variable d'une cellule à l'autre et dépendait de l'apparition

aléatoire d'une perméase dans la cellule [21].

En 1961, Mary Lyon décrit pour la première fois que le mosaïcisme de la fourrure observé chez les souris femelles est lié à l'inactivation aléatoire d'un des chromosomes X. Elle démontre que deux cellules possédant le même génotype peuvent présenter des phénotypes différents au niveau de la couleur de la fourrure. En effet, l'expression des gènes du chromosome X peut être variable entre deux cellules puisqu'elle dépend du chromosome qui est inactivé [22]. En d'autres termes, l'inactivation du chromosome X est donc un processus stochastique.

Comprendre le choix des cellules à la résolution de la cellule unique est un changement d'échelle qui a induit de profonds changements dans les paradigmes qui décrivaient les systèmes biologiques. En effet, à l'échelle de la cellule unique, les processus de décision cellulaire ne sont pas déterministes mais sont fondamentalement stochastiques.

En outre, les conclusions des modèles déterministes des processus de décision cellulaire sont établies sur un grand nombre de cellules et représentent un comportement moyen. Ils ne peuvent pas servir à décrire le comportement individuel d'une cellule [23]. D'ailleurs, en confrontant les études réalisées à l'échelle de la cellule unique au mythe de la cellule moyenne, on réalise qu'il existe une très forte hétérogénéité entre les cellules issues d'une population isogénique.

Pour autant, le passage à l'échelle de la cellule unique ne veut pas dire que les observations et conclusions obtenues en population sont inutiles. Pour expliquer des phénomènes complexes, il est en effet nécessaire de passer par des modèles simplifiés. Ces modèles permettent de comprendre les observations et pourront être complétés et complexifiés au cours du temps, et à mesure que les technologies et connaissances évoluent. Une maxime connue reprend bien cette idée : « tous les modèles sont faux mais certains sont utiles » [24].

Un processus déterministe peut être vu comme un processus stochastique suivant une loi de probabilité de densité infinie en un point puisque le résultat est toujours le même [25] (Figure 1.1 A). Le terme « stochasticité » a d'ailleurs été emprunté au domaine des mathématiques.

Un processus stochastique est décrit comme ayant un résultat différent à un signal ou des signaux similaires. Dans un exemple hypothétique où une cellule reçoit un signal et va pouvoir mettre en place trois réponses possibles, le processus stochastique le plus simple consiste en la possibilité équiprobable pour la cellule d'effectuer une des trois réponses (Figure 1.1 B). C'est

par exemple le cas lors du choix de l'activation d'un des deux programmes génétiques chez le phage lambda que nous allons détailler plus tard [26].

Mais un processus stochastique n'est pas nécessairement équiprobable, une des réponses peut avoir une probabilité de se produire plus importante que les autres, comme avec un dé pipé (Figure 1.1 C). Par exemple dans la rétine de la drosophile, bien que le choix du sous-type de récepteurs que chaque cellule va exprimer soit aléatoire comme nous allons le voir, *in fine* 70% des cellules de la rétine exprimeront le sous-type bleu et 30% le sous-type jaune, il y a donc un biais dans le processus de décision [27].

Enfin, un processus stochastique peut être totalement ou partiellement prédit par des modèles ou des lois de probabilité en fonction de nos connaissances sur les variables qui influencent le processus [28].

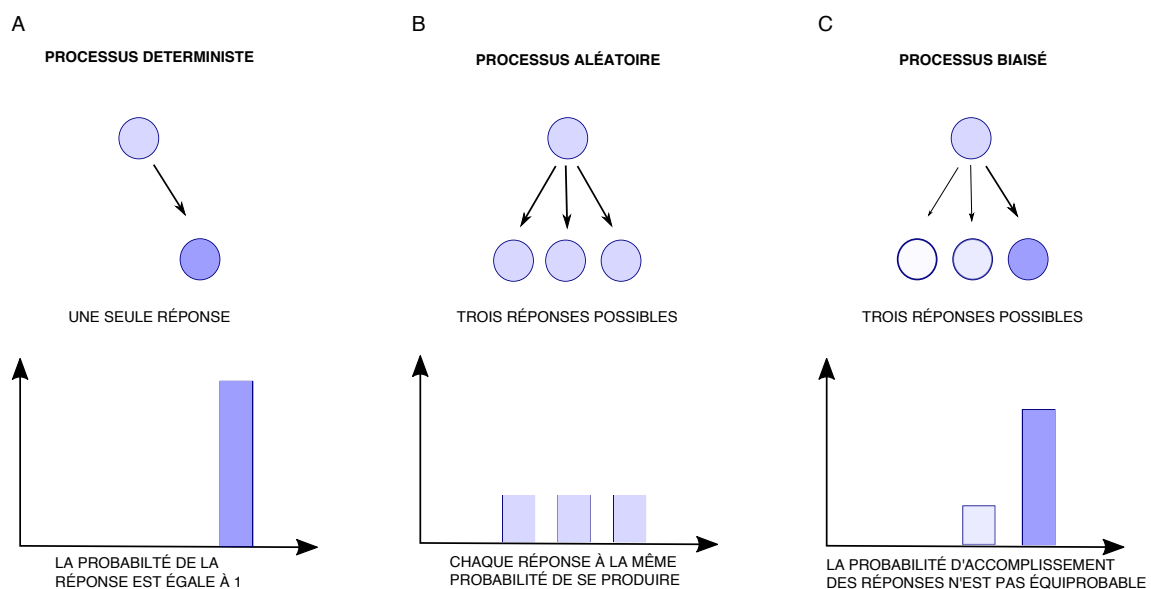


FIGURE 1.1 – Schéma de la théorie de probabilités.

(A) Un processus déterministe donne toujours la même réponse, sa probabilité de réalisation est donc égale à 1. (B) Lors d'un processus aléatoire chaque réponse a la même probabilité de se réaliser, la distribution de la probabilité est une loi uniforme. (C) Dans un processus aléatoire biaisé, certaines des réponses ont une probabilité plus importante que d'autres de se réaliser. Schéma adapté de Zechner et al. 2020 [28].

Le modèle stochastique le plus utilisé aujourd'hui est celui de Conrad Hal Waddington établi en 1957 (Figure 1.2 a). La métaphore de Waddington propose de représenter une cellule par une bille qui va se déplacer dans un paysage façonné par le réseau de gènes sous-jacent (Figure 1.2 b). La bille, démarre son trajet au sommet d'une montagne, à cet apex elle est alors multipotente, et à mesure qu'elle roule vers le bas de la montagne, son potentiel de choix décroît. Chaque

bifurcation qu'elle empruntera va limiter et restreindre les lignages qu'elle pourra suivre aux bifurcations suivantes, pour finalement devenir une cellule complètement différenciée dans un lignage/vallée spécifique tout en bas de la montagne. Ce modèle représente aisément le concept de choix aléatoire puisqu'à chaque embranchement la direction empruntée par la cellule dépendra de la topologie des pentes des vallées, topologie définie par le réseau de gènes sous-jacent : la cellule empruntera avec une probabilité plus élevée une pente abrupte qu'une pente douce [29].

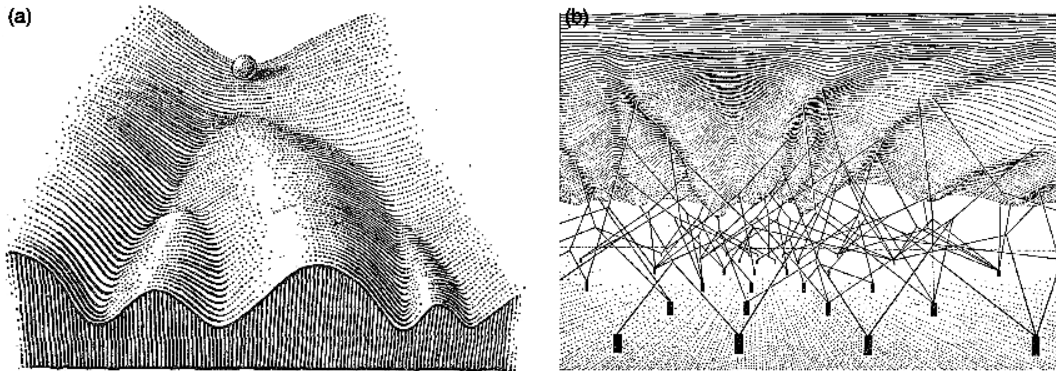


FIGURE 1.2 – Schéma du paysage de Waddington.

(a) Représentation du paysage épigénétique de Waddington. (b) Vision de l'arrière du paysage épigénétique. La forme de la pente est déterminée par la tension de plusieurs cordes interconnectées (produits génétiques en interaction) qui sont attachées à des chevilles plantées dans le sol (gènes). Issu de *Strategy of the gene de Waddington, 1957* [29].

L'apport majeur de passer d'approches en populations de cellules aux approches en cellules uniques réside dans la capacité d'observer l'hétérogénéité entre les cellules. Cette hétérogénéité est particulièrement visible au niveau de l'expression génique et est également appelée Stochasticité de l'Expression Génique (SEG).

De part la nature probabiliste et stochastique des interactions moléculaires lors de l'expression génique, il n'est pas surprenant que les processus de décision cellulaire, qui sont régis par des interactions moléculaires, suivent les mêmes lois thermodynamiques, et résultent de processus stochastiques [30, 31]. Ainsi, dans la suite de ce manuscrit, nous nous concentrerons sur la SEG.

1.2 La variabilité de l'expression génique

1.2.1 Les premières investigations sur la variabilité de l'expression génique

En 1990, Ko *et al.* ont examiné l'effet de différentes doses de glucocorticoïdes (hormones stéroïdiennes) sur l'expression d'un transgène codant pour la β -galactosidase et ont constaté que la variabilité de l'expression du transgène d'une cellule eucaryote à l'autre était particulièrement importante. De plus, l'augmentation de la dose de glucocorticoïdes entraînait une augmentation du nombre de cellules présentant un niveau élevé de l'expression de la β -galactosidase plutôt qu'une augmentation uniforme de son expression dans toutes les cellules. Autrement dit, l'effet de dose-dépendance observé est une conséquence de la modification de la probabilité qu'une cellule individuelle exprime le gène à un niveau élevé et non pas d'une augmentation globale et homogène de l'expression du gène dans toutes les cellules [32].

Il faudra ensuite attendre plusieurs années pour que les liens entre la stochasticité de l'expression génique et ses conséquences biologiques commencent à être observés et décrits. Lors de l'infection de *E. coli* par le phage lambda, deux réponses ont été observées. Dans la première, le virus se multiplie dans la cellule et entraîne l'activation du cycle lytique ce qui conduit à la destruction de la cellule. Dans la seconde, l'ADN du virus est intégré au génome bactérien, le provirus va alors être transmis à la génération cellulaire suivante au cours des divisions, c'est la lysogénie. Une étude en cellule unique a montré que le choix entre ces deux réponses est la conséquence de fluctuations stochastiques du niveau de l'expression des gènes régulateurs de ces deux processus [26].

Également à cette époque, plusieurs disciplines scientifiques se rapprochent pour aborder ces nouvelles questions et pour développer des modèles probabilistes dont le but est d'expliquer la variabilité observée. Par exemple, en modélisant l'expression stochastique d'un gène sur la base des paramètres biochimiques de son expression comme les taux de production des protéines ou encore les taux de dégradation des protéines associées à sa régulation [30].

Très tôt, le réseau de régulation des gènes sous-jacent, et les interactions qui en découlent sont envisagées comme l'origine probable de cette variabilité. L'utilisation d'approches pluridisciplinaires de biologie synthétique ont ainsi permis d'initier la caractérisation des causes de la variabilité de l'expression génique.

L'une des expériences les plus connues est le système « repressilator » d'Elowitz *et al.* repo-

sant sur la construction d'un réseau d'expression génique possédant trois noeuds de régulation [33]. Grâce à une boucle de répression, ce système d'expression génique synthétique est oscillant, c'est-à-dire activé périodiquement chez *E. coli* générant ainsi une boucle de rétro-action cyclique qualifiée d'horloge. Une protéine fluorescente est synthétisée suite à l'activation de cette voie génétique, qui est ensuite observée en microscopie à fluorescence dans chacune des cellules. Dans ce système simple, les auteurs ont observé que la fréquence d'oscillation de cette horloge artificielle est variable d'une cellule à l'autre, alors que les cellules sont cultivées dans les mêmes conditions. Les auteurs proposent que la raison la plus probable de cette variabilité est la fluctuation stochastique des composants qui régulent le système.

1.2.2 Les causes de la variabilité de l'expression génique

On sait aujourd'hui que la variabilité de l'expression génique a différentes sources. Certaines de ces sources ont été identifiées et d'autres restent à découvrir. Caractériser ces sources revient finalement à identifier quelles sont les variables qui vont influencer et/ou biaiser le processus de décision cellulaire. En reprenant la métaphore du dé, cela revient à comprendre si et comment le dé est pipé.

L'un des premiers papiers étudiant les causes de la variabilité de l'expression génique est celui d'Elowitz en 2002 [34]. Deux transgènes ont été insérés en une copie dans des loci opposés équidistants de l'origine de réplication du chromosome de *E. coli*. Le premier code pour la CFP (Cyan Fluorescent Protein) et le second code pour la YFP (Yellow Fluorescent Protein). Les deux transgènes sont sous le contrôle d'un promoteur identique. La variabilité de l'expression des deux gènes est quantifiée en utilisant comme signal les deux protéines fluorescentes synthétisées, par microscopie à fluorescence à l'échelle de chaque bactérie. Grâce à ce système, les auteurs ont pu décomposer la variabilité en deux composantes : la variabilité d'origine extrinsèque qui va affecter simultanément les deux transgènes, et la variabilité d'origine intrinsèque qui va affecter les deux transgènes indépendamment l'un de l'autre (Figure 1.3).

1.2.2.1 Les sources extrinsèques

Selon la définition d'Elowitz, les sources de variabilité d'origine extrinsèque sont les paramètres qui vont induire une corrélation de la variabilité entre les deux transgènes au sein d'une même cellule au cours du temps. En d'autres termes, la variabilité extrinsèque n'est pas spéci-

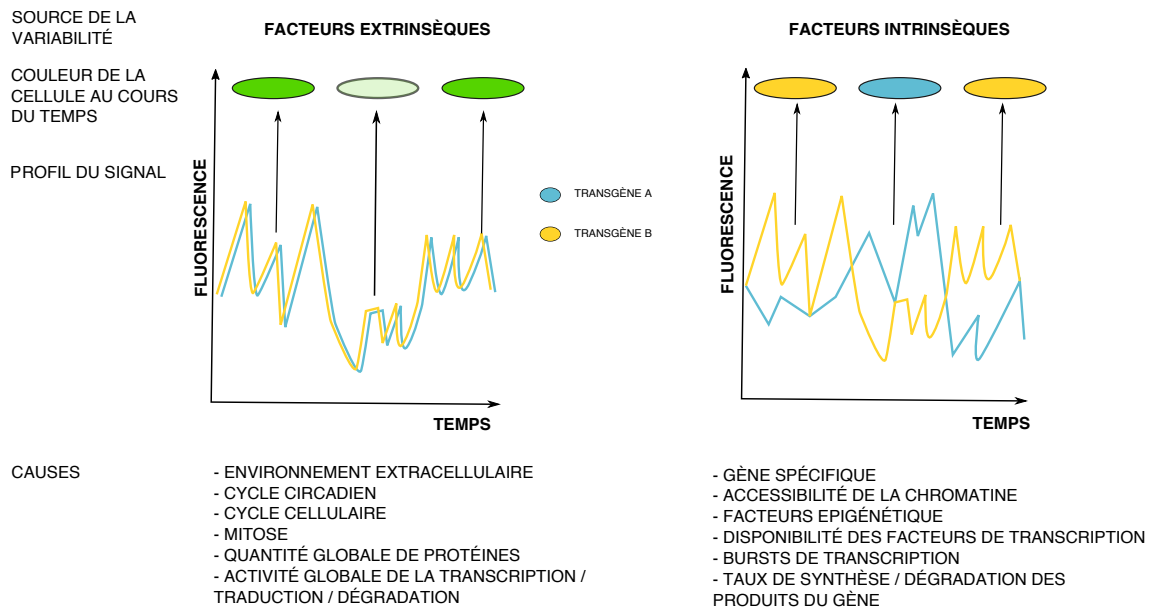


FIGURE 1.3 – Schéma des sources extrinsèques et intrinsèques de SEG.

La variabilité extrinsèque affecte simultanément les deux transgènes (transgène A en bleu, transgène B en jaune) induisant une variation corrélée entre eux du signal fluorescent. La superposition du signal des deux transgènes va entraîner une couleur verte de la bactérie, mélange des deux signaux fluorescents. La variabilité intrinsèque affecte chacun des transgènes indépendamment et donc leur variabilité au cours du temps n'est pas corrélée. La couleur de la bactérie va fluctuer du jaune au bleu en fonction du niveau d'expression de chacun des transgènes.

fique au gène étudié mais a un effet sur lui (Figure 1.3 panel gauche).

Dans les sources de variabilité extrinsèque, on retrouve :

- 1) l'environnement extracellulaire, par exemple les concentrations extracellulaires de morphogènes [35, 36];
- 2) le cycle circadien qui génère de la variabilité inter-cellules lors de l'expression des gènes cycliques [37];
- 3) le cycle cellulaire qui induit des changements dans le nombre global de protéines à un temps donné [38];
- 4) la répartition des molécules lors de la mitose qui n'est pas toujours parfaitement équitable [39] : cet effet de partition aléatoire générant d'autant plus de variabilité que le nombre de molécules à partager est petit [40];
- 5) l'activité globale des complexes permettant la transcription (les ARN polymérases), la traduction (les ribosomes) et la dégradation (les constituants du protéasome) qui sont plus au moins efficaces d'une cellule à l'autre, en fonction par exemple de l'abondance des protéines impliquées dans ces fonctions [31, 34, 41];
- 6) les dynamiques de diffusion des molécules dans les cellules qui génèrent également de la va-

riabilité extrinsèque *via* par exemple la diffusion passive [42] et la rétention nucléaire des ARNm [43].

1.2.2.2 Les sources intrinsèques

Dans le système d'Elowitz, lorsque les variations affectant les deux transgènes dans la même cellule ne sont pas corrélées au cours du temps, il s'agit d'une source intrinsèque (Figure 1.3 panel droit).

Les sources intrinsèques influencent séparément sur la variabilité de l'expression des deux gènes dans la même cellule. Si on considère un gène particulier, même si toutes les cellules d'une population sont exactement dans le même état, les réactions chimiques menant à la transcription et à la traduction du gène vont se produire à des moments différents de manière cellule-dépendante. Ces fluctuations sont déterminées localement par la séquence du gène, sa position dans le génome et les propriétés de régulation de la protéine qu'il code. En effet, les taux de transcription, traduction, dégradation, sont des facteurs intrinsèques lorsqu'ils sont observés de manière gène spécifique.

L'abondance et la disponibilité des facteurs de transcription participent de manière importante à la variabilité intrinsèque. Il y a finalement peu d'éléments cis-régulateurs de la transcription disponibles dans une cellule à un temps donné, ce qui rend l'expression génique par nature stochastique.

Enfin, l'expression d'un gène se produit de manière pulsatile, ce phénomène d'« explosion de transcription » ou burst, est décrit comme l'élément majeur favorisant le caractère stochastique de l'expression génique dans différents organismes, particulièrement dans les cellules de vertébrés [44-50].

Un burst est défini par sa durée, c'est-à-dire le temps où le promoteur est dans un état de transcription active, et par sa fréquence, c'est-à-dire le temps de bascule entre l'état actif et l'état inactif du promoteur. Plusieurs facteurs vont moduler, l'un ou ces deux paramètres et jouer sur la variabilité de l'expression d'un gène donné. Notamment l'état de la chromatine aux environs du gène qui va favoriser ou non l'accès des facteurs de transcription [51] et la présence de modificateurs de la chromatine, particulièrement les histones désacétylases [52].

D'autres mécanismes ont été proposés, et pourraient participer à induire de la variabilité dans l'expression des gènes. Par exemple, l'existence de complexes de pré-initiation qui se forment

sur la région promotrice de l'ADN et facilitent de multiples cycles d'événements de transcription par l'ARN polymérase II [53-55]; si ces complexes n'existent que pendant de courtes périodes, ils pourraient entraîner une transcription pulsatile. Également, la transcription n'a pas lieu de manière uniforme dans tout le noyau mais est concentrée dans certaines régions. Parmi ces foci de transcription, des études proposent le modèle des usines de transcription qui seraient situées au niveau des gènes actifs [56, 57]. Seul un nombre limité de ces usines (quelques centaines) serait responsable de la plupart des transcriptions d'ARNm dans la cellule; la compétition pour ces usines pourrait donc entraîner l'expression stochastique des gènes [58, 59].

Il est enfin important de noter que tout composant cellulaire subit des fluctuations dans sa propre concentration et va ensuite agir comme source de bruit extrinsèque pour les autres composants avec lesquels il interagit [60].

1.2.3 Les différents niveaux de variabilité

La variabilité résulte donc majoritairement de l'effet cumulé de la variabilité des facteurs extrinsèques et intrinsèques qui vont agir à différents niveaux de l'expression génique : l'ADN, l'ARNm et les protéines. Bien que tous les niveaux soient fortement inter-connectés et bien sûr dépendants les uns des autres, la variabilité entre ces différents niveaux n'est pas toujours corrélée. Déjà parce que chaque niveau n'a pas la même échelle, le nombre de protéines étant souvent de l'ordre de centaines de molécules, pour l'ARNm de l'ordre de dizaines de molécules, et pour les gènes eux-mêmes étant dans la majorité des cas présents en deux copies ou une seulement par cellule. Mais aussi parce que la synthèse et la dégradation d'un ARNm ou d'une protéine ne sont pas régulés de la même manière. Il est possible d'étudier la variabilité à chacun de ces trois niveaux, c'est-à-dire à l'échelle de l'ADN, des ARNm ou des protéines.

1.2.3.1 La variabilité au niveau de l'ADN

La variabilité de l'expression génique au niveau de l'ADN concerne, par exemple, l'expression de gènes de manière allèle spécifique. Dans ce cas, sur les deux chromosomes, un seul allèle du gène d'intérêt sera exprimé et l'autre allèle restera silencieux. C'est le cas, comme décrit plus haut, du mosaïcisme de la couleur de la fourrure des souris femelles [22], ou encore de la diversification des photorécepteurs dans les yeux de la drosophile [61]. Ce processus de « silencing » est stochastique, et va générer un phénotype différent d'une cellule à l'autre.

Un autre exemple particulièrement intéressant de variabilité de l'expression génique au niveau de l'ADN est le réarrangement génétique aléatoire des gènes codant pour les immunoglobulines dans les lymphocytes B. Les immunoglobulines sont composées de 3 parties : les parties V et J (sur la partie variable de la chaîne légère des immunoglobulines), et la partie D (sur la chaîne lourde). Au niveau génétique, plusieurs exons redondants sont présents dans le génome et ce pour les trois parties. C'est la sélection aléatoire d'un seul exon par partie qui rend possible la grande diversité des immunoglobulines. De plus des bases nucléotidiques peuvent être insérées aléatoirement lors du processus de recombinaison somatique ce qui génère encore plus de diversité en modifiant par exemple le cadre de lecture [62].

L'environnement chromosomique joue aussi un rôle dans la variabilité. Une étude utilisant des cellules HeLa, qui possèdent naturellement deux chromosomes 19 intacts et un chromosome 19 transloqué sur deux autres chromosomes (t6;19 et t13;19), a permis de suivre l'expression de 20 gènes portés par le chromosome 19 avec une technique dérivée du RNA FISH (iceFISH) et de comparer leurs expressions dans ces 3 différentes situations chromosomiques (wild type et les deux translocations). Les auteurs ont montré que dans le cas de la translocation t13;19 la plupart des 20 gènes étaient jusqu'à 5 fois plus transcrits que sur les chromosomes normaux, alors que dans la translocation t6;19, les gènes avaient le même taux de transcription que sur les chromosomes normaux. Cette étude montre que la position des gènes dans le génome peut influencer les taux de transcription probablement par des mécanismes de régulation au niveau des chromosomes favorisant un accès à la chromatine ou possiblement *via* la localisation des usines de transcription comme décrit plus haut [47].

1.2.3.2 La variabilité au niveau de l'ARN

Les variations inter-cellules du nombre de molécules d'ARNm ont d'abord été observées dans des populations homogènes de bactéries. Dans une étude de 2005, Golding *et al.* ont quantifié le nombre d'ARNm dans une cellule de *E. coli* en utilisant la technique de fusion MS2-GFP qui sera détaillée plus tard dans la thèse. Cette méthode permet de tagger les transcrits néosynthétisés, qui sont ensuite quantifiables par microscopie à fluorescence. Les auteurs ont montré que la synthèse d'ARNm se produit par burst, correspondant à une courte période de transcription active et une longue période sans transcription. Comme précisé ci-dessus, ces événements de bursts ont lieu de manière stochastique et sont l'une des sources principales de fluctuation du nombre de molécules d'ARNm pour un gène donné entre des cellules [63].

Raj *et al.* ont montré en 2006 que les ARNm sont également synthétisés par burst dans les cellules de mammifères. En utilisant du single molecule RNA FISH (smRNA FISH), les auteurs ont observé qu'en moyenne, un gène entraînait la production de 40 molécules d'ARNm par cellule. Cependant la variance est très importante (soit 1600). Cette observation démontre que l'expression génique chez les mammifères ne suit pas une distribution de Poisson comme attendue si les ARNm étaient produits à des taux constants car la variance serait alors égale à la moyenne. Les auteurs ont aussi montré que la quantité de facteurs de transcription disponible influence la durée des bursts [49] et d'autres études ont montré que ce paramètre influence également la fréquence des bursts [64, 65].

Enfin, la variation de l'expression génique est particulièrement visible au niveau ARN parce qu'il existe un phénomène d'amplification de la variabilité des ARNm par les cellules. Hansen *et al.* ont quantifié par smFISH en cellule unique différents ARNm endogènes. Dans 85% des cas, la variabilité intercellulaire de l'expression d'un messager est accrue après l'export de l'ARNm du noyau vers le cytoplasme. C'est-à-dire que le nombre de molécules d'ARNm pour un gène donné est encore plus variable dans le cytoplasme que dans le noyau entre deux cellules. Les auteurs ont montré à l'aide de modèles mathématiques que l'augmentation du niveau de la variabilité des quantités d'ARNm dans le cytoplasme par rapport au noyau est principalement liée aux variations des taux de dégradation et de traduction des ARNm dans le cytoplasme, ainsi que lorsque le taux d'export nucléaire est supérieur au taux de dégradation cytoplasmique des ARNm [66].

1.2.3.3 La variabilité au niveau des protéines

Les niveaux de protéines sont également variables, notamment parce qu'en conséquence de la synthèse pulsatile des ARNm elles sont également synthétisées par burst dans les cellules [67]. Mais d'autres facteurs contribuent à générer de la variabilité au niveau protéique, notamment l'efficacité de la machinerie traductionnelle. Dans une étude de 2002, les auteurs ont intégré un transgène GFP sous contrôle d'un promoteur inductible à l'IPTG dans le chromosome de *B.subtilis* et ont utilisé différentes concentrations d'IPTG pour faire varier l'efficacité de transcription. Afin de mesurer l'efficacité de la traduction, ils ont produit différentes souches de bactéries avec des mutations soit au niveau du site de liaison du ribosome soit dans le codon d'initiation du transgène. Ces altérations affectent le rendement de traduction dans les cellules. Les résultats de ces travaux ont montré que l'efficacité de la machinerie traductionnelle était

ainsi la source prédominante de variabilité observée au niveau de la quantité des protéines [41].

Cependant, les niveaux de protéines sont moins variables que les niveaux d'ARNm. Cette différence est due à plusieurs éléments. Premièrement, dans une cellule il y a un facteur 10 entre le nombre de molécules d'ARNm et celui des protéines. Or, sur un plus grand nombre de molécules, des variations faibles sont moins visibles. Deuxièmement, le bruit est « tamponné » au niveau protéique, c'est-à-dire que des mécanismes tendent à atténuer les variations dans le nombre de protéines. Parmi ces mécanismes, on retrouve les taux de dégradation des protéines qui sont généralement plus longs que ceux des ARNm, et qui réduisent les effets de bursts [49, 68] ou encore la localisation des protéines dans le cytoplasme [69].

1.2.4 La Stochasticité de l'Expression Génique : nuisance ou rôle biologique ?

Il a été proposé que cette variabilité est un « bruit », une nuisance dont les cellules doivent s'accommoder au cours de leur vie, notamment parce que d'un point de vue déterministe, des niveaux d'expression génique contrôlés avec précision semblent optimaux. De plus, cette hypothèse est étayée par l'existence de mécanismes moléculaires permettant de réduire la variabilité de l'expression génique. A l'inverse, un grand nombre d'études menées sur différents organismes et dans différents processus biologiques, démontrent un rôle essentiel de la variabilité de l'expression génique.

1.2.4.1 Les mécanismes de réduction du bruit

Chez la levure *S. cerevisiae*, la variabilité des gènes codant pour des protéines dites « essentielles » est minimisée à la fois au niveau transcriptionnel et traductionnel. En effet, ces gènes ont des taux de transcription et de dégradation des ARNm élevés et des taux de traduction assez bas comparés à des gènes dont la variabilité est plus importante (Figure 1.4). En d'autres termes, la durée de vie des protéines est beaucoup plus longue que l'intervalle de temps entre les bursts de production donc l'accumulation de protéines dans le temps compense la variabilité générée par l'expression pulsatile et stochastique des gènes codants pour celles-ci [70].

Il a été montré, dans un système de biologie synthétique chez *E. coli*, que les boucles de rétroaction négative dans les réseaux de gènes, où la protéine produite régule sa propre transcription, tendent à amortir la variabilité de l'expression du gène contrôlé par ce réseau [72]. Cependant, il

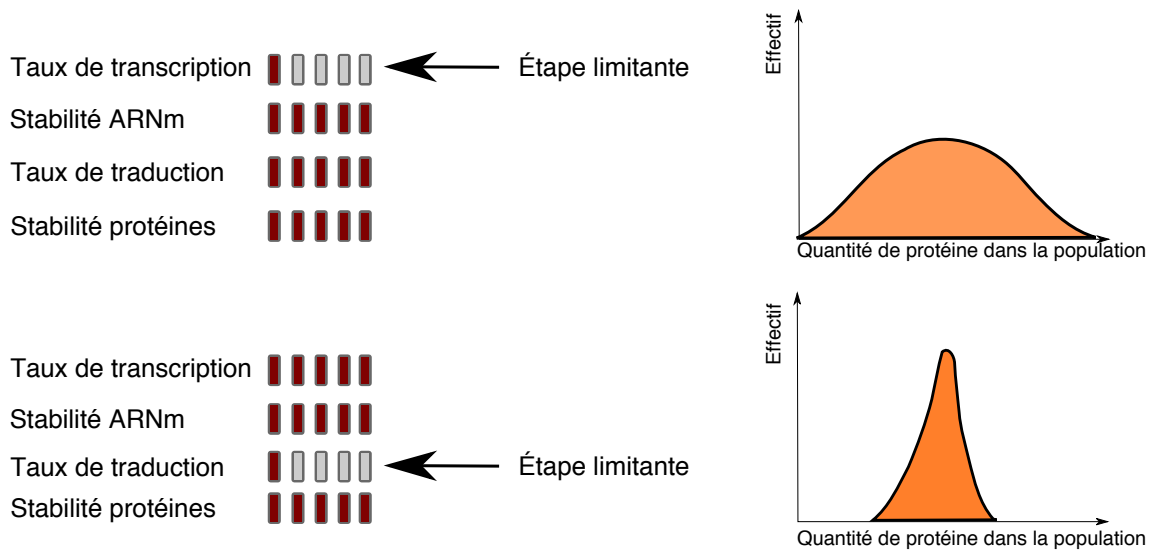


FIGURE 1.4 – Schéma des effets des taux de transcription et traduction sur la variabilité de l'expression génique.

Une transcription peu fréquente suivie d'une traduction efficace entraîne un bruit intrinsèque élevé dans les niveaux de protéines, puisque l'étape limitante est la transcription et qu'elle se produit par burst (haut) ; une transcription fréquente et une traduction inefficace entraînent un bruit intrinsèque faible, puisque l'étape limitante (la traduction) est compensée par la durée de demi-vie des protéines (bas). Schéma adapté de Mundt et al. 2018 [71].

a aussi été montré que cette propriété est dépendante de la fréquence des bursts de transcription. En effet, chez *Dictyostelium*, les gènes dont l'expression est réprimée par une boucle de rétroaction négative ont des niveaux de variabilité plus importants dans le cas où la fréquence des bursts est faible [73] (Figure 1.5).

Le transport des pré-ARNm peut être également un mécanisme important dans la réduction de la variabilité des ARNm. En effet, à l'inverse de l'exemple développé dans la partie précédente (1.2.3.2), si le taux d'export nucléaire est inférieur au taux de dégradation des ARNm (et est donc lent), le nombre d'ARNm dans le cytoplasme aura tendance à moins varier [66, 75] (Figure 1.6).

Enfin, récemment chez la souris, une équipe a mis en évidence que certains miRNA sont des acteurs importants pour la réduction de la variabilité de gènes cibles en accélérant le taux de dégradation des ARNm issus du gène [76].

Mais le coût de la suppression du bruit est très important. Des auteurs ont évalué ce coût en modélisant un système simplifié où une protéine X_2 est produite proportionnellement à la quantité d'un ARNm X_1 , X_1 est produit de manière stochastique, et donc X_2 aussi. Les auteurs

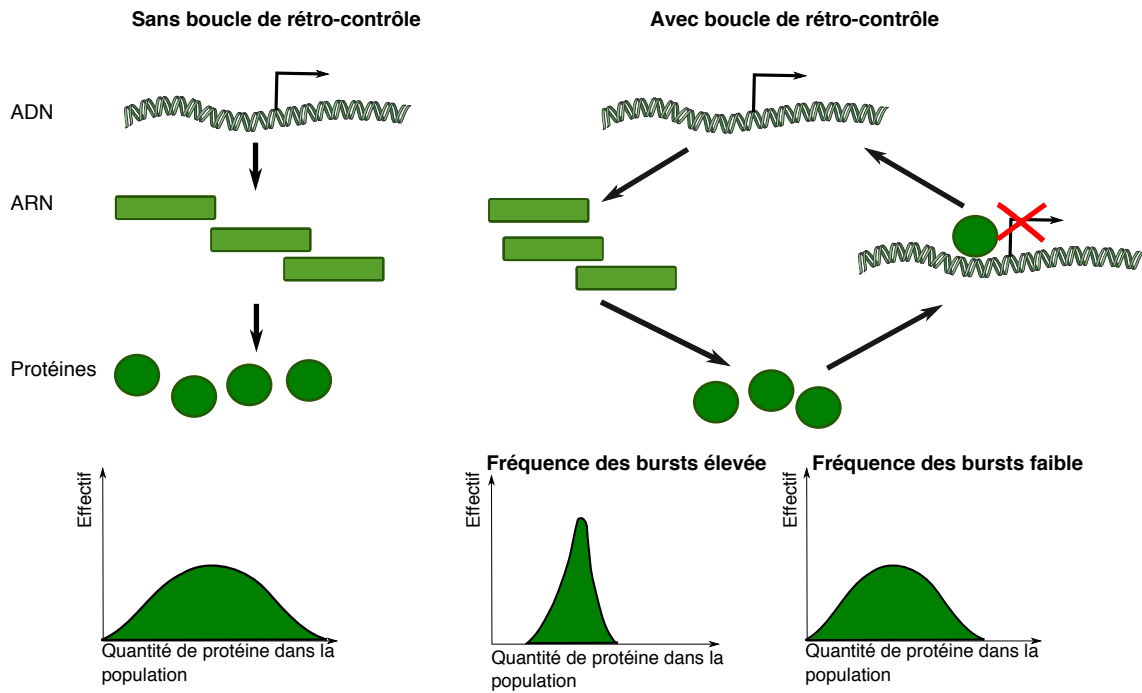


FIGURE 1.5 – Schéma des effets des boucles de rétro-contrôle négatif sur la variabilité de l'expression génique.

La variabilité de l'expression d'un gène donné est diminuée lorsque la protéine codée par ce gène régule négativement sa propre expression si la fréquence des bursts de transcription est élevée. Schéma adapté de Raser et al. 2005 [74].

ont montré que si la cellule veut réduire la variation dans le niveau de X_2 , elle va devoir produire 16 fois plus de molécules de X_1 [77]. Ainsi, le coût énergétique deviendrait rapidement trop élevé si une cellule devait réduire le bruit de tous ses gènes. Dans le premier exemple décrit ci-dessus de réduction du bruit lors de l'expression des gènes essentiels chez la levure, cette réduction se fait au prix d'une faible production de protéines pour une production très importante de molécules d'ARNm.

1.2.4.2 Le variabilité comme moteur des processus de décision cellulaire

A l'inverse, de nombreuses études démontrent un rôle fonctionnel de la variabilité de l'expression génique dans les processus de décision cellulaire.

S'adapter à l'environnement

La variabilité permet d'adopter des stratégies de « bet-hedging » permettant à des cellules, issues d'une population homogène et soumises au même environnement d'être dans des états transcriptionnels différents [78, 79]. Ainsi, en cas de changement brutal de l'environnement, certains de ces états pourraient être plus adaptés aux nouvelles conditions. Ce mécanisme pourrait ainsi permettre à des cellules de garder une forte capacité d'adaptation, sans que le coût ne soit porté

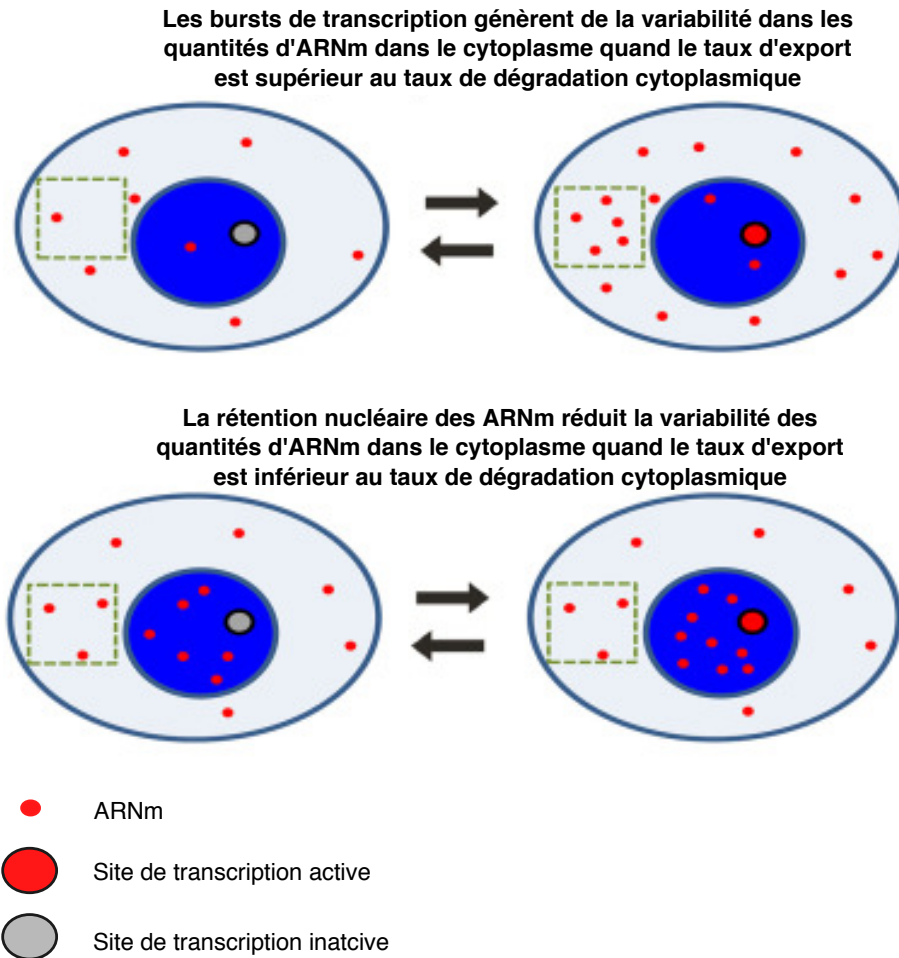


FIGURE 1.6 – Schéma des effets de la dynamique de l'export nucléaire des ARNm sur la variabilité de l'expression génique.

Détails dans le texte. Schéma adapté de Bahar Halpern et al. 2015 [42].

par leur ADN [80]. Le bet-hedging est particulièrement relevant pour des cellules autonomes, comme par exemple les bactéries.

Générer de la diversité

Dans des cellules d'organismes multi-cellulaires, la variabilité peut permettre de diversifier les types cellulaires. C'est le cas par exemple lors de la génération des neurones olfactifs chez la souris. Un neurone ne produit qu'un seul type de récepteur olfactif. Chaque cellule exprime aléatoirement un gène codant pour un récepteur parmi 1500 gènes possibles, ce qui génère une très grande diversité dans les signaux olfactifs que la souris va pouvoir traiter [81, 82]. Si ce processus était déterministe, le mécanisme de régulation serait sûrement extrêmement complexe et coûteux à l'échelle du tissu.

Un exemple similaire existe dans la distribution des types de photorécepteurs chez l'homme et chez la drosophile. Dans ces deux cas, au niveau de la rétine, la détermination stochastique

du photorécepteur favorise une répartition aléatoire de ceux-ci, et prévient la génération de clusters de cellules exprimant le même type de photorécepteur. En effet, la présence de clusters de photorécepteurs similaires pourrait être délétère pour la vision en diminuant la capacité de l'oeil à discriminer les contrastes ou les couleurs [27, 83].

Une autre illustration qui cette fois implique plusieurs types cellulaires est la spécification de l'épiderme pendant le développement de la drosophile. Dans un cluster homogène, une cellule va devenir un neuroblaste et toutes les autres cellules du cluster formeront l'épiderme par inhibition latérale. C'est une légère augmentation stochastique de l'expression du ligand Delta de la voie Notch par une des cellules qui détermine la cellule qui deviendra le neuroblaste. L'expression de Delta par cette cellule entraîne l'expression du récepteur Notch par les cellules avoisinantes les empêchant ainsi d'exprimer Delta. Toutes les cellules du cluster ont la capacité de devenir le neuroblaste. En effet, si le neuroblaste en place est détruit, et donc qu'il n'y a plus d'inhibition latérale, une autre cellule peut aléatoirement devenir à son tour le neuroblaste [84].

Induire la différenciation

La variabilité peut aussi induire la différenciation et ce dans différents modèles biologiques procaryotes et eucaryotes. Par exemple, chez *B. subtilis* l'acquisition de l'état de compétence (qui est qualifié d'état de différenciation) passe par une boucle de rétro-action positive. Le gène codant pour la protéine comK est exprimé à bas bruit. Puis, son niveau d'expression augmente légèrement de manière stochastique. Au-delà d'un certain seuil, son expression est alors amplifiée *via* une boucle de rétro-action positive et augmente rapidement permettant d'atteindre l'état de compétence. Lorsque comK est fortement présente, son expression va être ensuite réprimée par un répresseur qui mettra fin à la période de compétence. Une étude a montré que le bruit généré par la stochasticité de l'expression de *comK* jouait un rôle crucial dans l'initiation de la compétence. Il permet en effet d'augmenter le niveau de ComK jusqu'à un seuil où la rétro-action positive peut se mettre en place pour initier les périodes de compétence [79].

Dans les cellules eucaryotes, Chang *et al.* ont montré que la variabilité peut aussi influencer la différenciation dans des progéniteurs hématopoïétiques humains. Dans une population homogène de ce type cellulaire, les niveaux de SCA-1, marqueur de pluripotence, sont extrêmement variables, jusqu'à trois ordres de grandeur. Les auteurs ont trié par cytométrie les cellules en fonction de leur niveau de SCA-1. Ils ont ainsi isolé des populations exprimant soit faiblement, moyennement ou fortement *SCA-1*. Les auteurs ont observé qu'au bout de quelques jours toutes

les sous-populations triées reconstituaient la distribution initiale du niveau de SCA-1, c'est-à-dire que son expression était à nouveau très variable dans les trois sous-populations quelque fut le niveau d'expression de départ. Cependant, le niveau initial de SCA-1 a tendance à biaiser le choix de lignage des cellules [85].

Le moyen le plus direct pour étudier la variabilité d'expression génique en cellule unique et comprendre son rôle lors des processus de décision cellulaire est de quantifier directement les niveaux d'ARNm dans des cellules individuelles. En effet, la demi-vie des ARNm est généralement beaucoup plus courte que celle des protéines, et leurs niveaux reflètent donc plus précisément l'état d'activation d'un gène. Ainsi, dans la suite du manuscrit nous nous concentrerons sur la variabilité au niveau transcriptionnel.

1.3 Les outils pour étudier la variabilité de l'expression génique au niveau transcriptionnel

1.3.1 Les outils expérimentaux pour étudier la variabilité de l'expression génique en cellule unique

Différentes techniques permettent d'évaluer la variabilité de l'expression génique au niveau transcriptionnel. De manière générale, ces outils permettent de quantifier le nombre de molécules d'ARNm, correspondants à un gène donné, présents dans une seule cellule selon différentes résolutions : les ARNm d'un gène spécifiques, les ARNm de quelques gènes différents, les transcrits d'une centaine de gènes cibles voire à l'échelle de tout le transcriptome. Le choix entre ces différents outils va dépendre de la question biologique posée.

1.3.1.1 La détection et la localisation des ARNm issus d'un gène

Le Tagging MS2 permet à la fois de quantifier et de localiser les ARNm issus d'un seul gène à l'échelle de la cellule unique. La méthode de Tagging MS2 est issue de l'interaction naturelle d'une protéine de bactériophage, la protéine MS2, avec une structure tige-boucle du génome du phage. Ces séquences génomiques qui forment les structures tige-boucles peuvent être insérées dans une région non codante du gène d'intérêt qui sera alors transcrit avec ces structures particulières. La protéine MS2, exprimée par la cellule après transfection du gène correspondant, va reconnaître et se fixer sur ces structures tige-boucles. L'ARNm cible peut être ainsi détecté par microscopie à fluorescence en couplant la protéine MS2 à une protéine fluorescente comme la GFP (Green Fluorescent Protein) [86, 87] (Figure 1.7).

C'est par cette approche, qu'a été mise en évidence la stratégie « Monte Carlo », c'est-à-dire les choix aléatoires lors du développement des neurones olfactifs chez la souris décrits plus haut [81] ainsi que la spécification des photorécepteurs chez la drosophile [27].

Le tagging MS2 présente l'avantage majeur de quantifier très précisément le nombre de molécules d'ARNm du gène d'intérêt mais aussi de donner une information sur la localisation de ces molécules dans les cellules. Cette méthode peut aussi donner une information dynamique, si l'analyse est réalisée en microscopie à intervalles de temps (time-lapse), et permettre de détecter les ARNm en cours de synthèse.

Cependant, le Tagging MS2 demandent de pouvoir modifier génétiquement et de transfecter

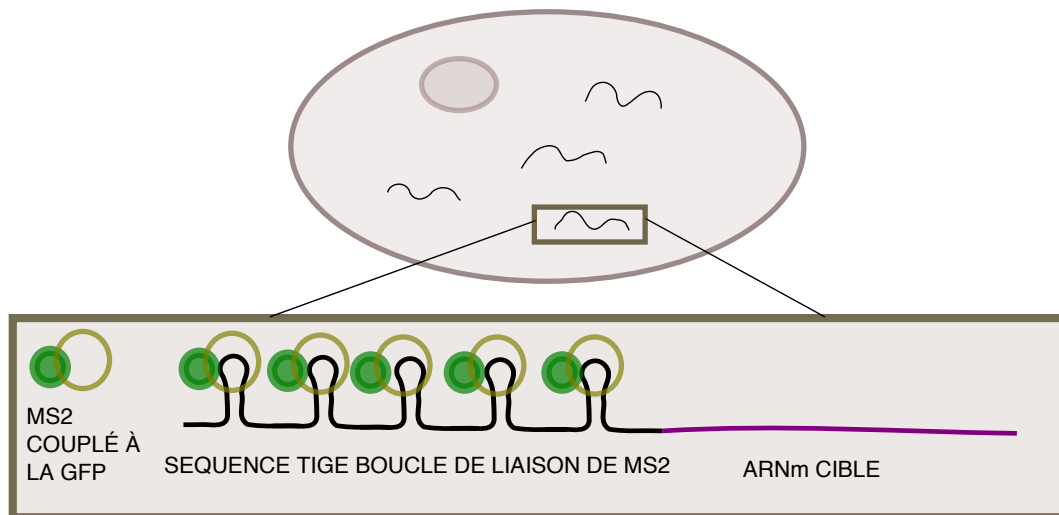


FIGURE 1.7 – Schéma du Tagging MS2.

Les séquences formant des structures tige-boucles sont insérées avant ou après la séquence codante du gène d'intérêt et seront spécifiquement reconnues par la protéine MS2 couplée à la GFP. Le signal est ensuite détecté par microscopie à fluorescence.

facilement les cellules étudiées, ce qui n'est pas toujours possible. De plus, il faut en général insérer plusieurs séquences générant les structures tige-boucles afin qu'une grande quantité de protéines MS2 s'y fixent et ainsi rendre le signal détectable. Également, même si l'insertion se fait dans les régions non codantes du gène, cela peut néanmoins impacter sa régulation. Enfin, un seul gène ne peut être étudié à la fois.

1.3.1.2 La détection et la localisation des ARNm issus de quelques gènes

Le sm-RNA-FISH (single molecule RNA Fluorescent *In Situ* Hybridization) permet également de faire à la fois de la quantification et de la localisation d'ARNm de quelques dizaines de gènes différents simultanément, dans une seule cellule.

Le sm-RNA-FISH, est une adaptation du RNA-FISH qui consiste à utiliser des sondes fluorescentes constituées d'une séquence d'ADN simple brin qui se lie par complémentarité sur les séquences des ARNm cibles. Plusieurs sondes se fixent sur la même cible ce qui permet d'amplifier le signal fluorescent et rendre la molécule d'ARNm détectable. Il est possible d'utiliser une dizaine de sondes différentes pour suivre simultanément l'expression d'une dizaine de gènes différents. Le signal fluorescent est ensuite analysé par microscopie à fluorescence sur cellules fixées [88, 89] (Figure 1.8).

C'est en utilisant le sm-RNA-FISH que Hansen *et al.* ont observé l'amplification cytoplasmique de la variabilité des ARNm décrit précédemment dans une lignée de cellules souches

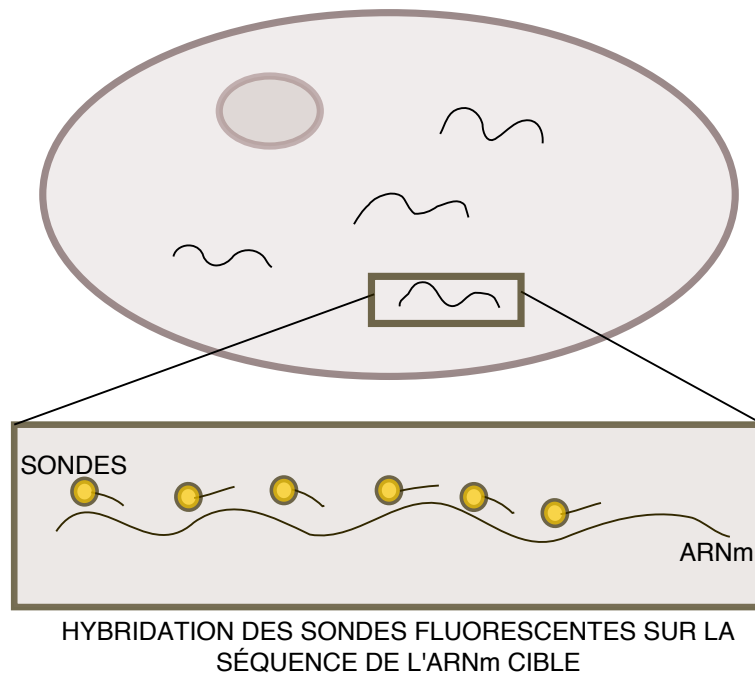


FIGURE 1.8 – Schéma du Single-molecule FISH.

Des sondes fluorescentes spécifiques sont introduites dans la cellule et se fixent sur la séquence de l'ARNm ou des ARNm d'intérêt par complémentarité. Le signal est ensuite détecté par microscopie à fluorescence sur cellules fixées.

embryonnaires de souris [66].

La quantification du nombre d'ARNm par sm-RNA-FISH est également particulièrement précise mais nécessite l'utilisation d'algorithmes complexes de déconvolution du signal. De plus, cette méthode peut être compliquée à mettre en place car la synthèse des sondes n'est pas triviale. Enfin, il n'est possible d'observer qu'un nombre limité de cibles en parallèle, ce qui, en fonction de la question posée, peut être limitant.

1.3.1.3 Les approches single cell RT-qPCR

La réaction de polymérisation en chaîne par transcription inverse (RT-PCR) est l'une des techniques principales pour étudier l'expression génique de manière générale. Elle permet la rétro-transcription d'ARNm en ADNc puis l'amplification spécifique des ADNc cibles. Avec cette technique, la quantification des ARNm est particulièrement précise.

Le passage de cette méthode à l'échelle de la cellule unique a été permise par le développement des approches microfluidiques permettant l'isolement des cellules dans des puces, leur lyse puis la détection et la quantification d'ARNm définis.

Single-cell digital PCR (sc-dPCR)

La PCR digitale ou « PCR en dilution limite » est basée sur la compartimentalisation des molécules. Brièvement, la cellule isolée est lysée et ses ARNm rétro-transcrits, puis un mélange réactionnel contenant des amorces ciblant le gène d'intérêt est ajouté à l'échantillon. L'échantillon est alors chargé dans la puce microfluidique. La puce est construite de telle sorte que chaque molécule d'ADNc va être isolée dans une chambre de réaction (aussi appelée partition). Chaque puce possède 10000 à 30000 partitions. La PCR est réalisée en parallèle dans chaque partition. Pendant la PCR, des molécules fluorescentes sont intercalées dans l'ADN puis les puces sont analysées par lecture de la fluorescence. Une amplification génère un signal fluorescent dans la partition, ce qui n'est pas le cas en absence d'amplification. On calcule ensuite le nombre de partitions où le signal est positif ce qui permet de connaître le nombre de molécules initial. Certaines puces permettent la capture des cellules, la RT, et la PCR de manière automatisée ce qui limite fortement le risque de perte de matériel lié à la manipulation des échantillons [90] (Figure 1.9).

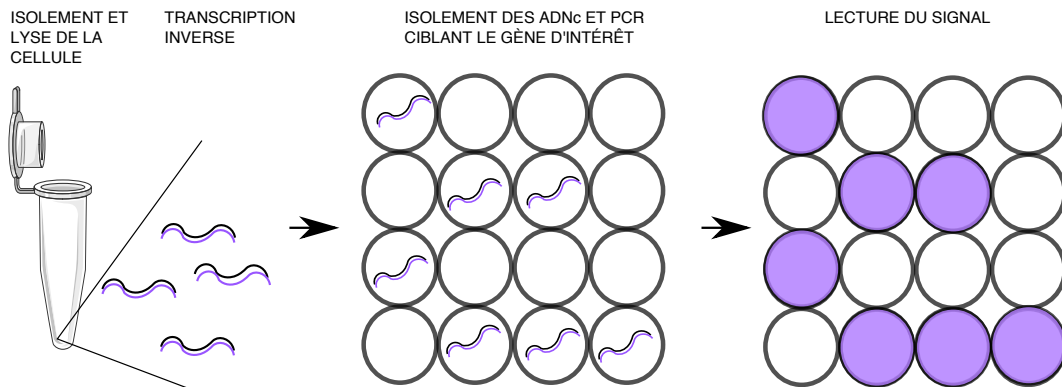


FIGURE 1.9 – Schéma très simplifié de la PCR digitale.

La cellule est isolée, lysée puis ses ARNm sont rétrotranscrits avant d'être chargés dans la puce microfluidique avec un mélange réactionnel contenant les amorces ciblant le gène d'intérêt. Chaque duplex ARNm/ADNc est isolé dans une partition. La PCR est réalisée en parallèle dans toutes les partitions et au cours de l'amplification un intercalant d'ADN fluorescent est incorporé.

Il est possible de multiplexer l'analyse pour détecter jusqu'à 5 cibles en parallèle, dans plusieurs centaines de cellules en même temps et la quantification est très précise. Cependant, ce multiplexage est compliqué et est limité par l'utilisation de couleurs de fluorescence différentes pour étudier des gènes différents. Il n'est donc pas possible d'étudier beaucoup de gènes simultanément.

Single-cell RT-qPCR Fluidigm (scRT-qPCR)

Le Single-cell RT-qPCR, et notamment la technique de Fluidigm, permet une analyse à plus haut débit car il est possible de quantifier simultanément l'expression de 96 gènes dans 96 cellules (Figure 1.10). Pour cela, les cellules sont initialement triées en plaque 96 puits par exemple par cytométrie en flux (une cellule par puits), dans laquelle elles sont lysées. Les ARNm sont rétro-transcrits puis une étape de pré-amplification est réalisée permettant de passer outre les problèmes liés à la faible quantité initiale de matériel. Les ADNc totaux de chaque cellule sont ensuite chargés dans la puce et répartis en 96 chambres où les 96 gènes cibles vont être amplifiés spécifiquement par PCR. Le résultat est un tableau (ou heat map) où les colonnes représentent les 96 gènes, les lignes représentent les 96 cellules et les intersections marquées par des spots fluorescents d'intensités différentes représentent l'expression du gène considéré dans la cellule considérée. Les données sont d'abord normalisées à l'aide d'ARN synthétiques, appelés Spikes, de concentration connue et ayant été intégrés dans chaque puits au moment de la lyse. L'analyse des résultats est ensuite assez semblable à ceux d'une RT-qPCR en population. Plus la valeur du Ct est faible, plus le signal est intense, et donc plus le gène a une expression élevée dans la cellule. A l'inverse, plus la valeur du Ct est élevé, moins le signal est fort et moins l'expression du gène est élevée [91].

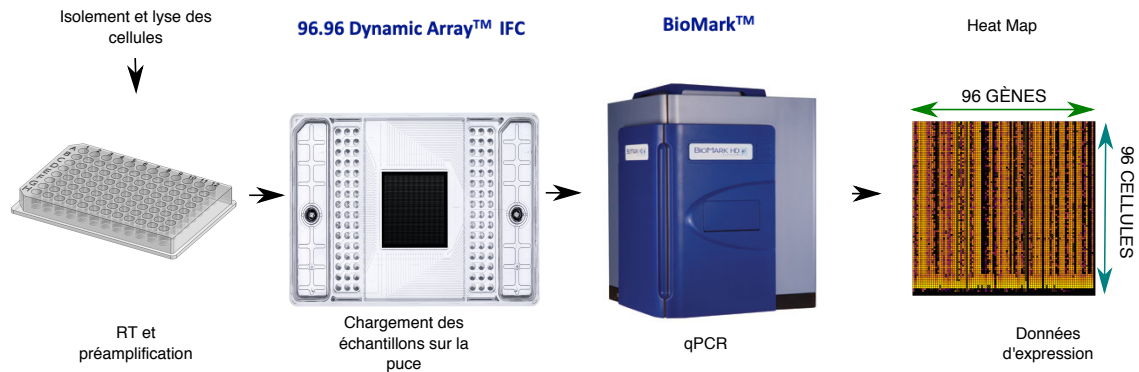


FIGURE 1.10 – Schéma de la Single-cell RT-qPCR.

Les 96 cellules sont isolées et lysées, puis les ARNm sont rétrotranscrits et pré-amplifiés avant d'être chargés dans la puce microfluidique avec un mélange réactionnel contenant les 96 amorces ciblant les 96 gènes d'intérêts. La PCR est réalisée à l'aide du BioMark (Fluidigm); au cours de l'amplification un intercalant d'ADN fluorescent est incorporé afin de quantifier le signal.

Les méthodes d'analyse de l'expression génique en cellules uniques basées sur la RT-PCR couplée à la microfluidique (sc-dPCR et scRT-qPCR (Fluidigm)) sont particulièrement sensibles, permettant de détecter des gènes très peu exprimés. Les réactions se produisant dans des vo-

lumes très faibles (ce qui limite les contaminations d'espèces moléculaires pouvant inhiber les réactions chimiques), ces méthodes favorisent un très bon ratio signal sur bruit permettant une quantification précise des molécules présentes initialement dans la cellule. Néanmoins, bien que la réaction responsable de la plus grande variabilité soit la RT, la pré-amplification est aussi une étape qui peut induire un biais. Enfin, il est possible d'étudier entre une dizaine et une centaine de gènes préalablement sélectionnés en parallèle ce qui peut être plus ou moins adapté selon la question biologique posée et induire un certain biais par le choix des gènes analysés. Également, ces techniques sont assez coûteuses et la multiplication des plaques reste limitée par cet aspect financier [92].

1.3.1.4 Le single-cell RNA sequencing (sc-RNA-seq)

Le single-cell RNA sequencing permet d'étudier la variabilité de l'expression génique à l'échelle de tout le transcriptome. Il existe un grand nombre de protocoles disponibles dont nous allons voir quelques spécificités. Pour des raisons de clarté, nous nous concentrerons sur 4 protocoles très utilisés qui illustrent assez bien le panel d'options disponibles : le Drop-seq, le Smart-seq2, le Mars-seq et le 10X. Globalement, toutes ces méthodes sont basées sur la capture des molécules d'ARNm de chaque cellule par leur queue polyA et l'insertion d'un code-barre unique pour chacune des cellules au niveau des ARNm. C'est grâce à ce code-barre qu'il sera possible, après séquençage, de réattribuer chaque ARNm à la cellule dont il provient. On peut décomposer les protocoles de sc-RNA-seq en 4 étapes : 1) l'isolement et la lyse des cellules ; 2) la Reverse Transcription et le barcoding des ARNm ; 3) l'amplification des ADNc ; 4) le séquençage (Figure 1.11).

1) L'isolement et la lyse des cellules

Les cellules peuvent être isolées par deux méthodes : par microfluidique (Drop-seq et 10X) ou par FACS (Fluorescence Activated Cell Sorting - Smart-seq et Mars-seq).

L'isolement par microfluidique se fait dans des gouttes aqueuses entourées d'huile dans lesquelles vont être encapsulées une cellule et une bille portant des amorces Oligo(dT) avec une séquence unique de code-barre cellulaire. Il y aura un code-barre cellulaire par bille et par extension un code-barre par cellule. La densité cellulaire est contrôlée de manière à ce que la probabilité que deux cellules se retrouvent dans la même goutte soit extrêmement faible ; le corollaire étant que beaucoup de gouttes sont en fait vides.

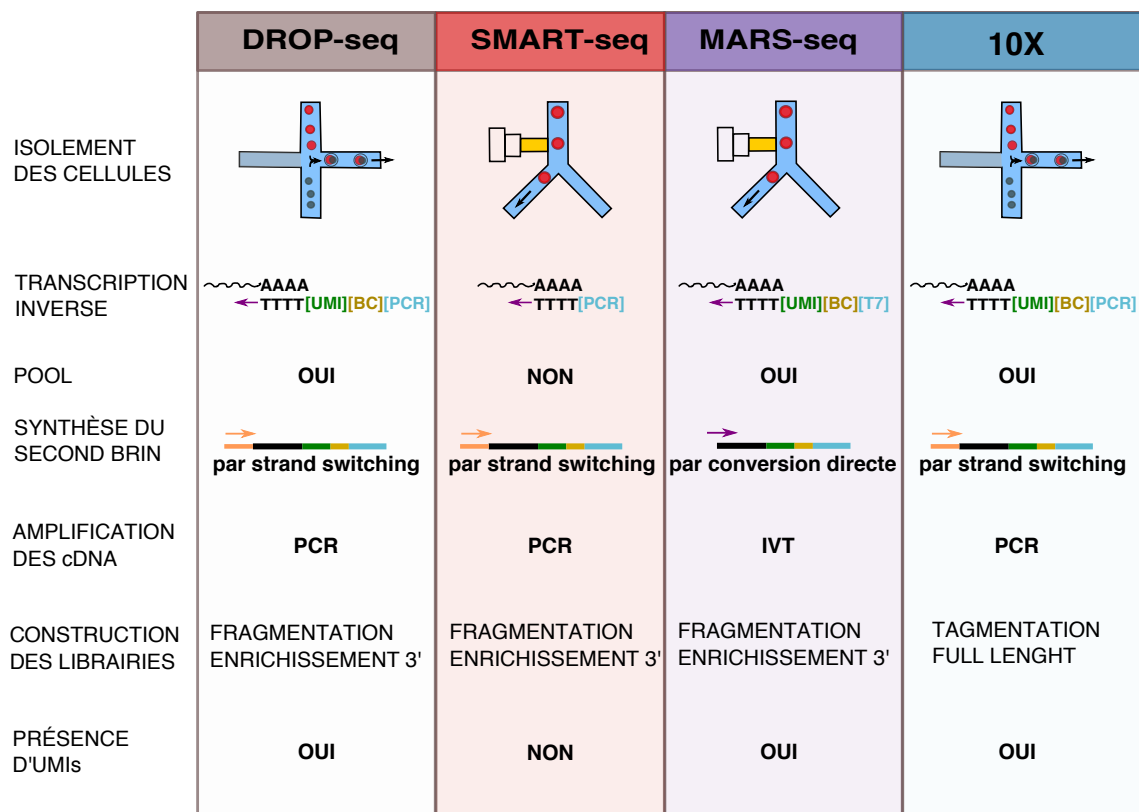


FIGURE 1.11 – Schéma général des méthodes de single cell RNA-seq.

Les cellules sont d'abord isolées par FACS ou par microfluidique, puis lysées. Les ARNm des cellules sont ensuite rétro-transcrits et les amorces contenant les code-barres cellulaires uniques (BC) et les UMIs (UMI) sont intégrées à la séquence des ARNm. Les cellules sont ensuite poolées et le second brin des ADNc synthétisé. Les ADNc sont ensuite amplifiés par PCR ou IVT, les molécules sont fragmentées et enfin les amorces permettant le séquençage sont insérées. Schéma adapté de Ziegenhain et al. 2017 [93].

L'isolement par FACS est réalisé en plaques 96 ou 384 puits. Comme pour les gouttes, chaque puits va contenir des amorces avec un code-barre cellulaire unique (un code-barre cellulaire par puits). L'avantage principal du tri par FACS est la possibilité de sélectionner en amont des sous-populations de cellules d'intérêt à l'aide de marqueurs membranaires par exemple. Le FACS permet également d'enregistrer des informations morphologiques de chaque cellule triée grâce au Forward Scatter Channel (FSC) et Side Scatter Channel (SSC) dont les signaux donnent une indication sur la taille et la granulosité des cellules, respectivement. Le tri par FACS en plaque implique par contre de « grands » volumes de réaction, de l'ordre de quelques microlitres, alors que les volumes de réaction en gouttes sont très inférieurs de l'ordre de quelques nanolitres. Toutefois, il est tout à fait possible de réaliser un enrichissement de cellules d'intérêt par tri par FACS puis d'encapsuler les cellules sélectionnées en gouttes.

Le module C1 de la société Fluidigm combine d'une certaine manière ces deux approches.

Après l'isolement de chaque cellule par microfluidique, l'appareil prend une photo de celle-ci, ce qui permet de récupérer des informations morphologiques de la cellule isolée mais aussi des informations sur l'expression de certaines protéines si un marquage a été préalablement effectué. Cependant avec le C1, pour espérer capturer 96 cellules, la population de départ doit être d'un millier de cellules, ce qui n'est pas toujours possible notamment pour les cellules rares.

Après l'isolement des cellules, aussi bien en plaques qu'en gouttes, les cellules sont lysées.

2) La Reverse Transcription et le barcoding des ARNm

Une fois les cellules lysées, les ARNm vont être rétro-transcrits (RT) à partir de leur queue polyA grâce aux amorces comportant une séquence Oligo(dT), le code-barre cellulaire unique à chaque cellule et un UMI (Unique Molecular Identifier - Figure 1.12). L'UMI est une séquence aléatoire de 4 à 10 paires de bases qui sera unique à chaque molécule d'ARNm. Ainsi les UMIs permettront de faire la différence entre un transcrit présent plusieurs fois (plusieurs ARNm d'un même gène avec des UMIs différents) et des amplifiats PCR (plusieurs ARNm d'un même gène avec le même UMI). Après la RT, comme chaque molécule d'ADNc est identifiée (e.g. porte un code-barre cellulaire et un UMI), il est possible de regrouper toutes les cellules pour la suite de la construction des banques, ce qui réduit grandement les coûts.

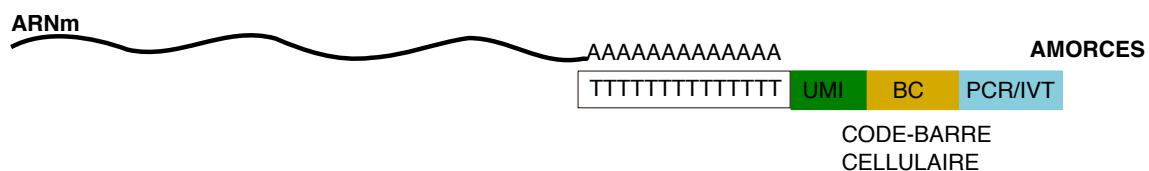


FIGURE 1.12 – Schéma du barcoding des ARNm.

Les amorces permettant le barcoding des ARNm sont composées d'un Oligo(dT) qui va se fixer par complémentarité sur la queue polyA des ARNm, d'un UMI (Unique Molecular Identifier) qui sera unique à chaque molécule, un code-barre cellulaire unique à chaque cellule (BC), et une séquence pour l'amplification ultérieure (par PCR ou IVT).

3) L'amplification des ADNc.

Après la RT et le regroupement des ADNc (simple brin) de toutes les cellules, le second brin est synthétisé et les ADNc sont ensuite amplifiés afin de les obtenir en quantités suffisantes pour être analysés. Les ADNc sont amplifiés soit par PCR soit par In Vitro Transcription (IVT). L'avantage de l'IVT par rapport à la PCR est que l'amplification est linéaire et non exponentielle. Ainsi les biais d'amplification sont mieux contrôlés. Les ADNc vont ensuite être fragmentés et

seule la partie 3' qui porte le code-barre va être conservée et analysée (on parle de banque 3'). Il existe aussi des protocoles permettant de conserver la partie 5' des ARNm mais nous n'en parlerons pas dans ce manuscrit.

D'autre part, quelques méthodes dites « full length » ont été mises au point et permettent de récupérer l'intégralité de la séquence des transcrits, nécessaire si l'on souhaite étudier par exemple les sites d'épissage des ARNm. C'est le cas des protocoles SMART-seq. Cependant l'inconvénient est qu'il faut utiliser une tagmentase qui va insérer aléatoirement les code-barres cellulaires à plusieurs endroits du transcrit et générer des petits fragments de taille plus ou moins variable. L'étape de tagmentation est réalisée à la fin de la construction des banques, après la dernière étape d'amplification par PCR pour que celle-ci soit uniforme sur tous les fragments et éviter des biais d'amplification qui pourraient favoriser des fragments de petites tailles au détriment de fragments de plus grandes tailles. Il n'est donc pas possible d'insérer les code-barres cellulaires et les UMIs lors de la première étape de RT, ni de pooler les cellules, ce qui augmente drastiquement le coût de production de ce type de banques, puisque les cellules sont traitées en parallèle tout au long du protocole. Le SMART-seq3 permet néanmoins l'utilisation d'UMIs, mais de la même manière que pour les autres protocoles SMART-seq, les échantillons ne peuvent être poolés qu'après l'insertion des code-barres lors de l'étape de tagmentation après l'amplification.

4) Le séquençage

Une fois les banques terminées, elles vont être préparées pour le séquençage. Des séquences spécifiques sont insérées à chaque extrémité de chaque molécule d'ADNc (soit par ligation soit par PCR). Ces séquences sont en fait des amorces qui vont être utilisées lors du séquençage pour ancrer les molécules sur la flowcell de séquençage et permettre le démarrage de la lecture des séquences. Lors du séquençage, un ADNc donné va être amplifié et donc « lu » plusieurs fois dans une petite zone de la flowcell. C'est le concept de la PCR en pont. Les bases incorporées lors de cette PCR particulière portent différentes couleurs. C'est à partir de la lecture optique lors de l'incorporation des bases que la séquence est déduite. Lire plusieurs fois la même séquence augmente la quantité et la qualité du signal (Figure 1.13).

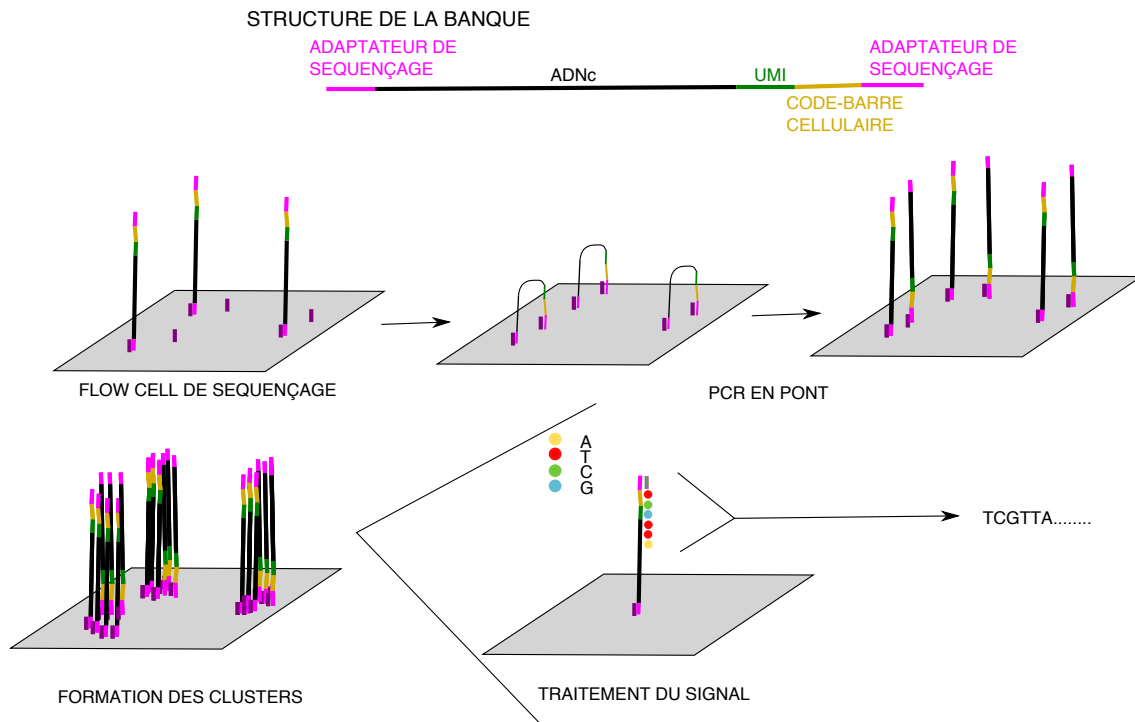


FIGURE 1.13 – Schéma du séquençage Illumina.

Les ADNc sont chargés sur la flow cell et vont s'hybrider sur les amorces Illumina complémentaires aux séquences insérées dans les ADNc. Par PCR en pont, les ADNc sont amplifiés et forment des clusters de séquences identiques. Lors de la phase de lecture, des nucléotides fluorescents sont intégrés à mesure de la synthèse du brin complémentaire permettant ainsi une lecture optique de la séquence.

Le choix des paramètres du protocole de séquençage sera déterminé par la structure des banques. Le séquençage peut être unilatéral ou bi-latéral (single-end ou pair-end) en fonction de l'information présente de chaque côté des ADNc (par exemple un code-barre de chaque côté des séquences nécessitera un séquençage pair-end). La longueur des reads, ou lectures, voulue, c'est-à-dire combien de bases vont être lues, va dépendre de la taille des fragments (en générale la longueur des reads varie entre 50 et 150pb). Le nombre de reads obtenu va lui dépendre de la profondeur de séquençage souhaité, qui elle même peut dépendre du nombre de cellules, de la rareté des molécules recherchées ou encore de l'abondance globale du transcriptome dans le modèle étudié. En général, on cherche à obtenir 200 000 reads bruts par cellule [94].

Les méthodes de sc-RNA-seq diffèrent sur plusieurs paramètres :

1) la sensibilité qui correspond au nombre de gènes différents qui vont pouvoir être détectés, plus précisément c'est le nombre minimal d'ARNm d'un gène nécessaire pour que ce gène puisse être détecté ; généralement les méthodes full-length sont plus sensibles et permettent de détecter un peu plus de gènes que les méthodes 3' ;

- 2) la précision, c'est-à-dire l'exactitude du nombre de molécules par gène détectées, les méthodes utilisant des UMIs ayant une meilleure précision que les autres puisqu'elles corrigent les biais introduits pendant l'amplification des ADNc ;
- 3) la robustesse, c'est-à-dire la variabilité technique due à la méthode. Cette variabilité est mieux gérée avec les méthodes utilisant des UMIs [93].

Le protocole sc-RNA-seq optimal dépend en fait principalement de la question biologique qui est posée mais aussi du système biologique (Figure 1.14). Combien de cellules sont disponibles ? Représentent-elles une population rare mais identifiable parmi une autre population ? Est-il nécessaire d'avoir la séquence complète des ARNm ? Faut-il connaître à l'avance la séquence des code-barres cellulaires que chaque cellule va porter pour identifier par exemple les relations de généalogie entre les cellules ? D'autres facteurs annexes mais non négligeables rentrent aussi en compte lors du choix du protocole, notamment le coût, l'accès à des machines spécifiques (C1 ou 10X), ou encore la possibilité d'utiliser un système robotisé de pipetage. Enfin, idéalement les analyses bio-informatiques que l'on souhaite réaliser en aval doivent aussi être prises en compte de manière à ce que le design expérimental soit le mieux adapté.

	DROP-seq	SMART-seq	MARS-seq	10X
Beaucoup de cellules disponibles	X	X		X
Peu de cellules disponibles ou population de cellules rares		X	X	
Séquence complète		X		
Coût	++	+++	+	++
Équipements particuliers	X			X

FIGURE 1.14 – Schéma des applications possibles en fonction des méthodes de single cell RNA-seq.

Les quatre méthodes de sc-RNA-seq présentées ici sont évaluées en fonction de différents types d'application (questions biologiques ou contraintes techniques). Schéma adapté de Ziegenhain et al. 2017 [93].

L'avantage majeur du sc-RNA-seq est qu'il permet de récupérer l'information de l'expression

globale d'un grand nombre de gènes, sans à priori et dans un grand nombre de cellules en parallèle. Cependant, ces méthodes sont moins sensibles que les approches scRT-PCR, notamment à cause du phénomène de drop out [95]. Le drop out est directement corrélé à la sensibilité de la méthode et induit de la perte d'information sur tous les transcrits. Cette perte d'information peut même conduire à la non détection des transcrits les moins abondants et ainsi générer des zéros dits techniques. En sc-RNA-seq, il peut être difficile de différencier un zéro technique d'un zéro biologique. Néanmoins, l'utilisation de transcrits synthétiques appelés ERCC Spikes (Evaluation of the External RNA Controls Consortium), qui consistent en 92 espèces d'ARNm polyadénylés synthétiques utilisés à différentes concentrations connues, permet de prendre en compte et de corriger en partie ce biais [96]. Cependant, il faut noter que les ERCC ne sont pas des contrôles parfaits parce qu'ils ne se comportent pas exactement comme des ARN messagers [97]. Ainsi, d'autres méthodes alternatives pourraient permettre de corriger les variations techniques des méthodes de sc-RNA-seq, comme par exemple utiliser des marqueurs de poids moléculaires ADN introduits dans chaque cellule à la place des ERCC [98].

Le sc-RNA-seq permet ainsi la quantification de l'expression de presque l'ensemble des gènes d'une seule cellule et donc de caractériser son état transcriptomique. La force du sc-RNA-seq est aussi de pouvoir quantifier l'hétérogénéité permettant de prendre en compte la variabilité de l'expression génique dans les études. La comparaison de différents états transcriptomiques en considérant cette variabilité est une information majeure pour identifier les paramètres qui contraignent la prise de décision cellulaire [99]. Les données omics issues des expériences de sc-RNA-seq sont des données complexes et multidimensionnelles. Les étapes permettant l'exploitation de ces données et la sélection des outils d'analyses sont alors cruciales pour donner un sens biologique pertinent aux résultats.

1.3.2 Les outils computationnels pour étudier les données d'expression obtenues en cellules uniques

Après le séquençage, les données se présentent sous la forme de fichiers Fastq, il s'agit d'un texte comprenant des centaines de milliers voir des millions de lignes, les lectures (aussi appelées reads). Ces reads sont les séquences des ADNc et comprennent des informations de qualité pour chaque base encodées en ASCII (American Standard Code for Information Interchange). Chaque caractère ASCII (une lettre, un chiffre ou un symbole) correspond à une valeur de qualité entre

0 et 40. Plus la valeur est haute, plus la base correspondante dans la séquence est de qualité. Si la valeur est très faible, il est possible que la base assignée soit fautive. Il est évident que ces données sont impossibles à traiter manuellement. Pour cela des pipelines bio-informatiques sont utilisés pour générer une matrice de comptage avec en colonne les cellules et en ligne les gènes, où chaque gène va être identifié, quantifié et réattribué à la bonne cellule. Ensuite, les données sont filtrées, normalisées, puis différents outils sont utilisés pour extraire les informations pertinentes relatives à la question biologique posée.

1.3.2.1 Les pipelines bio-informatiques

Un pipeline est un enchaînement de logiciels où chaque logiciel attend la sortie du précédent, exactement à la manière d'une chaîne de montage de voiture. Chaque logiciel va accomplir une tâche. Pour le traitement des données sc-RNA-seq, les étapes, d'un pipeline à l'autre, sont généralement les mêmes, ce sont les logiciels utilisés qui peuvent être différents. De la même manière qu'il n'existe pas de protocole sc-RNA-seq parfait, il n'existe pas de logiciel parfait mais certains sont plus adaptés pour traiter certains types de données ou pour répondre à certaines questions biologiques.

Le pipeline prend en entrée les fichiers Fastq (Figure 1.15). La première étape consiste à filtrer les reads de mauvaise qualité (grâce au codage ASCII) pouvant provenir de problèmes de lecture lors du séquençage. Ensuite, les reads sont démultiplexés par les code-barres cellulaires, c'est-à-dire que chaque read va être réattribué à la cellule à laquelle il appartient. Pour cette étape il est nécessaire de fournir au pipeline une liste des code-barres cellulaires ainsi que la position de ceux-ci dans le read. En général les code-barres sont situés au début de chaque read (correspondant aux 6 premières bases de la séquence pour un code-barre de 6 bases par exemple). Le logiciel va alors chercher les séquences des code-barres cellulaires au début de chaque read et les réarranger en fonction de ce code-barre. Il va éliminer les reads qui n'appartiennent à aucune cellule, provenant par exemple de contaminations. Chaque read est ensuite positionné sur un génome ou un transcriptome de référence, qui est fourni au pipeline. L'outil réalisant cette tâche est un mappeur. Il existe plusieurs mappeurs ayant des caractéristiques différentes, certains pouvant mapper les séquences sur des transcriptomes, d'autres sur des génomes ce qui permet d'avoir des informations sur les sites d'épissage par exemple. Certains mappeurs réalisent du pseudo-mapping, ils ne cherchent pas à positionner précisément chaque read sur le transcriptome de référence, mais ils vont chercher à estimer avec la plus haute probabilité la

compatibilité entre la séquence des reads et la séquence des transcrits. Le choix du mappeur n'est pas trivial et dépend à la fois de la structure des données et du type d'information que l'on souhaite récupérer. Enfin, les reads sont comptés, si les reads contiennent des UMIs, les doublons d'amplification PCR sont éliminés assurant une quantification plus exacte. La sortie du pipeline est une matrice gènes (ou transcrits) par cellule.

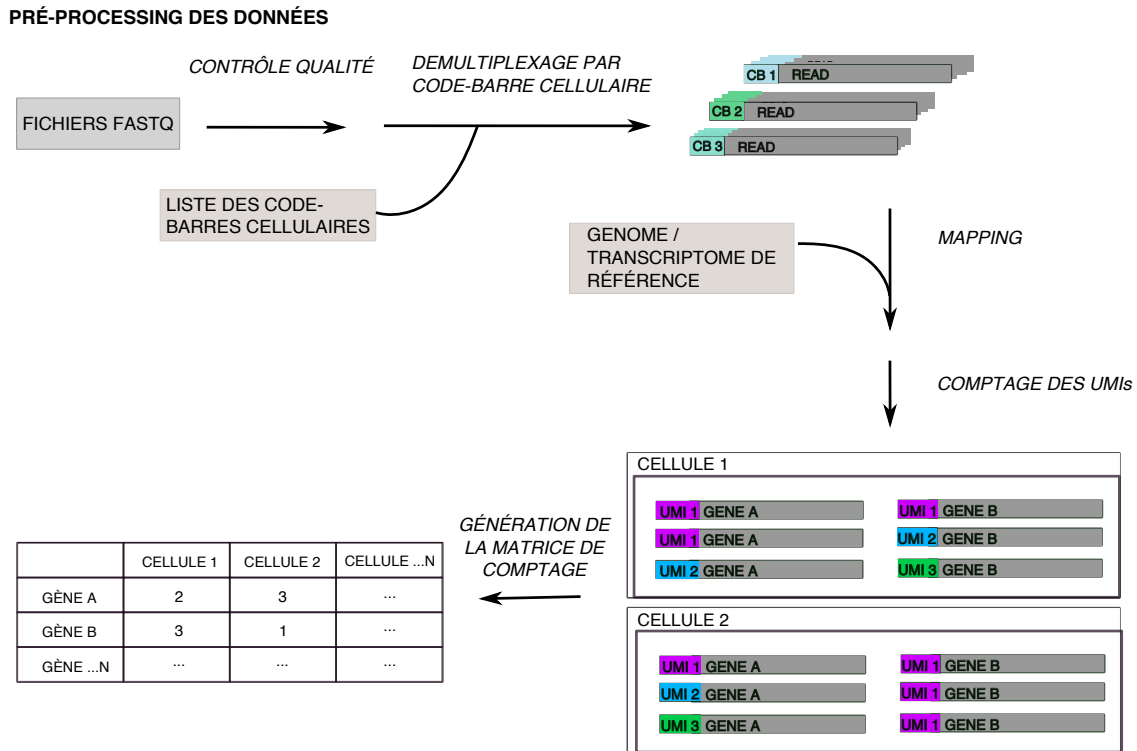


FIGURE 1.15 – Schéma général d'un pipeline bio-informatique de pré-processing des données de sc-RNA-seq.

Le pipeline prend en entrée les fichiers FastQ issus du séquençage, élimine les séquences de mauvaise qualité, démultiplexe les reads par leur code-barre cellulaire (CB), mappe les reads sur un génome ou transcriptome de référence, puis effectue un comptage afin de générer une matrice gènes \times cellules (prise en compte des UMIs dans cet exemple).

1.3.2.2 Les filtres qualités

Une fois la matrice brute générée et avant d'analyser les données, il est nécessaire d'éliminer les cellules générant des résultats de mauvaise qualité. Ces cellules peuvent avoir été mal lysées, ou la RT a pu mal fonctionner, ou encore il peut s'agir de cellules stressées ou mourantes. Si ces cellules sont conservées pendant l'analyse, elles peuvent générer des artefacts en les considérant par exemple comme un sous-type cellulaire différent (Figure 1.16 A).

Plusieurs paramètres peuvent être utilisés pour filtrer les cellules de mauvaise qualité :

- 1) Le nombre total de reads par cellule qui assure d'une couverture homogène lors du séquençage.

- 2) Le pourcentage de reads mappés sur le génome, chaque cellule devant posséder en moyenne le même niveau d'information.
- 3) Le nombre total de gènes détectés par cellule.
- 4) Le nombre d'UMIs par cellule qui correspond au nombre de molécules d'ARNm.
- 5) Le pourcentage de gènes mitochondriaux exprimés qui peut suggérer, lorsqu'il est élevé, que les cellules étaient particulièrement stressées ou mourantes lors de la lyse. Le seuil est souvent fixé à 5% mais le pourcentage de gènes mitochondriaux exprimés est particulièrement cellule-dépendant. Par exemple, dans les cardiomyocytes 45% des gènes exprimés sont des gènes mitochondriaux [100]. Ainsi, dans un mélange de différents types cellulaires, fixer un seuil sur ce paramètre, même plus élevé que 5%, pourrait conduire à l'élimination d'une sous-population particulière de cellules. Une alternative est de filtrer le pourcentage d'ERCC (ou spikes) parmi les gènes cellulaires si l'expérience a été conçue avec des ERCC. Si la majorité des gènes d'une cellule sont des ERCC elle est éliminée.
- 6) Si des ERCC ont été utilisés lors de la construction de la banque, la dernière étape de filtrage consiste à regarder la corrélation entre le nombre théorique attendu d'ERCC et le nombre retrouvé dans les données afin de s'assurer de l'efficacité de la RT, ainsi que de l'efficacité de détection des ARNm en fonction de leur abondance. Généralement le seuil est fixé à une corrélation supérieure à 60%.

Les cellules qui présentent des valeurs aberrantes pour ces paramètres sont donc généralement exclues des analyses ultérieures. Cependant, les seuils doivent être fixés avec précaution, notamment pour éliminer le moins de cellules possibles en cas d'incertitude. Par exemple, il faut être vigilant si les données comportent des types cellulaires différents qui pourraient avoir des profils d'expression très inégaux. Ainsi, il est préférable de déterminer les seuils en fonction du ou des types cellulaires qui sont étudiés et donc de manière expérience-dépendante.

1.3.2.3 La normalisation

Une fois les données filtrées, l'étape suivante consiste à les normaliser (Figure 1.16 A). Les enjeux de la normalisation des données sont :

- 1) Éliminer la variabilité qui est liée à la technique employée. Elle peut venir, par exemple, de variations dans la profondeur de séquençage entre les cellules ou de différences dans l'efficacité de la RT. Ces différences même peu importantes doivent être atténuées pour que la variabilité restante soit majoritairement une variabilité biologique.

- 2) Stabiliser la relation moyenne - variance pour que les gènes les plus exprimés n'écrasent pas tout le signal.
- 3) Réduire les différences techniques entre les cellules (effet de différence de taille entre les cellules, position dans le cycle cellulaire ou encore les effets « batch » provenant d'expériences différentes ou de points de cinétique différents).
- 4) Transformer les données par une méthode dite monotone pour les rendre compatibles avec l'utilisation d'outils d'analyse en aval sans déformer les données.

Le choix de l'algorithme optimal pour la normalisation dépend beaucoup de la méthode de sc-RNA-seq utilisée [101].

L'une des méthodes la plus utilisée est SCTransform du package Seurat [102]. Cette méthode modélise le nombre d'UMI (à l'aide d'un modèle binomial négatif régularisé) pour éliminer la variation due à la profondeur de séquençage, et ajuste la variance entre les gènes ayant des abondances similaires. Cet algorithme peut aussi corriger par régression linéaire des effets techniques identifiés pour éliminer la variabilité liée aux facteurs techniques comme par exemple les effets plaques ou jours.

Les counts, c'est-à-dire les valeurs d'expression des gènes, sont ensuite log-transformés. La transformation logarithmique est une méthode de transformation monotone qui permet d'atténuer l'écart entre la moyenne et la variance et réduit l'asymétrie des données sans déformer le signal d'une cellule à l'autre [102]. Cette étape est particulièrement importante car beaucoup d'outils utilisés en aval pour l'analyse des données supposent que les données sont normalement distribuées.

Il faut noter que les gènes très peu exprimés sont particulièrement difficiles à normaliser. En effet, beaucoup d'algorithmes de normalisation génèrent des biais dans les counts normalisés pour ces gènes, notamment parce que le signal mesuré pour les gènes peu abondants est dominé par le bruit technique. Une solution à cela peut être de filtrer ces gènes et ne garder que les gènes présents à hauteur d'au moins un UMI par cellule en moyenne [103].

Enfin, un jeu de données de sc-RNA-seq peut contenir plusieurs milliers de gènes mais tous ne sont pas informatifs pour le processus investigué et ne contribuent pas à la variabilité biologique. Les gènes les plus variables (Highly Variable Genes) sont donc généralement sélectionnés et conservés pour les analyses ultérieures. Cette sélection s'opère en modélisant la relation moyenne-variance pour chaque gène à l'aide de régressions linéaires ou polynomiales, puis la variance

résiduelle est extraite, elle correspond à l'écart entre la variance estimée par le modèle (variance théorique) et l'observation (variance observée). Plus la variance résiduelle est élevée, plus le gène est variable. On garde en général entre 200 à 2000 gènes les plus variables.

1.3.2.4 Les outils d'analyse

Classiquement, l'analyse des données en cellule unique commence par une représentation des données en 2 dimensions afin d'apprécier leur structure générale (Figure 1.16 B). Puis en fonction des questions posées, on peut chercher s'il existe des sous-groupes de cellules à l'aide d'algorithmes de clustering et déterminer si ces sous-groupes ont des profils d'expression génique différents entre eux (Figure 1.16 C-D). L'analyse bio-informatique est en fait une succession de choix entre différents outils, chaque outil ayant des propriétés différentes et permettant de traiter de manière plus ou moins adaptée différents types de données et de répondre à différentes questions biologiques [104, 105].

Les outils de réduction de dimensions

Les données sc-RNA-seq sont multidimensionnelles par nature. Les algorithmes de réduction de dimensions intègrent la matrice d'expression multidimensionnelle (autant de dimensions que de gènes dans la matrice) dans un espace à faibles dimensions (2 ou 3). Le but est de capturer la structure sous-jacente des données dans un nombre de dimensions aussi réduit que possible de manière à pouvoir les visualiser et les analyser (Figure 1.16 B).

Il existe plusieurs algorithmes de réduction de dimensions, parmi les plus utilisés on retrouve l'ACP, t-SNE et UMAP [106].

L'Analyse en Composante Principale (ACP) est une approche linéaire qui génère des dimensions réduites en maximisant la variance appelées composantes principales (CP). Ces composantes principales sont ordonnées par la proportion de la variance totale qu'elles expliquent, et classiquement seules les deux ou trois premières composantes principales sont analysées [107]. En général, les premières CP capturent les facteurs dominants la variabilité des données.

L'approche t-distributed Stochastic Neighbor Embedding (t-SNE) permet de visualiser des données à grande dimension en donnant à chaque point (chaque cellule) un emplacement dans un espace à deux ou trois dimensions. Contrairement à l'ACP, il s'agit d'une approche non-linéaire où les objets similaires sont modélisés par des points proches dans l'espace et les objets différents sont modélisés par des points éloignés [108]. t-SNE calcule d'abord la distance entre une cellule

ANALYSE DES DONNÉES

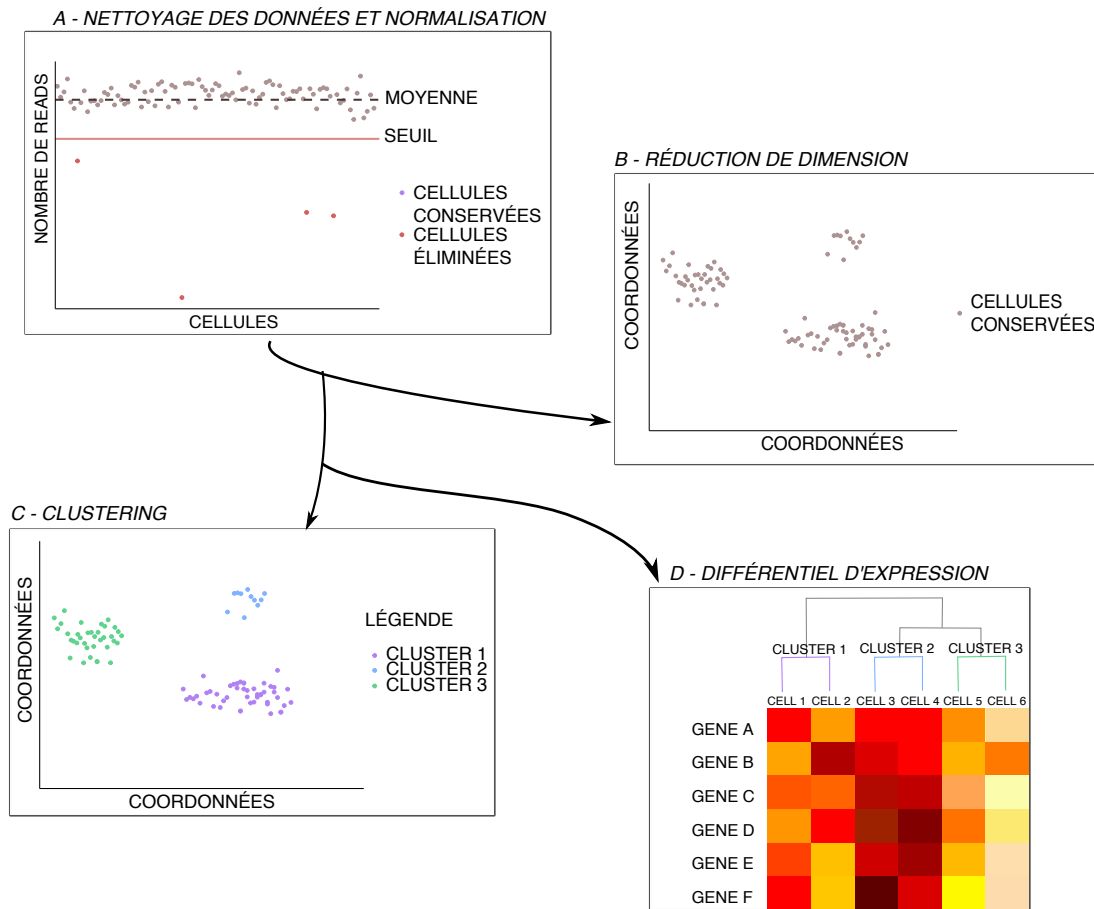


FIGURE 1.16 – Schéma général et simplifié des principaux outils d'analyse des données sc-RNA-seq.

La première étape de l'analyse des données de sc-RNA-seq consiste à filtrer et normaliser les données (A), puis appliquer une méthode de réduction de dimension (B), réaliser du clustering sur les données en dimensions réduites (C) et rechercher des profils d'expression génique différents entre les clusters (D).

et un nombre contrôlé de cellules voisines dans l'espace de grande dimension, puis l'algorithme s'assure que ces distances sont conservées lorsque les données sont déplacées point à point vers l'espace de plus faible dimension. Cependant, les distances entre des groupes de points éloignés ne sont pas nécessairement informatives.

L'approche Uniform Manifold Approximation and Projection (UMAP) est aussi une méthode de réduction de dimensions non linéaire [109]. UMAP commence par calculer un graphe représentant les données à grandes dimensions puis les intègre dans un espace à faibles dimensions.

Les approches non linéaires permettent de mieux représenter la structure locale des données dans un espace de dimensions réduites mais cela se fait au dépend de la représentation de la structure globale. Il est courant de d'abord réaliser une ACP puis une réduction non linéaire.

Les outils de clustering

Le clustering se fait en général sur les matrices des données après l'étape de réduction de dimensions. Cette analyse permet de regrouper les cellules sur la base de la similarité de leur transcriptome pour identifier des états cellulaires proches sans à priori sur l'expression de marqueurs spécifiques (Figure 1.16 C). Il est typiquement employé pour identifier des types cellulaires comme dans les projets Atlas. La similarité des profils d'expression est déterminée par des mesures de distances géométriques [110]. Très simplement, les valeurs d'expression génique pour une cellule sont utilisées comme des coordonnées pour la placer dans l'espace d'expression des gènes. Une fois toutes les cellules placées, il est possible de mesurer les distances entre chacune d'elles. Nous allons voir 3 types de clustering majoritairement utilisés.

Le clustering hiérarchique calcule des distances entre les cellules et les agglomère en groupes ou divise des groupes existants en plus petits groupes.

Le k-means identifie des centres de regroupement dans les données (les centroïdes), et affecte chaque cellule au centroïde le plus proche. Il teste itérativement k clusters et prend le nombre de clusters le plus optimisé. Cet algorithme est particulièrement utilisé pour identifier des types cellulaires rares.

La détection de communautés (k nearest neighbours graph) est un autre type de clustering qui est spécifiquement appliqué aux graphes. Au lieu d'identifier des groupes de points proches les uns des autres, la détection de communautés identifie des groupes de nœuds (les cellules) qui sont les plus densément connectés.

Cependant, l'un des défauts de la plupart des méthodes de clustering est qu'elles partitionnent toujours les données, qu'il existe vraiment ou non des groupes biologiques distincts. Si aucun groupe discret de cellules n'est présent dans les données (au sens biologique), le clustering n'est donc pas une méthode appropriée car elle induirait un biais d'interprétation en cherchant nécessairement des groupes.

Une méthode de clustering récente permet d'outrepasser cette difficulté. Il s'agit de l'outil PhiCLust qui estime la « clusterabilité » des données en prenant en compte le rapport signal sur bruit. En effet, si les données sont particulièrement bruitées, il y a une forte probabilité que les algorithmes de clustering classiques détectent plus de clusters que le nombre réel existant dans

les données [111].

Les outils de différentiel d'expression

L'analyse de différentiel d'expression (DE) permet de mettre en évidence des changements quantitatifs de l'expression génique entre différents groupes de cellules. Un grand nombre d'outils sont disponibles et proviennent soit de méthodes développées pour les analyses RNA-seq en population (bulk) soit de méthodes développées spécifiquement pour les données RNA-seq acquises en cellules uniques. Pour les gènes les plus exprimés, les méthodes « bulk » peuvent avoir des performances similaires à celles des méthodes en cellules uniques mais pour les gènes faiblement exprimés, leurs performances varient considérablement, notamment parce que les données en cellules uniques comportent beaucoup de 0 comme on l'a vu précédemment [112]. De manière générale, ces algorithmes vont comparer la distribution des gènes entre toutes les cellules de groupes différents deux à deux. Dans un premier temps, une hypothèse de distribution est posée (qui dépend de l'outil choisi), puis un test statistique est effectué pour tester cette hypothèse. Souvent, avec les données en cellules uniques, la distribution n'est pas connue, il est donc nécessaire d'utiliser une approche non-paramétrique, comme la méthode de Wilcoxon ou des approches basées sur des modèles (modèle Gaussien, Poisson...).

L'analyse de DE permet notamment d'identifier des gènes marqueurs au sein de groupes de cellules identifiés à l'aide du clustering (Figure 1.16 D).

1.3.2.5 Les métriques utilisées pour mesurer la variabilité de l'expression génique

L'apport majeur des approches de RNA-seq en cellule unique est la capacité à détecter la variabilité de l'expression génique. Mais paradoxalement, mesurer et surtout quantifier correctement cette variabilité n'est pas forcément évident. Les solutions à cette difficulté viennent de l'interdisciplinarité et de l'emprunt de différents types de métriques à d'autres disciplines scientifiques (physique et mathématiques particulièrement).

La variance est la mesure de la dispersion des échantillons autour de la moyenne. Autrement dit, la distance à la moyenne du groupe de cellules pour l'expression d'un gène donné. La variance est aussi utilisée par certains outils de différentiel d'expression.

Le coefficient de variation est le ratio entre la racine carré de la variance et la moyenne de la population. Plus la valeur du coefficient de variation est élevée, plus la dispersion autour de la

moyenne est grande [34]. Il est utilisé lorsque l'intensité du bruit est dépendante de l'intensité du signal. Le coefficient de variation est particulièrement utilisé pour estimer la variabilité d'origine extrinsèque [113].

Enfin, la surdispersion est une métrique définie par l'excès de variabilité observé dans les données par rapport à ce qui serait prédit par le bruit d'échantillonnage de Poisson (Poisson sampling noise) pour lequel la variance est égale et évolue avec la moyenne [114].

Ces méthodes ont en commun de mesurer la dispersion du signal, c'est-à-dire l'écart du nombre d'ARNm d'un gène donné, par rapport à la moyenne d'expression de toutes les cellules. Plus la dispersion est importante, plus le gène est variable.

Une autre méthode issue de la théorie de l'information prend plus en compte la distribution du signal. Il s'agit de l'entropie de Shannon qui permet de mesurer la quantité d'information portée par la distribution d'une mesure. Appliquée à l'analyse de la variabilité génique, si un gène est exprimé complètement aléatoirement et donc que sa distribution est uniforme dans une population de cellules, alors l'incertitude est maximale pour celui-ci et la valeur d'entropie également. A l'inverse, si l'expression d'un gène est plus informative, c'est-à-dire que sa distribution dans les cellules est plus dense en certains points, il est donc moins variable dans la population. En effet, sa variabilité est structurée dans la population et donc l'incertitude est moins importante et la valeur d'entropie aussi [115].

1.3.3 Les outils d'inférence à partir de données d'expression issues de cellules uniques

L'hétérogénéité cellulaire ne peut pas seulement être décrite par des analyses « statiques » comme la réduction de dimension ou le clustering. En effet, un des autres avantages des données en cellules uniques est que, bien que ces données soient elles-mêmes des photographies instantanées ou « snap shot » de l'expression génique de cellules à un temps t , il est possible de les utiliser pour faire des modèles stochastiques dynamiques des processus de décision cellulaire.

1.3.3.1 Les inférences de trajectoires

Les modèles d'inférence de trajectoires permettent de capturer les transitions entre les états cellulaires (par exemple au cours des processus de différenciation) à partir de données de sc-

RNA-seq temporelles. L'idée sous-jacente est que la différenciation s'accompagne de changements continus dans l'expression des gènes au cours du temps. Les algorithmes d'inférence de trajectoires vont trouver une trajectoire qui minimise les changements transcriptionnels entre cellules voisines. Les méthodes d'inférence de trajectoires diffèrent principalement par la complexité des chemins qui sont modélisés. Il faut prendre en compte la complexité des trajets à laquelle on s'attend pour choisir la méthode la plus adaptée aux données car aucune méthode ne donne des résultats optimaux pour tous les types de trajectoires [116].

1.3.3.2 Les inférences de réseaux de gènes

Il est possible de modéliser la dynamique des gènes qui entraînent les changements d'états des cellules. La régulation d'un gène passant par des interactions avec d'autres gènes, celles-ci peuvent être mises en évidence par les réseaux de régulation de gènes (GRN). Dans ce cadre spécifique, le terme « gène » ne désigne pas nécessairement la séquence d'ADN, mais peut aussi désigner les ARNm, ou souvent les protéines, et le terme « interaction » ne signifie pas forcément une interaction directe, mais plutôt une relation entre deux gènes qui peut comporter plusieurs intermédiaires et être de différentes natures.

Ces modèles peuvent permettre de poser et répondre à différentes questions, notamment les relations entre le « bruit » transcriptionnel et le GRN sous-jacent aux processus de décision cellulaire [117-119]. Ces modèles peuvent être ensuite validés par des expériences de perturbations *in vitro* ou *in vivo*.

1.3.4 Les apports des études en sc-RNA-seq pour étudier la variabilité de l'expression génique

Grâce aux données en cellules uniques, on peut implémenter les informations provenant de la dynamique des GRN au modèle de Waddington. L'état du GRN peut être représenté par le fait que le paysage où les cellules évoluent est en fait très dynamique. Lors des processus de décision, un ou des signaux extrinsèques modifient les paramètres et la structure du GRN se traduisant par un changement de la topologie du paysage. Ainsi, une cellule initialement bloquée dans un état attracteur défini par l'état du GRN va pouvoir sortir de son état suite au changement de configuration du réseau et potentiellement rejoindre un nouvel état attracteur. Pour reprendre la métaphore, l'état du GRN peut par exemple rendre les vallées plus planes ou les crêtes moins hautes, favorisant ainsi une exploration moins contrainte des cellules, puis finalement accentuer

les reliefs et bloquer les cellules dans un état différent de celui dans lequel elles étaient [120] (Figure 1.17).

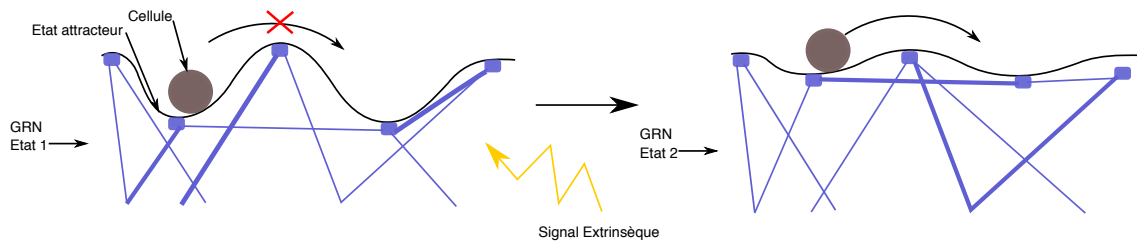


FIGURE 1.17 – Schéma du paysage de Waddington mis à jour.

L'état du GRN sous-jacent va changer suite à la réception d'un signal extrinsèque modifiant alors la topologie du paysage de Waddington. Ce changement de configuration du paysage peut permettre aux cellules d'accéder à d'autres états attracteurs.

Comme illustré dans les parties précédentes, la variabilité de l'expression génique dans le cadre des processus de décision cellulaire est étudiée à l'aide de différents outils dans de nombreux modèles biologiques. Parmi ces modèles, le système hématopoïétique est l'un des modèles eucaryotes de référence puisqu'il récapitule tous les processus de décision cellulaire (division, différenciation et mort programmée) et comporte plusieurs avantages techniques que nous allons voir [121].

1.4 L'hématopoïèse et l'érythropoïèse comme modèles de décision cellulaire pour étudier la variabilité de l'expression génique

Les globules rouges permettent l'acheminement des molécules d'oxygène aux différents tissus et organes des organismes. Ces cellules sont indispensables tout au long de la vie d'un organisme, de son développement embryonnaire jusqu'à sa mort. Leur production est donc continue et est de l'ordre de plusieurs millions de cellules par seconde. Les globules rouges proviennent de l'érythropoïèse qui est un des embranchements de l'hématopoïèse que nous allons détailler ci-dessous.

1.4.1 L'hématopoïèse chez les mammifères

1.4.1.1 L'origine embryonnaire des cellules hématopoïétiques

Les cellules hématopoïétiques sont produites initialement à des stades précoces du développement embryonnaire. Il existe deux types distincts de cellules hématopoïétiques embryonnaires, les cellules hématopoïétiques primaires et les cellules hématopoïétiques définitives (Figure 1.18).

Les cellules hématopoïétiques primaires dérivent des cellules du mésoderme. Ces cellules s'agrègent en îlots sanguins dans le sac vitellin, tissu extra-embryonnaire transitoire, et se différencient en cellules hématopoïétiques et en cellules endothéliales. Des études récentes montrent que ces deux types cellulaires peuvent ensuite se différencier en érythrocytes primitifs qui ont des caractéristiques phénotypiques distinctes des érythrocytes matures définitifs. Ils conservent leurs noyaux chez les mammifères et produisent les formes fœtales de l'hémoglobine [123].

Les cellules hématopoïétiques définitives apparaissent un peu plus tardivement lors du développement embryonnaire et proviennent d'une structure différente et indépendante des cellules hématopoïétiques primaires. Leur origine est intra-embryonnaire, elles proviennent de clusters de cellules hématopoïétiques qui sont produites au niveau de la structure aorte – gonade – mesonephros (AGM) [124].

Les cellules hématopoïétiques définitives ont des potentiels de différenciation plus importants que les cellules hématopoïétiques primaires. Elles sont multipotentes et peuvent générer les lignées érythroïdes, myéloïdes et lymphoïdes, alors que les cellules hématopoïétiques primaires sont restreintes à générer des progéniteurs érythroïdes et myéloïdes. Après la naissance, chez

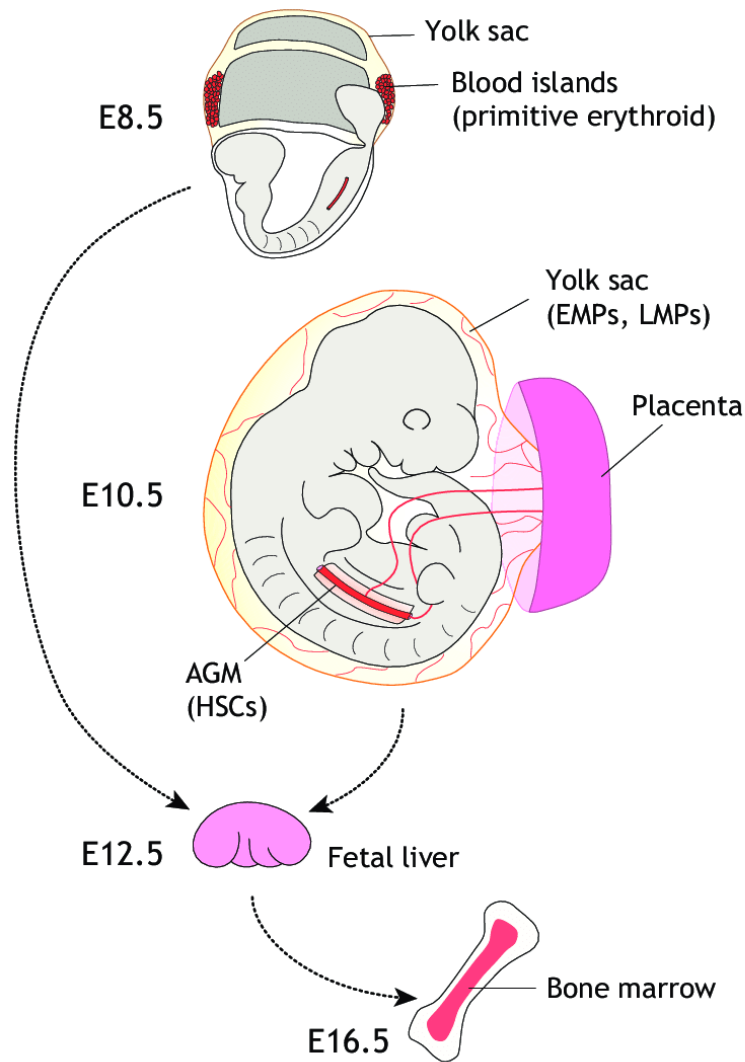


FIGURE 1.18 – Schéma de l'origine embryonnaire des cellules souches hématopoïétiques. Les cellules hématopoïétiques primaires dérivent de deux structures : l'une extra-embryonnaire, les îlots sanguins, l'autre intra-embryonnaire, la structure aorte – gonade – mesonephros (AGM). Les cellules provenant des deux structures vont ensuite coloniser les organes, thymus, foie et moelle osseuse, qui après le développement embryonnaire permettront la production continue des cellules souches hématopoïétiques. Schéma issu de Mevel et al. 2019 [122].

les mammifères, les organes qui produisent les cellules hématopoïétiques sont le foie, le thymus et la moelle osseuse. Pendant le développement embryonnaire, ces organes ne produisent pas de cellules hématopoïétiques *de novo* mais vont être colonisés par des progéniteurs sanguins extrinsèques provenant à la fois de l'hématopoïèse primaire et définitive [125, 126]. La contribution relative de chacun des progéniteurs au pool final de cellules souches hématopoïétiques (HSC) adultes reste à ce jour inconnue. Cependant, il est décrit que les érythrocytes provenant de l'hématopoïèse primitive ne persistent pas après le développement embryonnaire [127].

1.4.1.2 La description de l'hématopoïèse adulte chez les mammifères

L'hématopoïèse est le processus de génération de toutes les cellules sanguines. Les HSC adultes résident dans la moelle osseuse où elles sont capables de s'auto-renouveler de manière prolongée et de se différencier en une dizaine de types cellulaires différents [128] (Figure 1.19).

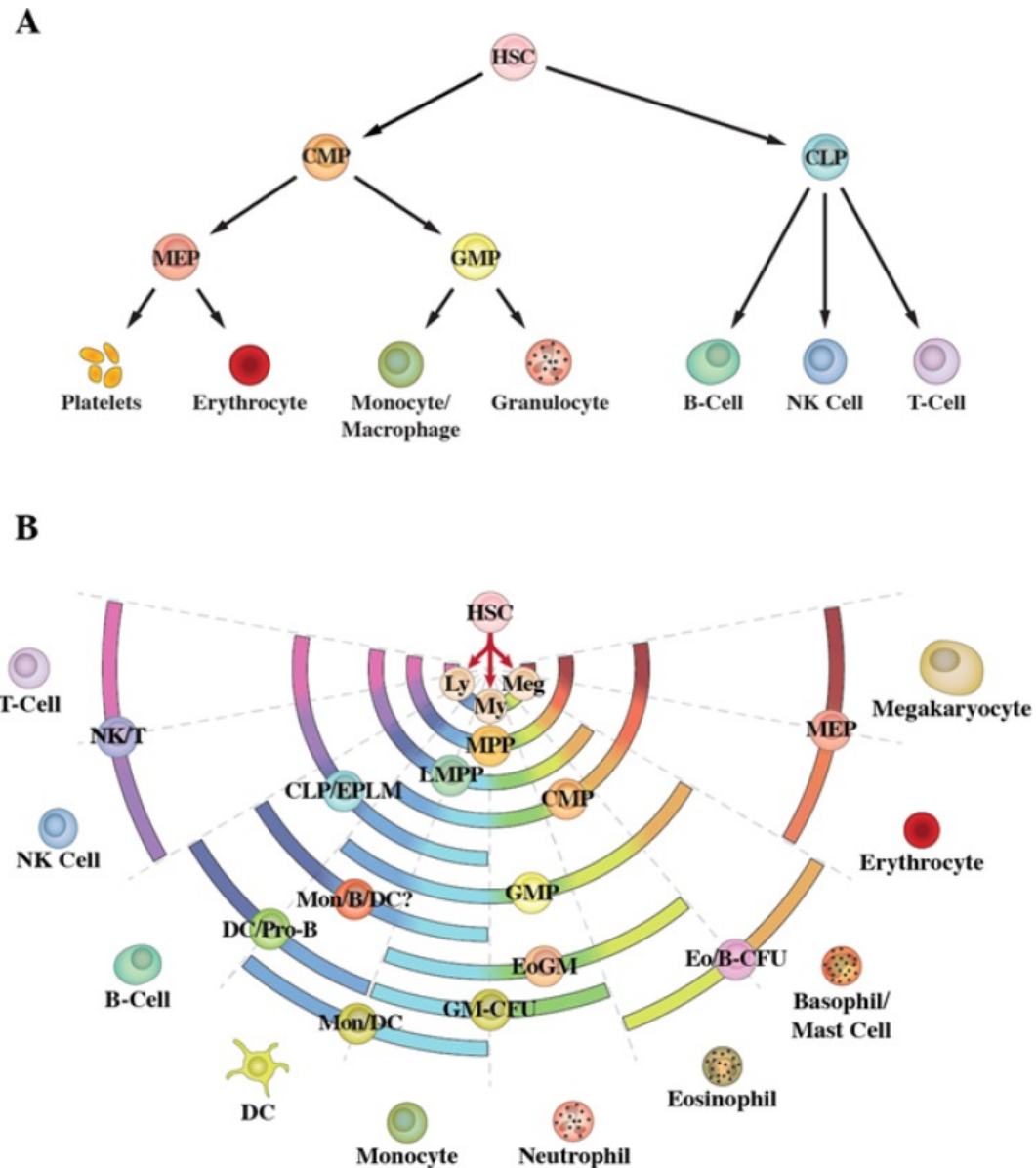


FIGURE 1.19 – Schéma simplifié de l'hématopoïèse adulte.

(A) Vision classique de l'hématopoïèse. Les HSC vont se différencier irrévocablement en précurseurs CLP (Common Lymphoid Progenitors) ou CMP (Common Myeloid Progenitors) qui vont à leur tour pouvoir se différencier dans les différents lignages des cellules sanguines : lymphoïde, myéloïde et érythroïde. (B) Vision contemporaine de l'hématopoïèse. Les cellules deviennent progressivement orientées vers la production d'un type cellulaire ou d'un autre en fonction de l'expression stochastique de facteurs de transcription et ce choix est renforcé par la présence et la réponse à des cytokines spécifiques. Schéma issu de Brown et al. 2015 [129].

Les LT-HSC (Long Term reconstituting HSC) sont les cellules qui ont la plus grande capacité

d'auto-renouvellement et qui peuvent générer les progéniteurs de toutes les lignées hématopoïétiques. Les HSC peuvent se différencier en CLP (Common Lymphoid Progenitors) ou CMP (Common Myeloid Progenitors), ce sont tous deux des progéniteurs oligopotents, c'est-à-dire partiellement engagés, restreignant les lignages qui peuvent en être dérivés. Les CLP après différenciation vont être à l'origine des lymphocytes T et B, des cellules NK (Natural Killer) et des cellules dendritiques lymphoïdes. Les CMP peuvent quant à eux, se différencier en deux types de progéniteurs encore plus engagés : 1) les GMP (Granulocyte-Myeloïdes Progenitors) qui donneront les basophiles, les éosinophiles, les neutrophiles, les monocytes, les macrophages et les cellules dendritiques myéloïdes ; 2) les MEP (Megacaryocytic-Erythroïdes Progenitors) qui engendreront les mégacaryocytes et les plaquettes d'une part et les globules rouges *via* l'érythropoïèse, d'autre part [130].

1.4.2 L'hématopoïèse des mammifères comme modèle de décision cellulaire

1.4.2.1 La pertinence du modèle

Le système hématopoïétique comme modèle d'étude des décisions cellulaires présente plusieurs d'avantages. Premièrement, il s'agit d'un tissu en partie liquide, les cellules sont donc facilement accessibles et un grand nombre de protocoles d'isolement des cellules bien établis sont disponibles. De plus, les cellules hématopoïétiques peuvent être cultivées *in vitro*.

Deuxièmement, le sang est l'un des tissus ayant le plus de capacité régénérative, le bon nombre de types cellulaires spécifiques devant être générés en permanence. C'est également un des tissus les plus plastiques, pouvant s'adapter efficacement aux conditions de l'environnement, par exemple en cas d'anémie ou d'infection [131].

Enfin, les trois types de décision cellulaire se produisent lors de l'hématopoïèse : 1) la division cellulaire, qui permet la prolifération et le maintien des HSC [128] ; 2) la différenciation de ces HSC qui ont le potentiel de générer tous les types cellulaires sanguins [123] et 3) la mort cellulaire programmée et la survie, modulée notamment par la voie Epo/EpoR [132], qui permet la régulation du nombre de HSC au niveau de la niche hématopoïétique [17].

Les perturbations dans les prises de décision des HSC sont à l'origine de troubles hématalogiques et il est donc essentiel, pour comprendre l'hématopoïèse normale et pathologique, d'étudier le contrôle moléculaire des différentes décisions que peuvent prendre ces cellules.

Enfin, ce modèle biologique est aussi utilisé pour établir des modèles computationnels qui

prennent compte des effets de la variabilité d'expression génique sur ces processus de décisions cellulaires [133].

1.4.2.2 Les modèles actuels de la différenciation hématopoïétique

Deux types de modèles ont été développés pour récapituler l'hématopoïèse. Un modèle instructif très hiérarchisé qui décrit la cascade de différenciation comme une perte successive des potentiels de lignage individuel par un processus de décision binaire ; dans ce modèle, les cytokines spécifiques auxquelles sont exposées les HSC sont le déterminant majeur de leur choix [134] (Figure 1.19 A). Et un modèle stochastique dans lequel les cytokines jouent un rôle majeur, mais ne spécifient pas directement le choix de lignage ; dans ce modèle, les cytokines permettent la survie et la prolifération des cellules qui finiront par entrer dans un lignage de manière aléatoire [135]. Appuyant ce deuxième modèle, il a été montré qu'*in vitro* des progéniteurs de toutes les lignées peuvent se développer à partir de HSC à une fréquence normale, même en l'absence de cytokines spécifiques à une lignée [136, 137] (Figure 1.19 B).

Il a été proposé que c'est l'expression stochastique des facteurs de transcription lignage-spécifique dans les HSC qui permet l'engagement initial dans une lignée, appuyé ensuite par des mécanismes d'activation et de répression de l'expression des gènes de la voie empruntée par les cellules [138] (Figure 1.19 B). En effet, les progéniteurs multipotents et les HSC peuvent exprimer des marqueurs de différentes lignées en même temps, bien que généralement à des niveaux faibles [139].

Cette vision est néanmoins discutée, notamment par Hoppe *et al.* qui suggèrent que le choix n'est pas simplement lié à l'expression stochastique d'un facteur de transcription en faveur d'un autre, mais qu'il s'agirait d'une régulation plus complexe impliquant d'autres facteurs de transcription et d'autres mécanismes de régulation qui restent à être découverts [140].

Il est important de souligner que ces modèles de la hiérarchie de la différenciation hématopoïétique ne reflètent que les connaissances actuelles et qu'ils continuent à évoluer au fil des nouvelles découvertes. En particulier, la première décision prise par les HSC qui mène à la spécialisation en CLP ou CMP est encore mal définie, notamment à cause de notre définition biologique du type cellulaire qui s'appuie sur la présence ou l'absence d'expression concomitante de différentes protéines membranaires détectables par les méthodes actuelles, particulièrement le FACS [141]. Cette définition est aujourd'hui remise en question principalement grâce aux études

menées à l'échelle de la cellule unique qui révèlent une forte hétérogénéité de l'expression génique au sein des HSC [142]. De plus, plusieurs études récentes démontrent que l'hématopoïèse est un processus très continu et non pas constitué d'un enchaînement d'étapes discrètes [143, 144].

D'autre part, l'hématopoïèse chez les mammifères est un processus particulièrement complexe, qui implique non seulement les HSC mais également l'environnement où ces cellules évoluent appelé niche hématopoïétique qui participe fortement à la plasticité du système [128].

Les connaissances acquises sur la régulation de l'hématopoïèse grâce à l'utilisation de modèles animaux a permis de développer un modèle cellulaire primaire humain, les cellules CD34+. Ces cellules sont récoltées à partir de sang de cordon ombilical humain et peuvent ensuite être cultivées *in vitro* [145]. Mais ce système cellulaire possède plusieurs limitations. Notamment, les CD34+ consistent en un mélange hétérogène de cellules souches et de progéniteurs hématopoïétiques et les cellules ne peuvent pas être maintenues en état d'auto-renouvellement *in vitro*. Ces limitations sont entre autres liées au fait que comme décrit ci-dessus les premières étapes de la différenciation hématopoïétiques sont encore mal comprises.

Un autre modèle de différenciation qualifié de plus simple mais tout aussi complet que l'hématopoïèse est l'un de ses embranchements, l'érythropoïèse et particulièrement l'érythropoïèse aviaire comme nous allons le voir.

1.4.3 L'érythropoïèse comme modèle de décisions cellulaires

1.4.3.1 L'érythropoïèse humaine

Les cellules de sang de cordon CD34+ peuvent être différenciées *in vitro* dans la voie érythrocytaire et générer des globules rouges [146]. Chez l'homme les progéniteurs les plus précoces identifiés *ex vivo* sont les BFU-E (Burst Forming Unit-Erythroid). Ils se différencient ensuite en CFU-E (Colony Forming Unit-Erythroid) [147]. Les CFU-E font 3 à 5 divisions en 2 à 3 jours au cours desquelles ils se différencient. Cette différenciation est caractérisée par des changements morphologiques et moléculaires majeurs, notamment une diminution de la taille des cellules, une condensation de la chromatine et une hémoglobination, qui va ensuite aboutir à leur énucléation et à l'expulsion d'autres organites. Après expulsion du noyau, ces cellules sont appelées les réticulocytes (Figure 1.20 A). Enfin, les réticulocytes sont libérés de la moelle osseuse dans la circulation sanguine, et deviennent des érythrocytes circulants qui transportent l'oxygène dans

le sang périphérique [148].

L'érythropoïèse humaine *in vitro* est majoritairement utilisée pour produire des globules rouges à l'aide d'un cocktail de cytokines assez complexe contenant notamment de l'EPO [149]. Mais les limites majeures de l'érythropoïèse humaine *in vitro* comme modèle cellulaire sont la difficulté d'obtenir des cinétiques de différenciation reproductibles, de maintenir les progéniteurs érythrocytaires dans un état d'auto-renouvellement et la complexité du signal de différenciation [150].

L'utilisation de l'érythropoïèse aviaire comme modèle permet de résoudre une partie de ces difficultés, comme nous allons le voir.

1.4.3.2 Le poulet un organisme modèle

Le poulet est un organisme modèle en biologie [151] particulièrement utilisé pour étudier les processus de décision cellulaire [1, 2] et notamment l'hématopoïèse car ce processus est conservé chez tous les vertébrés [152, 153]. Il s'agit d'ailleurs d'un des organismes modèles pionniers dans ce domaine, en particulier pour étudier l'hématopoïèse embryonnaire de part l'accessibilité de l'embryon dans l'œuf [154]. Des expériences de greffe de cellules d'embryons de poussins sur le sac vitellin d'embryons de cailles ont fourni les premières preuves que l'hématopoïèse primaire ne contribue pas à l'hématopoïèse définitive [155]. De plus, ce modèle a donné les premières évidences d'une origine intra-embryonnaire de l'hématopoïèse définitive [156], origine similaire constatée par la suite chez les mammifères [157].

1.4.3.3 L'érythropoïèse chez le poulet

L'érythropoïèse aviaire suit la même dynamique que l'érythropoïèse humaine (Figure 1.20 A). La différence majeure entre l'érythropoïèse humaine et aviaire est que les érythrocytes matures aviaires n'expulsent pas leur noyaux (Figure 1.20 B). Bien qu'au stade érythrocyte mature, ces noyaux soient transcriptionnellement inactifs, cette caractéristique en fait un très bon modèle pour étudier les changements de structure de la chromatine [151].

L'érythropoïèse aviaire se révèle être un modèle particulièrement intéressant. Il s'agit du premier système de différenciation primaire érythropoïétique à avoir été développé *in vitro* et dont la différenciation est monolignage. L'érythropoïèse aviaire a également été très étudiée dans le cadre de la transformation leucémique des cellules de poulet par le Avian Erythroblastosis Virus

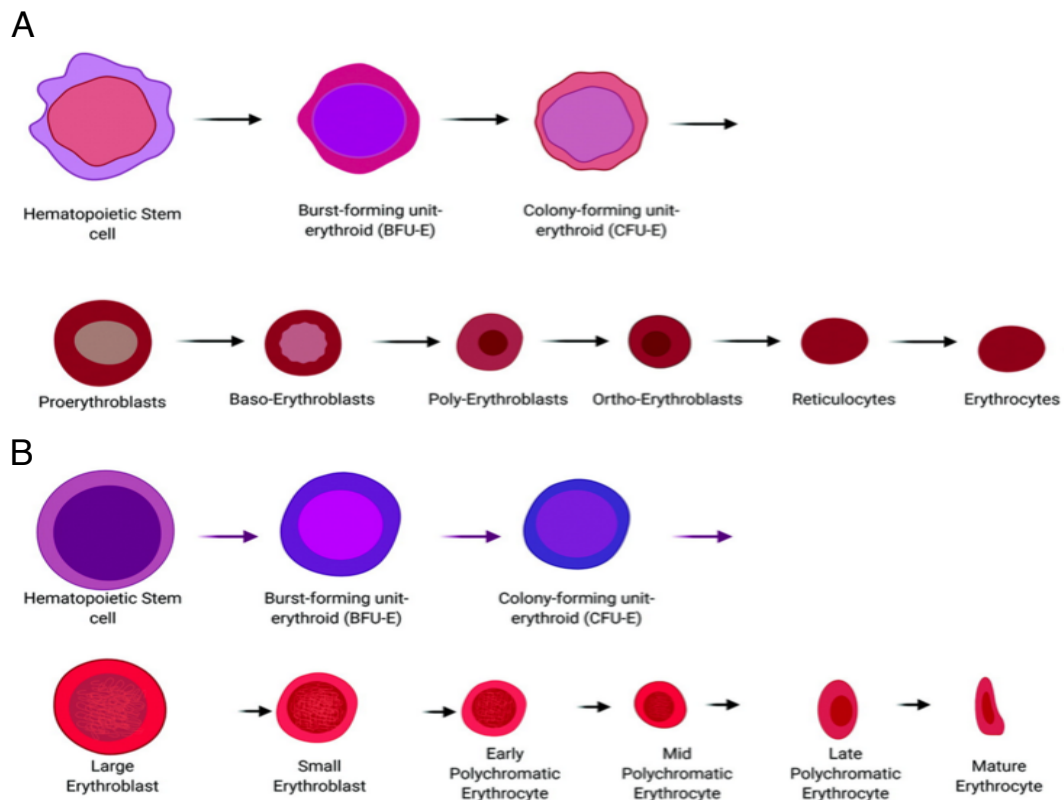


FIGURE 1.20 – Schéma simplifié de l'érythropoïèse.

Erythropoïèse humaine (A) et aviaire (B) détaillées dans le texte. Schéma issu de Beacon and Davie 2020 [151].

(AEV) qui a permis d'acquérir des connaissances sur les voies de signalisation et la régulation transcriptionnelle régissant la différenciation érythrocytaire [130].

1.4.3.4 Le modèle cellulaire aviaire T2EC

Un modèle biologique robuste consistant en des progéniteurs érythrocytaires aviaires, appelé T2EC (TGF- α /TGF- β -induced erythrocytic cells), a été établi dans l'équipe pour étudier la différenciation érythrocytaire [158]. Chez le poulet, il est en effet possible de récupérer des progéniteurs partiellement engagés à partir de moelle osseuse d'embryons. Ces progéniteurs peuvent être maintenus *in vitro* en état d'auto-renouvellement en présence de TGF α , TGF β et de dexaméthasone, et peuvent être induits à se différencier en érythrocytes matures par un changement du milieu contenant de l'insuline et du sérum anémié de poulet [158, 159].

Dans le cadre de l'utilisation de l'érythropoïèse comme système d'étude de la différenciation, ce modèle présente plusieurs avantages : il s'agit d'un système biologique primaire non modifié génétiquement dont la différenciation est inductible et mono-lignage. La cause première de la

différenciation est donc connue et est permise par la modification de l'information contenue dans l'environnement extracellulaire. De plus, la différenciation en un seul lignage affranchit de la variabilité induite par l'engagement des cellules dans différents lignages. Ce système est donc parfaitement adapté à l'étude de la variabilité dans la prise de décision cellulaire.

1.4.4 La variabilité de l'expression génique lors de la différenciation érythrocytaire aviaire et la différenciation hématopoïétique humaine

Une étude précédemment réalisée dans l'équipe a permis de mettre en évidence le rôle de la stochasticité de l'expression génique dans le processus de différenciation érythrocytaire aviaire [160]. Des progéniteurs érythrocytaires aviaires (T2EC) ont été induits à se différencier pendant 48h, et une fraction des cellules a été collectée à 6 temps différents lors de cette cinétique pour analyser leurs transcriptomes à l'échelle de la cellule unique par scRT-qPCR. L'étude de la dynamique transcriptionnelle de 92 gènes potentiellement impliqués dans la différenciation à l'échelle de la cellule unique a permis de mettre en évidence que la différenciation érythrocytaire s'accompagnait d'une augmentation significative et transitoire de la variabilité de l'expression génique, quantifiée par l'entropie de Shannon. La variabilité atteint un maximum entre 8h et 24h post-induction de la différenciation, avant de décroître (Figure 1.21). Ces résultats suggèrent un point critique de décision autour de 24h qui est par ailleurs accompagné de l'engagement irréversible des cellules dans la différenciation à 48h. En effet, après 24h de différenciation, les cellules peuvent se remettre à proliférer lorsqu'elles sont replacées dans du milieu d'auto-renouvellement, ce qui n'est pas le cas après 48h de différenciation, les cellules ayant perdu leur capacité à se diviser.

Pour définir si cette variabilité de l'expression génique avait un rôle moteur dans la différenciation érythrocytaire, l'équipe a ensuite réalisé une autre étude visant à moduler la stochasticité de l'expression génique à l'aide de drogues et à observer leurs effets sur la dynamique et le taux de différenciation des progéniteurs érythrocytaires aviaires [161]. Dans cette étude, la variabilité de l'expression génique a également été quantifiée par l'entropie de Shannon. Le traitement des cellules par deux drogues, l'Artemisinine et l'Indométhacine, entraîne une diminution de la valeur d'entropie indiquant une baisse de la variabilité intrinsèque de l'expression génique. Lorsque des progéniteurs érythrocytaires sont induits à se différencier en présence de ces drogues, leur taux de différenciation est significativement plus faible que celui des cellules contrôles. À l'inverse, des cellules cultivées en présence de MB3, drogue qui augmente la variabilité de l'expression génique

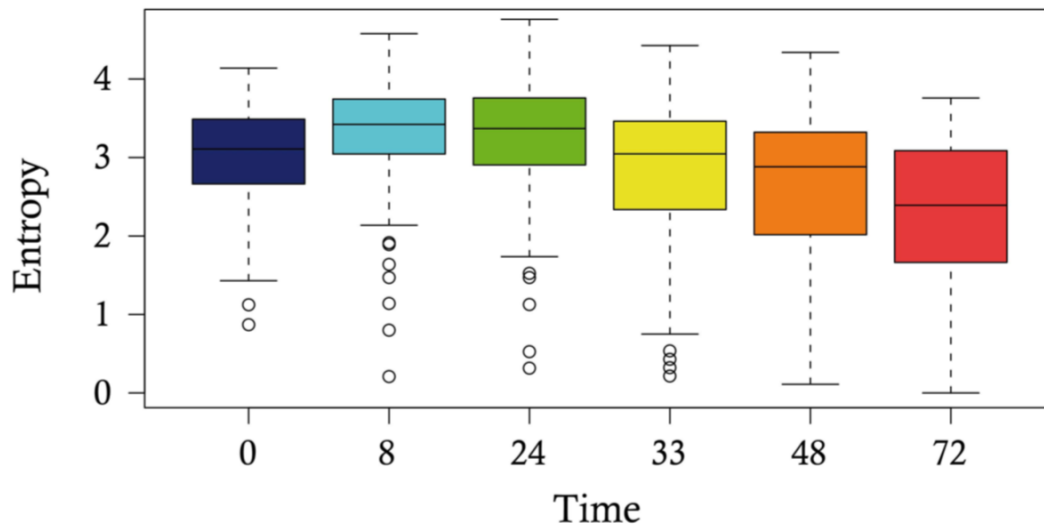


FIGURE 1.21 – *Mesure de l’hétérogénéité inter-cellulaire des progéniteurs érythrocytaires aviaires à l’aide de l’entropie de Shannon.*

L’entropie de Shannon a été calculée pour chaque point de temps et pour chaque gène. Les boxplots représentent la distribution des valeurs d’entropie des gènes dans chaque groupe de cellules. Figure issue de Richard et al. 2016 [160].

[70], présentent alors un taux de différenciation supérieur à celui des cellules contrôles.

Une augmentation transitoire de la variabilité de l’expression génique a également été décrite dans le modèle d’hématopoïèse humaine, lors des étapes très précoces de décision des HSC. Des cellules souches hématopoïétiques de sang de cordon placées en culture en présence de cytokines spécifiques voient leur profil d’expression génique changer aléatoirement pendant deux jours, jusqu’à l’émergence de deux profils transcriptomiques distincts qui se traduisent aussi par deux phénotypes différents. Il est très intéressant de noter que certaines cellules, décrites comme hésitantes, changent plusieurs fois de phénotypes. L’analyse transcriptomique impliquant le tri en FACS puis la lyse des cellules l’information de la morphologie n’a pas pu être directement corrélée à l’état transcriptomique des cellules. Ainsi, il n’a pas été possible aux auteurs de démontrer avec certitude que ces fluctuations transcriptomiques étaient accompagnées de fluctuations de phénotypes. Néanmoins, il ne serait pas étonnant que les deux soient liés [162].

Confirmant ces résultats, une autre étude utilisant cette fois-ci des données de sc-RNA-seq issues de cellules hématopoïétiques humaines, prélevées à des stades un peu plus tardifs de décision (CLP, CMP et MEP, GMP voir figure 1.19), a mis en évidence un pic de variabilité de l’expression génique mesuré également par l’entropie de Shannon. Les auteurs ont montré que l’augmentation de l’entropie est majoritairement portée par les gènes codant pour les facteurs de

transcription lignée-spécifique, sans que ceux-ci ne soient pour autant les plus différenciellement exprimés [137].

Ce pic de variabilité est également observé dans d'autres modèles de différenciation [163-165]. L'ubiquité de ce phénomène lors des processus de différenciation supporte ainsi l'hypothèse que la variabilité de l'expression génique est l'un des moteurs de la différenciation et qu'elle pourrait contribuer à faire sortir les cellules de leur état d'auto-renouvellement [166].

Paradoxalement, les études en cellule unique ont également mis en évidence que certains gènes, très variables dans une population de cellules isogéniques, sont soumis à une forme de mémoire transcriptionnelle. Cette mémoire transcriptionnelle, dans ce contexte, est caractérisée par une hérédité des niveaux de fluctuation de manière transcrit spécifique entre des cellules apparentées. Cette mémoire transcriptionnelle fait partie de la mémoire non génétique.

1.5 La mémoire non génétique

1.5.1 La description de la mémoire non génétique

La mémoire cellulaire est bien plus que la copie exacte du matériel génétique de la cellule mère lors de la mitose. En effet, il existe une mémoire non génétique ou épigénétique, qui est la transmission d'une information entre des générations de cellules sans que cette information ne soit portée par la séquence d'ADN. Plus précisément, c'est un changement dans l'expression d'un gène qui n'implique pas une mutation génétique et qui est hérité en l'absence du signal qui a initié ce changement d'expression. Cette mémoire, comme la variabilité de l'expression génique, est mesurable à l'échelle de la cellule unique.

Par exemple, Kaufmann *et al.* ont génétiquement modifié la levure *S. cerevisiae* pour qu'elle exprime une version modifiée de la voie GAL où il n'y a pas de boucle de rétro-action négative. Dans ce modèle, les cellules transitionnent aléatoirement entre deux états de la voie au cours du temps, actif et inactif et ce en l'absence de signaux extracellulaires. Ces transitions sont peu fréquentes et résultent de fluctuations stochastiques des concentrations des protéines régulant la voie. Les états actifs et inactifs sont détectés grâce à des rapporteurs fluorescents en aval de la voie. Les résultats de leurs travaux ont montré que lorsqu'une cellule se divise, les cellules filles adoptent à la fois l'état d'expression de la cellule mère au moment de sa division (actif ou inactif) mais surtout, elles vont passer d'un état à l'autre de manière synchronisée au cours du temps. Cette mémoire du timing de transition est par ailleurs maintenue sur plusieurs générations cellulaires. Ces résultats sont particulièrement intéressants puisque si on observe les dynamiques de transition sans prendre en compte les informations de généalogie, elles paraissent entièrement stochastiques [167].

Un des modèles de prise de décision des lymphocytes, le ACH (Autonomous Competition Hypothesis), propose que les choix de se diviser, se différencier ou mourir sont des décisions indépendantes et que chaque processus est en compétition avec les autres. Beaucoup d'études *in vitro* confirment ce modèle qui récapitule très bien l'hétérogénéité des décisions observée au sein des populations de lymphocytes. Récemment, des études ont examiné les effets de parenté sur ces prises de décisions. Les résultats de ces études sont particulièrement intéressants puisqu'ils montrent que le timing de décision est corrélé entre cellules soeurs, même si leur décision est différente, comparé à des cellules non apparentées. En d'autres termes, deux cellules soeurs vont prendre leur prochaine décision à des temps similaires alors que la nature de la décision peut

être différente, l'une pouvant se diviser et l'autre pouvant mourir [168].

La mémoire non-génétique est aussi observée *in vivo* lors du développement d'un organisme. Ferraro *et al.* ont suivi l'activité de transgènes exprimés de manière stochastique dans des cellules d'embryons de drosophiles. Ces transgènes peuvent être actifs (transcrits) ou inactifs dans la cellule mère. L'état d'expression des transgènes a ensuite été observé dans les cellules filles après la mitose. Les résultats montrent que la probabilité d'une réactivation rapide de ces transgènes après la mitose est 4 fois plus élevée lorsque la mère avait une transcription active du transgène [169].

Ces exemples mettent en évidence la transmission de profils d'expression génique sans modification génétique, et donc l'existence d'une mémoire non génétique lors de processus biologiques.

1.5.1.1 La mémoire au niveau de la chromatine

La mémoire non génétique est beaucoup étudiée à l'échelle de la chromatine, particulièrement la transmission « d'états chromatiniens » qui concernent notamment la transmission de modifications épigénétiques.

Un des exemples les plus évidents de transmission d'états chromatiniens est l'inactivation des chromosomes. Par exemple, lors de l'inactivation du chromosome X, des longs ARN non-codants (lncRNA) appelés Xist vont se fixer aléatoirement sur un des deux chromosomes X. Ces ARN vont ensuite recruter PRC2. PRC2 est un complexe protéique qui triméthyle un résidu d'histone (H3K27), cette tri-méthylation induit la compaction de la chromatine où H3K27me3 est présent et entraîne donc la répression des gènes du chromosome inactivé. Ce processus de mémoire est très stable et est propagé au cours des générations cellulaires, si bien qu'une fois l'un des deux chromosomes inactivé, c'est toujours le même chromosome X qui sera inactivé au cours des divisions cellulaires [170].

Une autre marque épigénétique transmissible est la méthylation de l'ADN. Des groupements méthyles sont associés à la molécule d'ADN et peuvent modifier l'activité d'un segment d'ADN sans en changer la séquence. Il s'agit de modifications covalentes, donc très stables, de la chromatine. Ces groupements méthyles sont maintenus pendant la réplication de l'ADN. Puis, la

méthylation est ensuite propagée post-réplication par l'enzyme DNMT1 méthylant les CG non méthylées qui sont situés face des CG hémi-méthylés (meCG/GCme) [170]. Ainsi la méthylation de l'ADN est propagée de manière clonale. Néanmoins, le rôle de la méthylation de l'ADN dans la régulation de la mémoire d'expression génique clonale au sein de populations cellulaires n'est pas très clair [171].

Enfin les modifications de résidus d'histones peuvent aussi être transmises au cours des générations cellulaires. Les résidus d'histones peuvent porter différentes marques (phosphorylation, méthylation, acétylation....) qui vont participer à moduler le degré de condensation de la chromatine.

Lors de la réplication de l'ADN, les histones sont évincées à mesure que la fourche de réplication avance. Après la réplication, ces histones vont être recyclées et repositionnées sur la chromatine et de nouvelles histones sans marques, vont être insérées en plus. Les histones recyclées conservent leurs modifications et les nouvelles histones environnantes vont acquérir les mêmes marques selon deux modèles, soit par auto-propagation des marques soit parce que des enzymes vont être recrutées au niveau des histones déjà modifiées et vont pouvoir modifier les nouvelles histones de la même manière [170].

Conceptuellement, les modifications d'histones ne sont pas héréditaires notamment parce que les enzymes qui les modifient ne sont pas spécifiques à des histones particulières ou certains locus où se trouvent ces histones, mais aux résidus portés par les histones. Il n'y a pour l'instant pas de preuve expérimentale que les modifications d'histones peuvent s'auto-propager de cellule mère à cellule fille sur le même locus génétique. Néanmoins des corrélations ont été observées entre le maintien des niveaux de l'expression de certains gènes entre des cellules apparentées et l'enrichissement de certaines marques épigénétiques [172, 173].

Les motifs de modification d'histones sont souvent très corrélés à la régulation de l'expression des gènes. En effet, les histones participent à la régulation de la transcription en rendant la chromatine plus ou moins accessible, en fonction des marques qu'elles portent, à d'autres protéines se fixant à l'ADN qui vont elles mêmes favoriser le recrutement de facteurs de transcription et autres complexes protéiques régulant la transcription.

Aujourd'hui le terme « épigénétique » est presque systématiquement associé aux modifica-

tions épigénétiques, c'est-à-dire aux modifications bio-chimiques de l'ADN qui induisent des changements de la structure de la chromatine. Bien que ces modifications soient pour certaines corrélées à des mécanismes de mémoire, la mémoire non génétique et donc le terme épigénétique dans sa définition initiale est plus large et englobe tous les phénomènes qui participent à la transmission des niveaux d'expression d'un gène [174].

1.5.1.2 La mémoire au niveau des ARNm

La mémoire non génétique peut aussi être évaluée au niveau du transcriptome, c'est la mémoire transcriptionnelle. Il s'agit de la transmission de niveaux d'expression de gènes qui est médiée ou non par des changements de la chromatine.

Dans une étude récente, Phillips *et al.* ont étudié la propagation de l'activité transcriptionnelle de 9 gènes rapporteurs à courte durée de vie, dans le temps et au cours de générations cellulaires dans des cellules embryonnaires de souris phénotypiquement homogènes. Les analyses ont été faites par microscopie en temps réel à l'échelle de la cellule unique. Les fluctuations de l'activité transcriptionnelle des transgènes est variables entre les cellules de la population. Cependant, l'activité transcriptionnelle entre des cellules soeurs est très fortement corrélées et est synchronisée, c'est-à-dire que le niveau d'expression des gènes rapporteurs est similaire entre les cellules soeurs et co-varie au cours du temps. La corrélation de l'expression des gènes est également transmise de cellule mère aux cellules filles et pour certains gènes elle perdure jusqu'à une quinzaine de générations cellulaires [175]. Cette étude met en évidence une mémoire transcriptomique clonale et gène-spécifique.

A plus haut débit, Mold *et al.* ont analysé par Smart-seq3 les transcriptomes de cellules uniques issues de clones de cellules T mémoire humaines. Ils ont montré que chaque clone issu d'une même cellule T mémoire présente une signature transcriptionnelle distincte et que ces signatures transcriptionnelles sont hérissables. Ils suggèrent que cette mémoire est probablement liée à des profils d'accessibilité à la chromatine variable d'un clone à l'autre [176].

Enfin, la mémoire des niveaux de l'expression génique se répercute également sur les niveaux des protéines impliquant d'autres mécanismes moléculaires que ceux décrits ci-dessus.

1.5.1.3 La mémoire au niveau des protéines

Dans une étude investiguant la mémoire des niveaux de protéines, des tags endogènes YFP (Yellow Fluorescent Protein) ont été intégrés au niveau de différentes protéines dans une lignée de cellules humaines. Le niveau des protéines a été suivi en microscopie à fluorescence en time-lapse. Les résultats montrent que les niveaux de certaines protéines taggées sont transmis de cellules mères aux cellules filles et sont maintenus pendant plusieurs générations cellulaires. Ces variations héréditaires qualifiées de « bruit à mémoire longue » par les auteurs possèdent une autre propriété intéressante : le niveau des protéines est corrélé entre des protéines appartenant aux mêmes fonctions, par exemple les protéines ribosomales [177].

Pour identifier les mécanismes impliqués dans ce type de mémoire, Corre *et al.* ont comparé des clones de cellules qui avaient été transfectés avec des transgènes générant des ARNm et des protéines fluorescentes très stables avec des clones transfectés avec des transgènes générant des ARNm et des protéines peu stables. Les cellules mères des clones exprimant les transgènes produisant des ARNm et des protéines stables engendrent des cellules filles avec un niveau similaire à la cellule mère de protéines fluorescentes ; au contraire, dans les clones exprimant les transgènes produisant des ARNm et des protéines à demi-vies courtes, le niveau de protéines n'est pas conservé entre cellule mère et cellules filles. Les auteurs ont montré que la dynamique des taux de dégradation des ARNm et des protéines est corrélée entre des cellules apparentées. De plus, ces taux expliquent partiellement comment la mémoire est transmise, donc *via* un mécanisme de mémoire passif qui correspond à un délai de dégradation [68].

Au niveau mécanistique, ces corrélations dans les niveaux et dynamiques d'activité transcriptionnelle et protéique pourraient en partie être causées par la transmission de facteurs qui contrôlent la dynamique de l'expression des gènes considérés de la cellule mère aux cellules filles. Mais bien que la répartition des molécules lors de la division cellulaire soit approximativement binomial, les corrélations entre les nombres d'ARNm et les nombres de protéines sont plus faibles au début du nouveau cycle cellulaire parce que la division cellulaire a tendance à randomiser les concentrations relatives des molécules [63]. Il apparaît donc que la mémoire non génétique est très probablement maintenue par une combinaison de mécanismes passifs et actifs. Parmi les mécanismes passifs, pourraient intervenir comme on l'a vu, la partition des molécules ou les temps de demi-vie des ARNm et des protéines. Cependant, les temps de demi-vie des ARNm et

des protéines peuvent être modifiés au cours d'un processus biologiques ; ces changements des temps de demi-vie pourraient faire intervenir des mécanismes actifs et donc de manière gènes spécifiques [178].

Enfin, d'autres mécanismes actifs impliqueraient des facteurs cis-régulateurs ou trans-régulateurs qui pourraient aussi intervenir de manière gène-spécifique.

1.5.1.4 La compréhension de la dynamique de l'expression génique des cellules dans le contexte de la multicellularité

La stochasticité de l'expression génique peut permettre de générer des états transcriptionnels différents ce qui permet notamment une flexibilité pour s'adapter à l'environnement. Ensuite, ces états transcriptionnels pourraient être maintenus clonalement par un mécanisme de mémoire épigénétique et notamment transcriptionnelle.

La mémoire transcriptionnelle pourrait donc être impliquée dans divers processus biologiques pour propager des états transcriptionnels dans des populations de cellules. Par exemple, lors du développement embryonnaire chez la souris, les cellules de l'épiblaste expriment des niveaux variables du gène *myc*, et par compétition cellulaire seules les cellules produisant les plus hauts niveaux de *Myc* survivent. La raison pourrait être que ces hauts niveaux d'expression de *Myc* leur procurent une activité anabolique plus élevée, mieux adaptée pour la suite du processus de développement [179].

Il a aussi récemment été montré qu'il existe une transmission clonale d'un fort niveau d'expression d'un gène codant pour un récepteur permettant la résistance « non génétique » des cellules de mélanomes à certaines drogues [172, 180]. Une autre étude suggère que la mémoire transcriptionnelle permet, parallèlement aux modifications génétiques mais à une échelle de temps plus rapide, de diversifier les programmes transcriptionnels qui favorisent par exemple l'EMT (Transition Épithélio-Mésenchymateuse) dans les populations de cellules cancéreuses [181].

Ainsi, étudier la mémoire transcriptionnelle sur de très petits clones de cellules semble être l'échelle d'étude la plus adaptée pour comprendre les mécanismes sous-jacents à la mémoire et potentiellement mieux comprendre son rôle dans les processus biologiques.

1.5.2 La mémoire transcriptionnelle et mémoire de l'état cellulaire

La mémoire transcriptionnelle peut aussi être définie comme la mémoire d'un état transcriptomique global et apparaît essentielle à l'adaptation et à la plasticité cellulaire. Cette mémoire d'un état transcriptomique par lequel la cellule est passée peut permettre aux cellules de répondre plus rapidement à des stimuli déjà rencontrés ou de permettre aux cellules de conserver une certaine plasticité de manière à pouvoir différencier des fluctuations environnementales d'un « vrai » signal (comme par exemple un signal de différenciation). Cette forme de mémoire transcriptionnelle est particulièrement décrite chez les eucaryotes unicellulaires.

1.5.2.1 Le priming

Chez *S. cerevisiae*, l'exposition antérieure à différentes sources de carbone ou à des conditions de stress induit un état « primé », c'est-à-dire une mémoire transcriptionnelle qui permettra aux cellules de répondre plus rapidement si elles rencontraient à nouveau la même source de carbone ou stress. Une étude a montré que cette mémoire transcriptionnelle est médiée notamment par des changements d'interactome des exosomes, c'est-à-dire un changement des protéines qui interagissent avec les exosomes, résultants en des changements des taux de dégradation des ARNm cytoplasmiques dans les cellules « primées » [182].

Dans les cellules de mammifères, cette mémoire d'état fait sens lors de phénomènes de plasticité cellulaire. Par exemple, l'analyse transcriptomique à haut débit de cellules précurseurs de la crête neurale (CNCC), qui sont des progéniteurs multipotents, montrent qu'ils présentent des signatures moléculaires reflétant à la fois la réactivation transitoire d'un état de pluripotence (notamment via l'expression de *OCT4*) et l'orientation vers le type cellulaire « crête neurale ». Les auteurs suggèrent qu'il s'agit d'un « priming » des régions régulatrices distales des gènes qui seront exprimés lors du développement de l'ecto-mésenchyme, dérivant des cellules de la crête neurale [183]. Ce système de « priming » des régions régulatrices des gènes lors du développement a également été observé dans d'autres systèmes biologiques [184].

D'autre part, dans les lymphocytes T mémoire, suite à la ré-exposition à un antigène déjà rencontré, les gènes précédemment activés vont être transcrits plus rapidement et plus efficacement que dans les lymphocytes naïfs. Cette capacité qu'ont les cellules T à se souvenir des réponses transcriptionnelles passées est appelée « mémoire transcriptionnelle adaptative » et est médiée notamment par des changements épigénétiques. Ces modifications de la chromatine ont

lieu lors de la rencontre initiale avec l'antigène et marquent les gènes impliqués dans la réponse immunitaire des lymphocytes T [185].

1.5.2.2 La plasticité cellulaire et dé-différenciation

Enfin, cette forme de mémoire transcriptionnelle est très intéressante lors des processus de plasticité cellulaire et de dé-différenciation (induite ou forcée).

Dictyostelium est un protiste qui peut physiologiquement inverser complètement sa différenciation en 24 heures environ. C'est donc un organisme modèle pour étudier la dé-différenciation physiologique. La différenciation se produit lorsque l'environnement devient pauvre en nutriments : les amibes s'agrègent alors pour former une structure multicellulaire où elles vont prendre deux identités différentes. Cette différenciation peut être reversée avant que les cellules n'atteignent un point de décision critique appelé l'effacement (erasure), au-delà duquel les cellules perdent cette capacité [186].

Chez les eucaryotes, en cas de blessures, des cellules partiellement engagées peuvent aussi ré-acquérir des phénotypes « souches » leur permettant ainsi de repeupler les compartiments de cellules souches, par exemple au niveau des cryptes intestinales [187], ou encore de l'épithélium pulmonaire [188].

L'exemple le plus évident de dé-différenciation induite est celui des iPSC (Induced Pluripotent Stem Cells), correspondant à des cellules différenciées qui sont reprogrammées à l'aide d'un cocktail de facteurs de transcription spécifiques leur permettant de revenir à un état pluripotent. Cette dé-différenciation récapitulerait les intermédiaires du développement mais en sens inverse. Cependant, si les cellules retiennent une mémoire de leur engagement précédent, cela pourrait engendrer des biais de choix de lignage. En effet, elles pourraient être « marquées » et se différencier plus facilement dans le type cellulaire dont elles proviennent initialement [189, 190].

La mémoire non-génétique est donc très probablement une variable « cachée » qui joue un rôle important dans les processus de décision cellulaire. Il a d'ailleurs été proposé que la mémoire transcriptionnelle est l'un des mécanismes principaux qui favorise la précision et la coordination des programmes d'expression des gènes lors des processus biologiques [169, 176]. Cependant, les quelques études qui interrogent la mémoire transcriptionnelle présentées ci-dessus ne questionnent pas directement la relation entre processus de décision cellulaire, tel que la différen-

ciation, et mémoire transcriptionnelle. Ainsi dans ma thèse je me suis principalement intéressée à la transmission des niveaux d'ARNm au cours d'un processus de différenciation cellulaire. Cela a nécessité la mise en place d'outils expérimentaux et l'identification d'outils computationnels adaptés. En effet, étudier la mémoire transcriptionnelle à l'échelle du transcriptome consiste concrètement à récupérer à la fois les informations de généalogie et de transcriptomique en cellules uniques sur des très petits clones de cellules apparentées ce qui représente un vrai challenge technique. Premièrement parce que les méthodes d'isolement de cellules les plus utilisées en amont des analyses transcriptomiques nécessitent un grand nombre de cellules qui sont mélangées et donc pour lesquelles l'information de généalogie est perdue. Et également parce qu'il n'existe pas de méthode satisfaisante permettant de suivre la généalogie de cellules avec la résolution d'une ou deux divisions cellulaires combinée à des approches de sc-RNA-seq.

1.6 Les outils pour étudier la mémoire transcriptionnelle

1.6.1 Le suivi des cellules apparentées

Pour étudier la mémoire transcriptionnelle, il faut comparer les transcriptomes de cellules apparentées sur une ou plusieurs générations cellulaires, autrement dit des cellules sœurs ou cousines. Les méthodes de cell tracking permettent de déterminer les relations de lignage entre une cellule mère et ses descendantes voire entre toutes les cellules d'un organisme en développement.

1.6.1.1 Les méthodes manuelles

La manière la plus directe de conserver l'information de généalogie entre des cellules est de suivre leurs divisions par microscopie (Figure 1.22).

Initialement, le suivi direct était limité à des organismes suffisamment petits pour être transparents, comme *C. elegans*. En suivant les cellules de cette façon pendant le développement de *C. elegans*, Sulston *et al.* ont pu mettre en évidence le lignage complet de cet organisme [192].

Aujourd'hui avec le développement des marqueurs cellulaires radioactifs, enzymatiques et fluorescents, ces méthodes manuelles sont adaptables à un grand nombre d'organismes. Ces méthodes sont prospectives, c'est-à-dire qu'en général, une cellule fondatrice est marquée et ses descendantes sont suivies au cours du temps par imagerie directe.

Bien que ces méthodes semblent simples à mettre en place, car elles nécessitent peu de matériel, elles exigent néanmoins un certain savoir-faire. De plus, la résolution est relativement faible c'est-à-dire qu'on ne peut suivre qu'un nombre limité de cellules pendant un temps limité parce que les marqueurs cellulaires vont être dilués au cours des divisions. Enfin, si l'objectif final est d'analyser le transcriptome des cellules, les cellules devront être récupérées manuellement à l'aide de capillaires ou d'un robot pipeteur ou par micro-dissection mais le rendement, le nombre de cellules récupérées, reste faible.

1.6.1.2 Les méthodes de « tagging » génétique

Les méthodes de « tagging » génétique permettent une meilleure résolution et un rendement plus important pour avoir accès à la généalogie cellulaire. Le « tagging » génétique consiste à introduire une marque héréditaire dans l'ADN d'une cellule qui sera transmise à ses descendantes. Ces méthodes permettent de reconstruire *a posteriori* les relations de parentés des cellules *via*

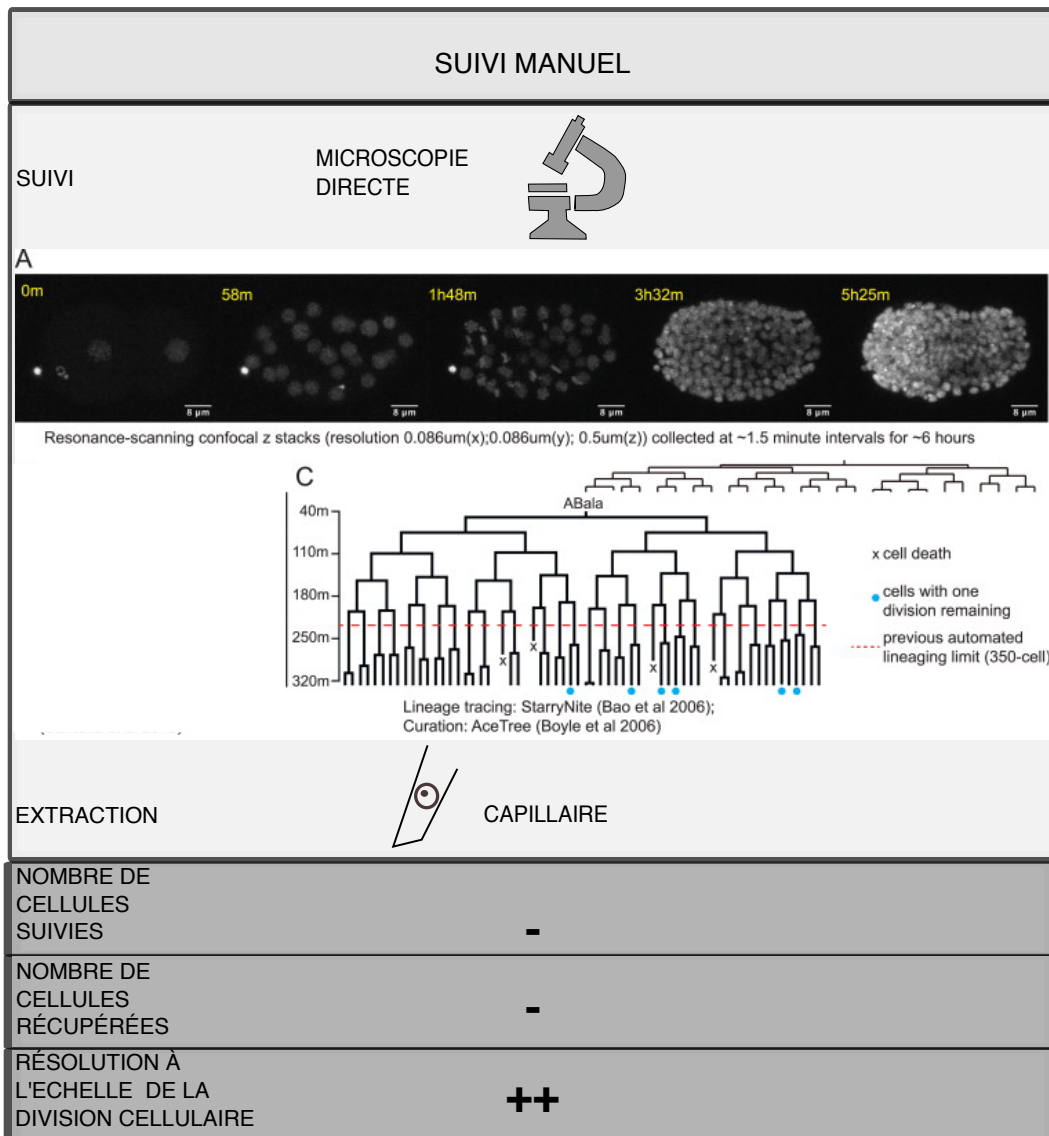


FIGURE 1.22 – Schéma du suivi « manuel » de cellules en division. Les divisions des cellules sont suivies en direct à l'aide d'un microscope. L'image centrale est issue de Richards et al. 2016 [191].

l'analyse des données de fluorescence ou par sc-RNA-seq (Figure 1.23).

Une des approches de « tagging » génétique consiste à insérer des blocs contenant des séquences d'ADN codant pour différentes protéines fluorescentes (Figure 1.23 panel de gauche). Ces séquences sont flanquées par des sites Lox et vont être aléatoirement intégrées dans le génome par recombinaison grâce à une CRE-recombinase. Ainsi, dans chaque cellule, va être généré un code-barre fluorescent unique qui sera transmis à ses descendantes. Ces code-barres peuvent être introduits à différents stades de développement et dans différents types cellulaires. C'est la méthode utilisée pour générer les « brainbow » [193]. Pour combiner cette méthode à de la transcriptomique en cellule unique, les cellules vont être triées et leurs code-barres fluorescents

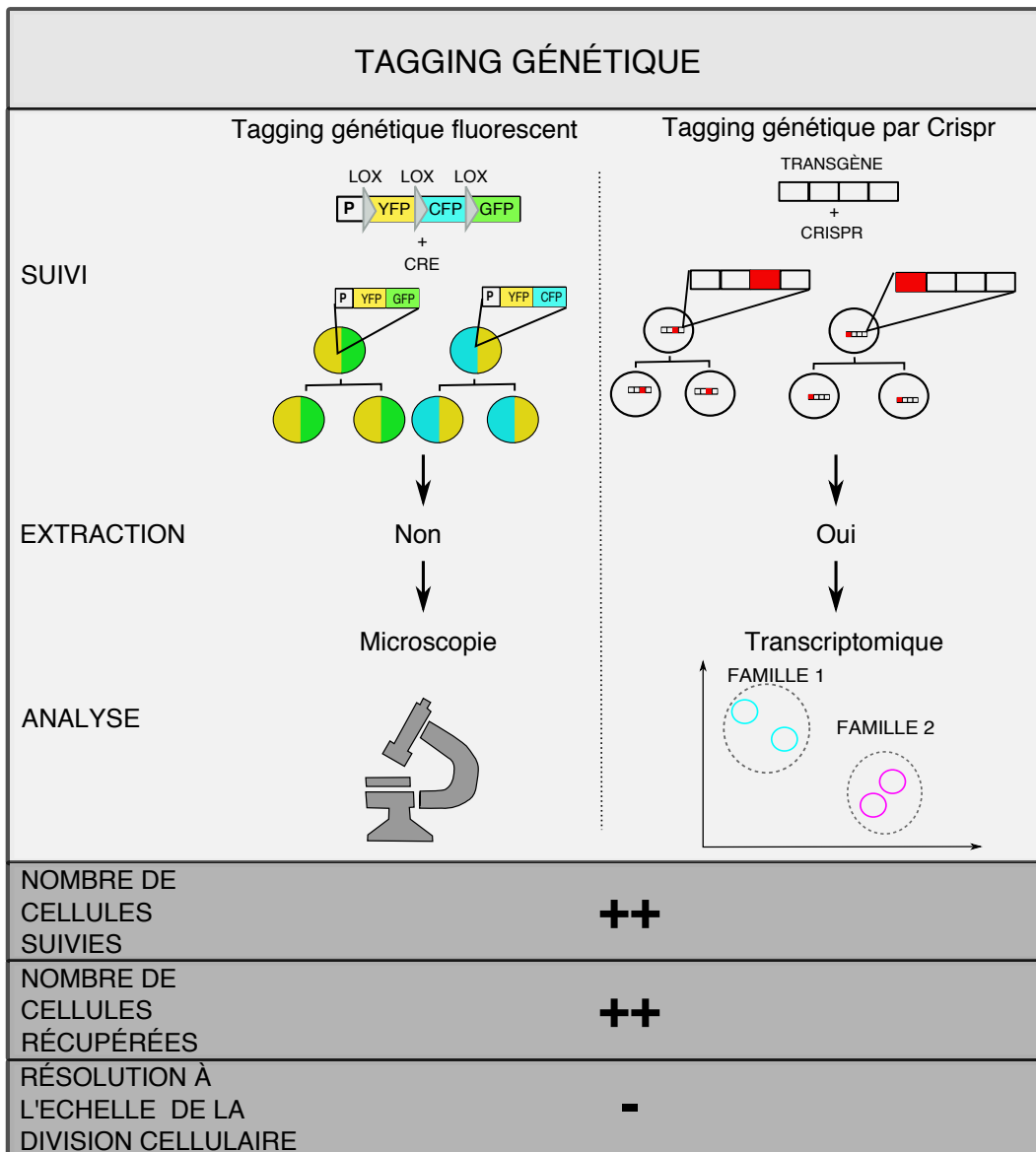


FIGURE 1.23 – Schéma du tagging génétique.

Le tagging génétique consiste à introduire des code-barres (correspondant à des code-barres fluorescents ou cicatrices) dans le génome d'une cellule fondatrice qui seront transmis à ses descendantes. Ces code-barres seront ensuite analysés pour reconstruire les relations entre cellules, soit par analyse de la fluorescence ou analyse des cicatrices génétiques en fonction de la méthode employée.

enregistrés lors du tri. Les relations généalogiques entre les cellules seront reconstruites à posteriori grâce à la similarité des code-barres fluorescents. La limite majeure de cette technique est le nombre de couleurs différentes qu'il est possible de multiplexer en parallèle pour discriminer suffisamment les différents clones de cellules.

Une autre famille de méthode qui combine le « tagging » génétique et le sc-RNA-seq plus directement permet un niveau de multiplexage plus important (Figure 1.23 panel de droite). Pour cela des transgènes sont introduits dans les cellules mères. A l'aide de différents outils, notamment

par CrispR-Cas9, des cicatrices génétiques aléatoires vont être générées dans les séquences de ces transgènes. Ainsi, dans chaque cellule, les cicatrices seront uniques et elles seront transmises aux descendantes de manière clonale. Ces transgènes vont être transcrits dans les cellules. Les banques de sc-RNA-seq vont ensuite être construites. Enfin, à partir des homologies de séquences des ARNm issus des transgènes, il sera possible de reconstruire les relations de parentés entre les cellules. En d'autres termes des cellules portant les mêmes mutations, aux mêmes endroits dans la séquence des ARNm des transgènes sont très probablement issues d'une même cellule fondatrice au départ.

Par exemple, la méthode LARRY de Weinreb *et al.* utilise des code-barres ADN uniques qui vont générer des ARN polyadénylés lors de leur transcription par les cellules. L'utilisation de cette méthode a permis de reconstituer les relations entre des petits clones de cellules et quelques groupes de cellules soeurs. L'analyse des ces cellules a permis de mettre en évidence que les niveaux de variation d'expression génique sont hérités et biaisent les choix de lignage des progéniteurs hématopoïétiques murins [194].

Ainsi, les méthodes de « tagging » génétique peuvent être facilement couplées à du sc-RNA-seq et permettre le suivi de nombreuses générations de cellules allant parfois même jusqu'à l'échelle d'un organisme entier comme cela a été fait chez le poisson zèbre [195].

Cependant, le « tagging » génétique n'est pas adaptable à tous les systèmes biologiques car il nécessite de lourdes modifications génétiques du système biologique et la résolution n'atteint pas l'échelle de la division cellulaire.

1.6.1.3 Les approches de microfluidique

L'utilisation de puces microfluidiques est une méthode alternative de cell tracking qui n'implique pas de modification génétique et qui peut être particulièrement polyvalente (Figure 1.24).

De manière générale, les systèmes microfluidiques consistent en des puits, des chambres ou des pièges qui sont des zones de culture miniaturisées où les cellules vont être isolées et cultivées. Les cellules sont introduites dans les puces *via* des petits tubes puis isolées dans les zones individuelles de culture de la puce. Elles vont ensuite proliférer et leurs divisions vont être suivies ou enregistrées par microscopie en time-lapse. Certaines puces permettent enfin d'extraire les cellules .

Le système MaSC développé par Schmitz *et al.* permet la capture de cellules uniques dans

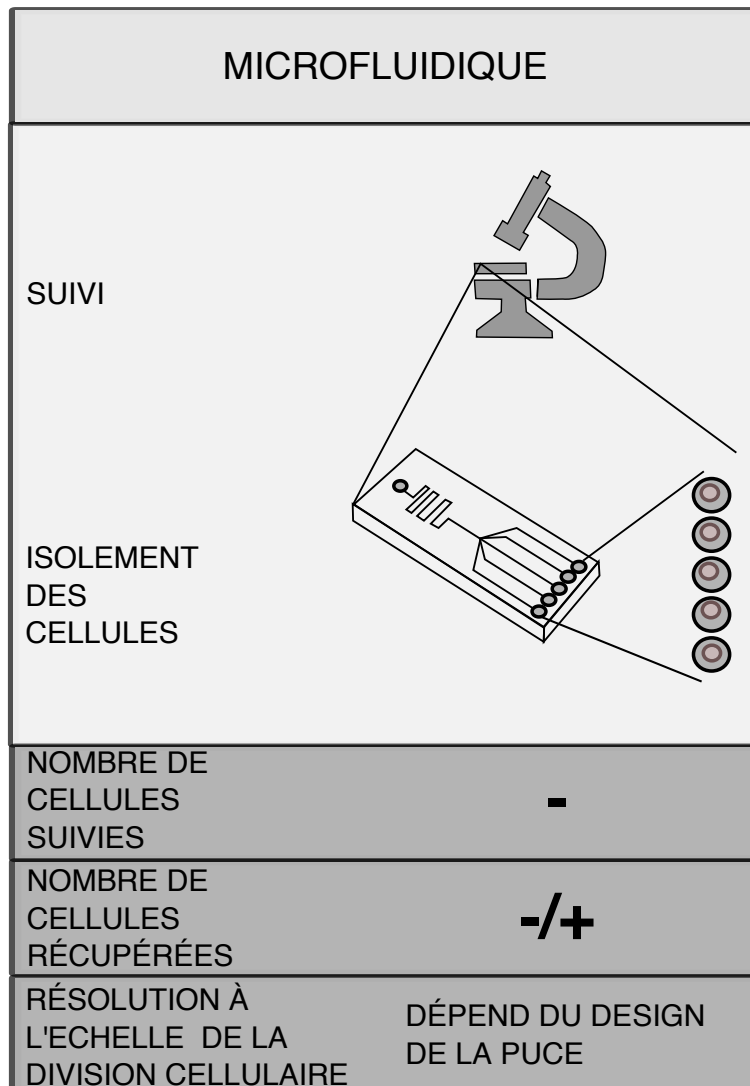


FIGURE 1.24 – Schéma du suivi de cellules par puce microfluidique.

Les puces microfluidiques permettent la capture de cellules mères uniques et leur division jusqu'à l'obtention de petits clones de cellules : pour certaines puces, l'extraction des cellules uniques pour réaliser ensuite des analyses de sc-RNA-seq est possible. Les divisions des cellules sont généralement suivies par microscopie en time-lapse.

une chambre et leur culture pendant environ 150h. La cellule et ses descendantes sont suivies par microscopie en time-lapse ce qui permet d'observer leur taux de croissance et leur temps de division qui sont variables d'une cellule à l'autre, ainsi que des événements cellulaires rares comme la polynucléation ou des réversions de mitose. Cependant, les cellules ne peuvent pas être extraites de cette puce [196].

Dans le système développé par Mulas *et al.*, des clones de cellules ES de souris sont cultivés à partir d'une cellule unique capturée dans une capsule d'hydrogel puis les clones sont induits à se différencier. Les clones peuvent être ensuite extraits en faisant sortir la capsule d'hydrogel de la puce permettant de réaliser des analyses fonctionnelles sur les cellules. Cependant avec cette

puce, les auteurs n'ont pas réalisé d'analyses transcriptomiques [197].

Enfin, Kimmerling *et al.* ont développé une puce dans laquelle des cellules uniques sont piégées et à chacune de leur division, l'une des cellules filles va être déplacée dans un nouveau piège. Le suivi des divisions en time-lapse permet de comprendre les relations de parenté entre les différentes cellules. Les cellules sont ensuite extraites pour réaliser des expériences de sc-RNA-seq, mais il n'est pas possible de sélectionner une cellule particulière. Les résultats des analyses transcriptomiques ont mis en évidence une mémoire transcriptomique entre cellules soeurs et cellules cousines [198].

Grâce à la microfluidique, les cellules sont isolées dans des espaces précis et observées en direct, il n'est donc pas nécessaire ni de les modifier génétiquement ni de les marquer avec des marqueurs fluorescents. Il est aussi possible de modifier leur environnement tout en observant en continu la réponse de chaque cellule sur de longues périodes de temps si besoin. Enfin, dans certains cas, les cellules peuvent être extraites de la puce pour analyser leurs transcriptomes. La quantité de cellules qu'il est possible de récupérer à partir d'une puce est variable et dépend du design de chaque puce mais peut être élevée.

Cependant la microfluidique est une discipline complexe qui demande une grande expertise et l'accès à ce type de technologie n'est pas nécessairement évident. Il s'agit généralement de collaborations avec des laboratoires experts qui ne commercialisent pas leurs puces. De plus, dans les cas où l'extraction des cellules est possible, les volumes d'extraction peuvent ne pas être compatibles avec du sc-RNA-seq en aval. En effet, les cellules sont généralement extraites *via* des petits tubes et cette extraction est réalisée à l'aveugle (il n'est pas toujours possible de suivre par microscopie la cellule lorsqu'elle est dans le tube d'extraction). De ce fait, afin d'être sûr que la cellule sorte de la puce, il faut extraire l'équivalent de trois fois le volume du tube, volume qui est déterminé par le diamètre et la longueur du tube ; des limites physiques impactent donc le volume minimal d'extraction. Enfin, avec les puces actuelles, il n'est pas toujours possible d'extraire avec précision une cellule donnée.

1.6.1.4 Les approches par FACS

En cytométrie en flux ou FACS (Fluorescence-activated cell sorting), les cellules passent devant un faisceau laser et dévient la lumière de celui-ci. En fonction de l'empreinte optique de la cellule, on peut en déduire des informations morphologiques sur celle-ci. En utilisant des

Cell-traceurs, qui sont des molécules fluorescentes, il est possible de faire du suivi de généalogie à l'aide d'un FACS. Les Cell-traceurs diffusent passivement à travers la membrane plasmique des cellules et sont liés de manière covalente aux lysines des protéines cytoplasmiques par des estérases cellulaires. A chaque mitose, les protéines marquées vont être partitionnées environ en deux et le signal fluorescent sera ainsi aussi divisé par deux à chaque division (Figure 1.25).

Ainsi, les Cell-traceurs sont majoritairement utilisés pour faire des mesures de prolifération de populations de cellules sur différents types cellulaires.

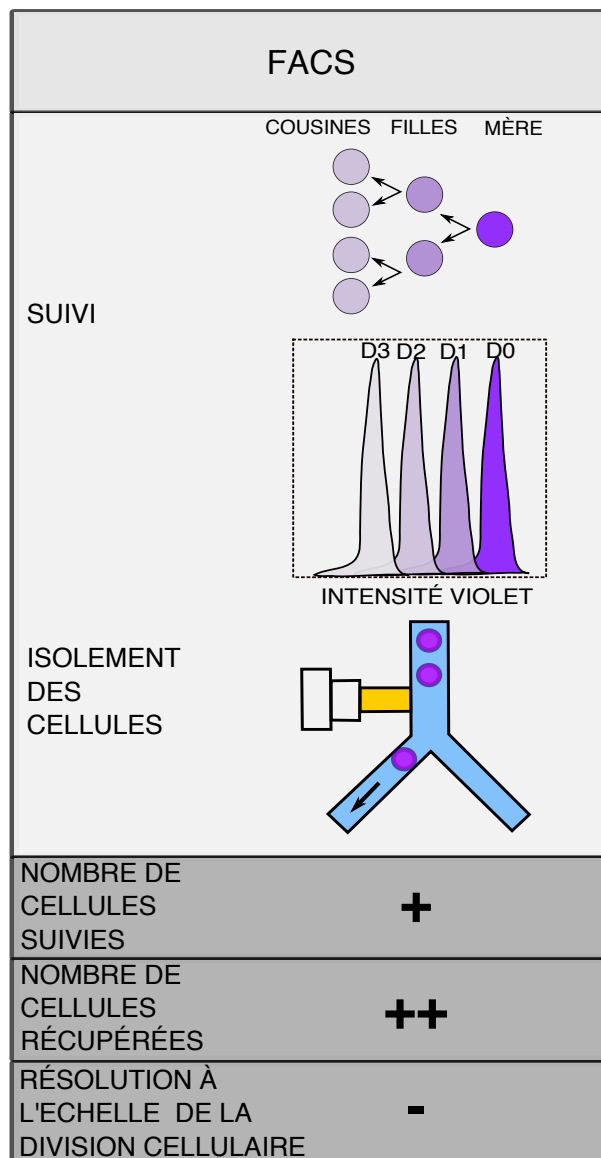


FIGURE 1.25 – Schéma du suivi des divisions de cellules marquées par des Cell traceurs, par FACS.

Les cellules sont marquées avec des Cell traceurs. L'intensité des Cell traceurs va diminuer à mesure que les cellules se divisent et se partitionnent les protéines marquées ce qui permettra de suivre le nombre de divisions cellulaires effectuées par une population de cellules. Les cellules sont ensuite isolées par FACS.

Ce type de marquage est simple à réaliser et il est possible de suivre la prolifération de plusieurs populations cellulaires en parallèle en réalisant un système de code-barres fluorescents avec plusieurs Cell-traceurs de couleurs et d'intensités différentes en même temps. Enfin couplé à du tri, il est relativement aisé de récupérer les cellules pour étudier leur transcriptome par sc-RNA-seq.

Néanmoins, tel qu'employé aujourd'hui les Cell-traceurs servent à compter le nombre de divisions effectué par une grande population de cellules, il n'est donc pas possible d'obtenir une résolution à l'échelle d'une division cellulaire, et donc de connaître les relations entre deux cellules soeurs entre elles ou entre des cellules cousines entre elles de manière précise avec cette technique. De plus, le tri de cellules au FACS nécessite généralement d'avoir une population de cellules non-adhérentes ou dissociées et une très grande quantité de cellules initiales. Il faut en moyenne 4 fois plus de cellules que le nombre que l'on souhaite récupérer.

Dans le cadre de ma thèse, nous avons développé des méthodes utilisant des Cell-traceurs pour retrouver les relations de généalogie entre des cellules : des cellules soeurs (après une division cellulaire) et des cellules cousines (après deux divisions cellulaires). Ces méthodes seront développées dans la partie résultat de ce manuscrit (chapitre 2).

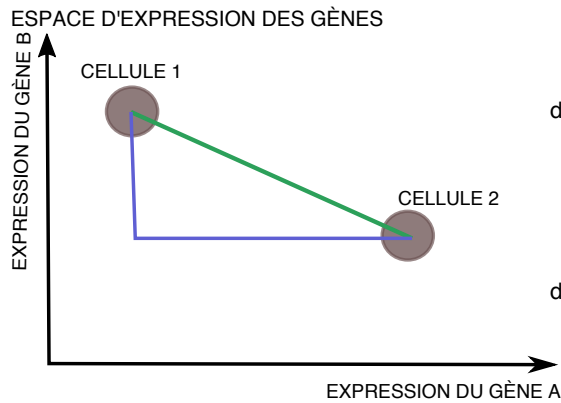
1.6.2 Les métriques qui permettent de mesurer la proximité des transcriptomes de cellules

Lors de l'analyse de données à la fois généalogiques et transcriptomiques, il faut 1) retrouver les relations généalogiques entre les cellules (avec les méthodes décrites ci-dessus) ; 2) préparer et normaliser les données de sc-RNA-seq (comme vu dans la partie 3 de l'introduction) ; 3) choisir les métriques adaptées qui vont permettre de déterminer la proximité des transcriptomes des cellules.

1.6.2.1 Les distances géométriques

Les distances géométriques permettent intuitivement d'évaluer la distance entre deux points. Appliquées à la mesure des transcriptomes de cellules apparentées il s'agit d'utiliser les valeurs d'expression des gènes d'une cellule comme des coordonnées pour la placer dans un espace d'expression des gènes. On peut ensuite mesurer la distance entre chacune des cellules.

A



Distance Euclidienne

$$d(C1, C2) = \sqrt{(C1_{\text{geneA}} - C2_{\text{geneA}})^2 + (C1_{\text{geneB}} - C2_{\text{geneB}})^2}$$

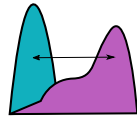
Distance de Manhattan

$$d(C1, C2) = |C1_{\text{geneA}} - C2_{\text{geneA}}| + |C1_{\text{geneB}} - C2_{\text{geneB}}|$$

B

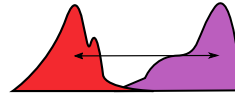
Distance de Wasserstein

Comparaison de la distribution de l'expression du gène A entre deux groupes de cellules (bleues et violettes)



Distribution du gène A

Comparaison de la distribution de l'expression du gène A entre deux groupes de cellules (rouges et violettes)



Distribution du gène A

Le groupe de cellules violettes a une expression du gène A plus proche du groupe de cellules bleues que du groupe de cellules rouges

FIGURE 1.26 – Schéma des différentes distances géométriques.

(A) Les distances de Manhattan et Euclidiennes peuvent être utilisées pour évaluer la proximité des transcriptomes entre des cellules. (B) La distance de Wasserstein peut être utilisée pour comparer la distance entre des distributions d'expression de gènes dans deux groupes de cellules.

La distance la plus utilisée est la distance euclidienne mais il a été montré que son application est limitée pour le traitement de données à grande dimension telles que les données sc-RNA-seq parce que l'élévation au carré peut conduire à une distorsion des données, et particulièrement parce que les données de sc-RNA-seq contiennent beaucoup de 0. La distance de Manhattan est plus robuste et moins sensible à la sparsité des données, elle est donc plus adaptée pour évaluer la proximité des transcriptomes de cellules apparentées à partir de données de sc-RNA-seq [199] (Figure 1.26 panel A).

Une autre manière de comparer la proximité entre des cellules apparentées à l'aide de distances est la distance de Wasserstein [200]. Cette métrique va permettre de comparer la distance entre cellules mais peut aussi être utilisée pour comparer la distribution d'un gène entre des groupes de cellules. Plus la distance est faible plus l'expression du gène est semblable entre les groupes et inversement, plus la distance est grande plus l'expression du gène est différente entre

les deux groupes (Figure 1.26 panel B).

1.6.2.2 Les modèles à effet mixte

Une autre approche pour déterminer si l'expression d'un gène est corrélée entre deux cellules soeurs est d'étudier sa variance. Pour cela, il s'agit de définir des groupes de cellules appartenant à la même « famille ». Chaque couple de cellules soeurs formera donc un groupe famille et la variance pour chaque gène sera mesurée au sein des groupes (variance intra-groupe). Si la variance intra-groupe est élevée pour le gène donné, il n'est pas corrélé, si au contraire la variance est faible on peut supposer que l'expression est proche au sein des familles et donc entre les cellules soeurs. On veut donc tester si l'effet du groupe, et donc l'effet famille, a un impact sur la variance de l'expression des gènes. Les modèles à effets mixtes sont un type de modèles statistiques largement utilisés dans diverses disciplines scientifiques telles qu'en physique, en biologie ou encore en sciences sociales pour étudier la variance.

Les modèles à effets mixtes permettent de modéliser le niveau d'expression d'un gène en fonction de différents paramètres. Dans le cadre de mon travail de thèse, on définira deux paramètres : la condition biologique, les cellules sont soit en auto-renouvellement soit en différenciation et l'appartenance à la même famille de cellules. Cet effet famille sera défini comme un effet aléatoire. En effet, dans le cadre des modèles à effets mixtes, définir un effet aléatoire revient en fait à décomposer la variance résiduelle et extraire la variabilité qui est liée à cet effet. Dans notre cas, en définissant l'effet famille comme un effet aléatoire, cela va permettre d'extraire la variance de l'expression du gène qui est lié à la parenté des cellules. Et plus précisément, la variance de cet effet famille correspond à la corrélation intra-classe, c'est-à-dire à la corrélation de l'expression du gène entre les cellules de la même famille.

On peut donc extraire les gènes dont l'expression est corrélée entre des cellules apparentées, c'est-à-dire les gènes soumis à une mémoire non génétique grâce à l'utilisation de ce type de modèle.

1.7 Aperçu général de la thèse

Comme précisé dans la partie introductive de ma thèse, la mémoire transcriptionnelle est une des variables impliquées dans la régulation de l'expression génique. Très peu d'études sont néanmoins disponibles sur la mémoire transcriptionnelle lors d'un processus de différenciation. Ainsi, comprendre si et comment la mémoire transcriptionnelle est maintenue au cours des générations cellulaires lors de la différenciation peut permettre d'affiner notre compréhension de ce processus de décision cellulaire.

Un des enjeux de ma thèse a donc été d'étudier l'évolution de la mémoire transcriptionnelle au cours du processus de différenciation érythrocytaire aviaire et donc comprendre de quelle manière se réconcilie l'augmentation de la variabilité observée au début de la différenciation et les contraintes de la mémoire transcriptionnelle.

Au début de notre travail, nous avons formulé 3 hypothèses sur le devenir de la mémoire transcriptionnelle lors de l'induction de la différenciation (Figure 1.27). Pour illustrer ces hypothèses, des cellules à l'état indifférencié sont placées dans l'espace d'expression génétique (cellules bleues). Puis, suite à l'induction de la différenciation, soit :

- 1) la mémoire transcriptionnelle a un effet plus fort que la variabilité de l'expression génique, ce qui fait que les cellules sœurs suivent le même chemin dans l'espace d'expression génique les menant vers l'état différencié (cellules rouges - sphère de gauche),
- 2) la variabilité de l'expression génétique a un effet plus fort que la mémoire transcriptionnelle, ce qui fait que chaque cellule sœur suit un chemin différent de celui des autres (sphère de droite),
- 3) la mémoire s'efface progressivement, ce qui se traduit dans notre représentation par des cellules sœurs qui commencent à suivre le même chemin puis se séparent progressivement les unes des autres (sphère du milieu).

Pour identifier l'hypothèse la plus juste, nous avons développé des méthodes expérimentales nous permettant de récupérer des cellules apparentées tout en conservant l'information précise de leur généalogie sur une et plusieurs générations cellulaires et de comparer leurs transcritomes. Nous avons également développé un pipeline bio-informatique permettant l'analyse de ces données sc-RNA-seq dans l'équipe. De plus, nous avons adapté et optimisé le protocole de sc-RNA-seq MARS-seq, ainsi qu'un protocole de séquençage compatible avec la structure de nos banques. Ces optimisations nous ont permis de générer des données de sc-RNA-seq exploitables

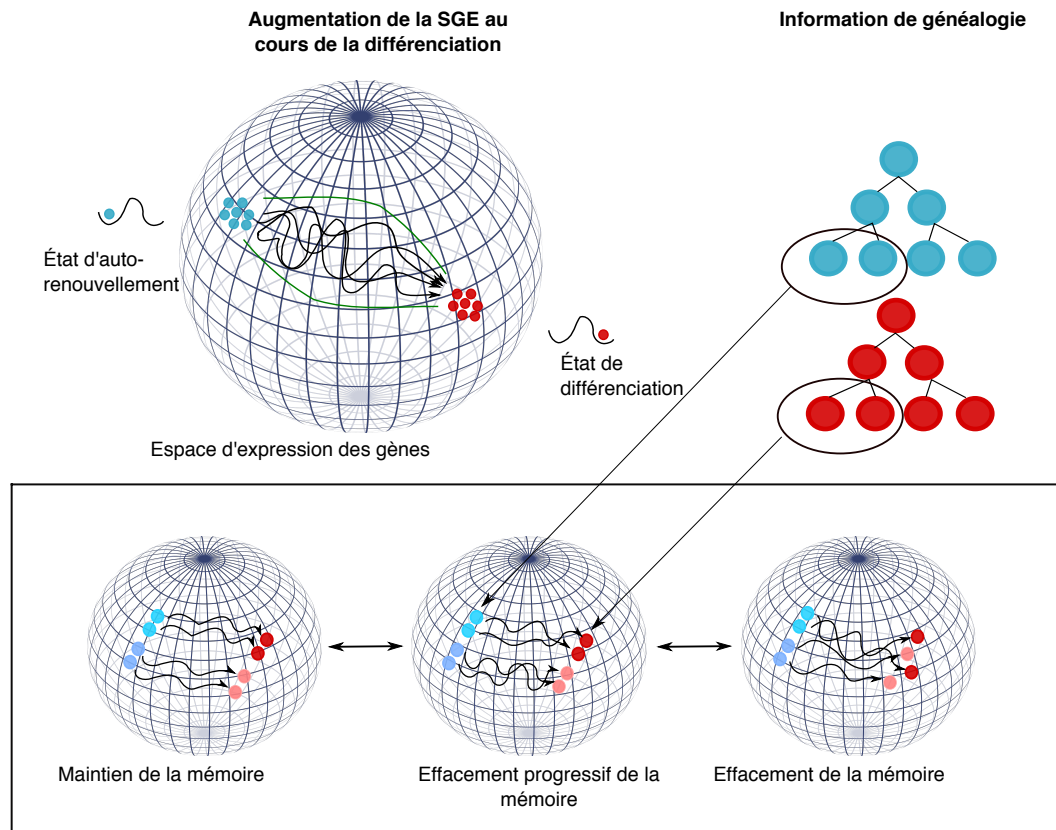


FIGURE 1.27 – Schéma de la problématique de la thèse.

La sphère représente l'espace d'expression des gènes où les cellules bleues en état d'autorenouvellement vont recevoir le signal de différenciation. La SEG augmente alors transitoirement dans ces cellules, représentée par le fait que chaque cellule va emprunter une route différente jusqu'à atteindre l'état différencié en rouge. Les 3 petites sphères représentent les différentes hypothèse : Sphère de gauche - la mémoire transcriptionnelle est fortement maintenue pendant la différenciation ; Sphère centrale - la mémoire s'efface progressivement au cours de la différenciation ; Sphère de droite - la mémoire transcriptionnelle est effacée au début du processus de différenciation.

sur une petite généalogie de cellules.

Dans un premier temps, nous nous sommes intéressés à la mémoire transcriptionnelle dans des progéniteurs érythrocytaires aviaires maintenus en l'auto-renouvellement et pendant leur différenciation au cours d'une et de deux générations cellulaires. Ce projet a nécessité le développement de deux types d'approches innovantes : les premières basées sur de la cytométrie en flux utilisée à la limite de sa résolution, et la seconde basée sur le développement d'une puce microfluidique spécifique dédiée à l'analyse transcriptionnelle sur plusieurs générations cellulaires.

Dans un second temps, en collaboration avec Souad Zreika (doctorante dans l'équipe), a été d'étudier la plasticité de notre modèle de différenciation mais sous un angle peu commun, c'est-à-dire à la mémoire transcriptionnelle comme mémoire d'un état transcriptomique. Comme précisé dans l'introduction bibliographique, une étude précédemment réalisée dans l'équipe avait mis en

évidence un point d'engagement particulier à 24h de différenciation en-deçà duquel les cellules pouvaient se remettre à proliférer si elles étaient re-placées dans du milieu d'auto-renouvellement. La ré-acquisition d'un phénotype moins engagé par ces cellules, qualifiées de cellules en réversion, avait ouvert la question de leur état moléculaire. Ainsi, nous avons caractérisé les transcriptomes de ces cellules en réversion par sc-RNA-seq et scRT-qPCR ; à l'aide d'outils computationnels et statistiques adaptés, nous avons pu comparer leur état transcriptomique à celui de cellules en différenciation ou indifférenciées.

Chapitre 2

Mémoire transcriptionnelle au cours des générations cellulaires

2.1 Introduction

La première partie de ma thèse visait à caractériser la mémoire transcriptionnelle au cours des générations cellulaires et comprendre comment les cellules concilient la contrainte de la mémoire transcriptionnelle et l'augmentation de la variabilité de l'expression génique nécessaire au processus de différenciation. Dans cet objectif, nous avons utilisé comme modèle cellulaire des progéniteurs érythrocytaires primaires aviaires (T2EC). Ces cellules peuvent être induites à se différencier en changeant simplement la composition de leur milieu de culture. Nous avons analysé les transcriptomes de cellules apparentées par sc-RNA-seq après une et deux divisions cellulaires dans un contexte d'auto-renouvellement et dans un contexte de différenciation.

2.2 Optimisations

Ce projet fortement pluridisciplinaire a nécessité un grand nombre d'optimisations, expérimentales et bio-informatiques que nous allons voir rapidement.

2.2.1 Isolement des cellules soeurs

Dans un premier temps il a fallu mettre en place une méthode permettant de cultiver des cellules mères et de récupérer les cellules filles issues de leur division en préservant leur relation généalogique. Initialement, nous nous étions orientés sur de l'isolement manuel à l'aide de capil-

lares en verre. Cependant, cette méthode particulièrement fastidieuse s'est révélée inappropriée pour trois raisons :

- 1) Le rendement était très faible (1 à 2 cellules par heure) notamment parce que post-mitose les deux cellules soeurs avaient tendance à coller l'une à l'autre ; il était donc nécessaire de les séparer délicatement et nous ne disposions pas de bras pipeteur qui aurait pu nous permettre d'aller plus vite.
- 2) Le volume d'isolement des cellules était très important et peu reproductible (environ $0.5\mu\text{L}$) ; ce volume contenant beaucoup de protéines, puisque les cellules étaient isolées à partir de milieu de culture, ce qui pouvait aussi impacter la reverse transcription en aval qui est l'une des étapes les plus sensibles.
- 3) Le temps d'isolement étant très long, il n'était pas exclu que cela puisse stresser les cellules et impacter leur expression génique.

Dans l'ensemble la qualité des données obtenues par cette méthode n'étant pas optimale, nous avons donc cherché à mettre en place une autre méthode permettant un isolement rapide des cellules dans un très faible volume. Avec l'aide de Sébastien Dussurgey, Ingénieur de la plateforme de cytométrie de la SFR Biosciences, nous avons ainsi développé une stratégie originale d'isolement des cellules soeurs par cytométrie en flux. Ce travail m'a permis de recevoir le prix de jeune cytométriste lors du congrès de l'Association Française de Cytométrie édition 2021.

2.2.2 Isolement des cellules cousines

Dans un second temps, nous avons développé une deuxième méthode permettant elle de récupérer des cellules cousines, soit après deux divisions cellulaires, en préservant à la fois les relations cousines mais aussi les relations soeurs deux à deux au sein des quatre cousines. Cette méthode est basée sur les résultats de la première méthode détaillée dans la partie suivante, montrant que les Cell-traceurs (dont le CFSE fait parti) sont répartis presque équitablement entre les cellules soeurs lors de la mitose. En effet, nous avons observé que la division d'une cellule mère marquée au CFSE générait deux cellules filles présentant une corrélation d'intensité de CFSE extrêmement élevée.

Ce projet a été réalisé en collaboration avec Sébastien Dussurgey, Ingénieur à la plateforme AniRA cytométrie de la SFR Biosciences, Catherine Koering Ingénieure de recherche et Elodie Vallin technicienne dans l'équipe, Laurent Modolo Ingénieur de recherche au pôle bio-computing du LMBC et Fanny Brunard stagiaire de master 1 que j'ai encadré pendant 7 semaines.

Pour ce projet, nous avons obtenu un petit financement (6000€) suite à un appel d'offre de la SFR Biosciences.

2.2.3 Bio-informatique

Troisièmement des optimisations bio-informatiques ont dû être faites avec le développement et l'adaptation d'un pipeline bio-informatique permettant de traiter et d'analyser les données de transcriptomique; pour cette partie, j'ai été supervisée par Laurent Modolo, Ingénieur de recherche du pôle bio-computing du LBMC. Et enfin des optimisations pour identifier les métriques de comparaison des transcriptomes de cellules soeurs; ce travail réalisé avec l'aide de Laurent Modolo et Franck Picard (DR CNRS dans notre équipe) qui m'a permis de me former aux outils bio-informatiques et mathématiques d'analyse de données.

Les résultats issus de l'ensemble de ces travaux feront l'objet d'un article actuellement en cours de rédaction.

2.3 Résumé des résultats

Nous avons développé une première méthode originale permettant de cultiver des cellules mères (T2EC) marquées au CFSE puis triées avec un trieur basse-pression pour maximiser la survie des cellules isolées en plaque 384 puits. Les cellules mères ont été triées soit dans du milieu d'auto-renouvellement, soit dans du milieu de différenciation. Après division (environ 24h de culture) les deux cellules soeurs résultantes de la division des cellules mères sont triées avec un FACS conventionnel (AriaII - BD) et isolées dans du tampon de lyse contenant les amorces de sc-RNA-seq dont la séquence des code-barres cellulaires uniques à chaque cellule est connue à l'avance. Les relations de sororité, c'est-à-dire les relation soeurs, sont retrouvées informatiquement et chaque cellule est identifiée grâce à la séquence du code-barre cellulaire qu'elle porte. Les relations de généalogie entre les cellules sont ensuite confirmées grâce à l'intensité du marquage CFSE qui est très corrélée entre les cellules soeurs et peu corrélée entre des cellules appariées aléatoirement. Enfin, les transcriptomes de ces cellules ont été obtenus par sc-RNA-seq et analysés grâce à des outils bio-informatiques et statistiques.

En parallèle, une étude très similaire à la notre a été réalisée par nos collaborateurs de l'EPHE de Paris. Ils ont analysé le transcriptome de cellules soeurs CD34+ humaines dérivées de sang de cordon ombilical. Comme expliqué dans l'introduction bibliographique, ces cellules consistent en

des progéniteurs hématopoïétiques et des cellules souches qui peuvent être différenciées *in vitro* grâce à un mélange de cytokines. 24 heures après la stimulation, l'augmentation de la variabilité transcriptionnelle produit un profil de transcription mixte appelé état « multilineage primed », puis à la fin du premier cycle cellulaire deux profils de transcription différents émergent dans la population cellulaire. Il s'agit du premier signe du processus de décision cellulaire dans ce modèle. Ces changements se produisent au cours du même cycle cellulaire mais il n'est pas clair si et comment ils sont liés à la division cellulaire. Les cellules CD34+ sont cultivées dans un petit device appelé SmartAliquoter permettant l'isolement de cellules uniques par dilution ; après la première division, les cellules soeurs sont séparées et isolées avec un capillaire en verre manipulé à l'aide d'un bras pipeteur. Le transcriptome de ces cellules a été analysé par scRT-qPCR.

Nous avons ainsi pu récupérer 30 couples de T2EC en auto-renouvellement, 32 couples de T2EC en différenciation et 43 couples de cellules humaines CD34+.

Nos résultats ont montré que :

- 1) nos méthodes d'isolement (par cytométrie ou manuelle) permettent d'obtenir et d'analyser, dans de bonnes conditions, les transcriptomes de cellules soeurs tout en conservant l'information de leur généalogie ;
- 2) les cellules soeurs, dans les deux modèles biologiques ont des transcriptomes statistiquement plus proches entre elles comparées à des cellules non apparentées, proximité mesurée à l'aide de distances géométriques ; pour le modèle T2EC, cette proximité entre cellules soeurs est observable aussi bien en état d'auto-renouvellement qu'au cours de la différenciation ;
- 3) la distance de Manhattan moyenne, bien que toujours plus petite entre cellules soeurs (en comparaison à des cellules non-apparentées), est plus grande dans les cellules soeurs en différenciation comparée aux cellules soeurs en auto-renouvellement ;
- 3) Cette proximité dans les niveaux de l'expression des gènes concerne un sous-ensemble de gènes, qualifiés de gènes mémoires et identifiés à l'aide d'un modèle mathématique à effets mixtes.
- 4) Cet effet mémoire ne peut pas être entièrement expliqué par des durées de demi-vie plus longues des ARNm des gènes « mémoire ».

Pour confirmer ces résultats nous avons développé une seconde méthode permettant de comparer les transcriptomes de cellules cousines, c'est-à-dire après une division supplémentaire au cours du processus de différenciation érythrocytaire.

Nos résultats montrent que notre stratégie de marquage utilisant des Cell-traceurs et notre outil d'analyse bio-informatique permettent de retrouver avec confiance des groupes de cellules cousines grâce aux différents code-barres fluorescents générés, et permet l'analyse de leur transcriptome par sc-RNA-seq en aval.

Suite au séquençage des banques et au traitement des données, nous avons pu analyser 8 groupes complets de cousines en auto-renouvellement et 5 groupes complets de cellules cousines en différenciation. Les analyses des résultats ont montré que :

- 1) les distances de Manhattan confirment que les cellules soeurs, quelque soit la condition biologique, sont toujours plus proches entre elles que des cellules appariées aléatoirement ;
- 2) comme observé précédemment, les cellules soeurs en différenciation sont moins proches entre elles que les cellules soeurs en auto-renouvellement, bien que non significatif ;
- 3) les distances de Manhattan moyenne des cellules cousines sont statistiquement plus petites que celles de cellules non-apparentées, et ce dans les deux conditions biologiques, indiquant une proximité des transcriptomes également entre cellules cousines ;
- 4) la distance de Manhattan moyenne entre les cellules cousines en différenciation est statistiquement plus grande que celle des cellules cousines en auto-renouvellement indiquant que les cellules cousines en différenciation sont statistiquement moins proches entre elles que les cellules cousines en auto-renouvellement.

2.4 Principales conclusions

Dans leur ensemble, nos résultats montrent que la mémoire transcriptionnelle est plus robuste que le changement d'état cellulaire puisque les cellules soeurs sont toujours plus proches entre elles que des cellules non apparentées. La robustesse de notre observation est appuyée par l'utilisation de deux modèles cellulaires différents et deux technologies d'analyse du transcriptome en cellules uniques différentes.

Néanmoins, dans les T2EC, les cellules soeurs en différenciation sont moins proches entre elles, au niveau transcriptomique, que les cellules soeurs en auto-renouvellement. Et après une division supplémentaire, cette différence devient statistiquement significative. Ces résultats suggèrent que la mémoire, bien que présente, serait moins forte au cours de la différenciation. Cette éloignement plus rapide des cellules soeurs au cours de la différenciation pourrait être dû à l'augmentation de la variabilité précédemment observée au cours de ce processus.

2.5 Publication - article 1

1 **DIFFERENTIATION IS ACCOMPANIED BY A**
2 **PROGRESSIVE LOSS IN TRANSCRIPTIONAL MEMORY**

3 Camille Fourneaux^{*,1}, Laëtitia Racine^{*,2}, Catherine Koering^{&,1}, Sébastien
4 Dussurgey^{&,3}, Elodie Vallin¹, Alice Moussy², Romuald Parmentier², Fanny
5 Brunard¹, Daniel Stockholm², Laurent Modolo¹, Franck Picard¹, Olivier
6 Gandrillon^{1,4}, Andras Paldi² and Sandrine Gonin-Giraud^{%,1}.

7 1 - Laboratory of Biology and Modelling of the Cell, Université de Lyon, Ecole
8 Normale Supérieure de Lyon, CNRS, UMR5239, Université Claude Bernard
9 Lyon 1, Lyon, France.

10 2 - Ecole Pratique des Hautes Etudes, PSL Research University, UMRS938,
11 CRSA, Paris, France.

12 3 - Univ Lyon, ENS de Lyon, Inserm, CNRS SFR Biosciences US8 UAR3444, UCBL,
13 50 Avenue Tony Garnier, F-69007 Lyon, France.

14 4 - Inria Center Grenoble Rhone-Alpes, Equipe Dracula, Villeurbanne, France.

15

16 * Those authors contributed equally. & Those authors contributed equally.

17 % Corresponding author sandrine.giraud@ens-lyon.fr

18 Abstract

19 Cell differentiation requires the integration of two opposite processes, a stabi-
20 lizing cellular memory, especially at the transcriptional scale, and a burst of
21 gene expression variability which follows the differentiation induction. There-
22 fore, the actual capacity of a cell to undergo phenotypic change during a dif-
23 ferentiation process relies upon a modification in this balance which favors
24 change-inducing gene expression variability. However, there are no experi-
25 mental data providing insight on how fast identical cells transcriptomes would
26 diverge on the scale of the very first two cell divisions during the differentia-
27 tion process.

28 In order to quantitatively address this question, we developed different
29 experimental methods to recover related cell transcriptomes, after one and
30 two divisions, while preserving the information about their lineage at the
31 scale of a single cell division. We analyzed the transcriptome of related
32 cells from two differentiation biological systems (human CD34+ cells and
33 T2EC chicken primary erythrocytic progenitors) using two different single-
34 cell transcriptomics technologies (sc-RT-qPCR and scRNA-seq).

35 We identified that the gene transcription profiles of differentiating sister-
36 cells are more similar to each-other than to those of non related cells of the
37 same type, sharing the same environment and undergoing similar biologi-
38 cal processes. More importantly, we observed greater discrepancies between
39 differentiating sister-cells than between self-renewing sister-cells. Further-
40 more, a continuous increase in this divergence from first generation to second
41 generation was observed when comparing differentiating cousin-cells to self
42 renewing cousin-cells.

43 Our results are in favor of a continuous and gradual erasure of transcrip-
44 tional memory during the differentiation process.

45 Introduction

46 During the division, the mother-cell endures a period of transient instability
47 – the mitosis – which is accompanied by dramatic cellular and epigenome
48 reorganizations [1]. The existence of maintenance mechanisms ensure the
49 conservation of cell’s identity throughout the division, referred as cellular
50 memory. Indeed, passive mechanisms, such as the roughly (random) equal
51 partitioning of all of the components of the mother-cell, contributes to main-
52 tain the similarity of the sister-cells and, in general, to the stable maintenance
53 of the cell clone’s phenotype. Furthermore, active mechanisms, such as the
54 conservation of gene transcription profiles after division by chromatin-related
55 epigenetic mechanisms, together with the long half-life of proteins ensure the
56 overall phenotypic similarity of sibling cells [2–4]. This cellular memory, pro-
57 moted by both passive and active mechanisms, is a major force contributing
58 to the stability of cell phenotypes in a multicellular organism.

59 A small number of studies have addressed the question of the link be-
60 tween division and cellular memory, using different approaches ranging from
61 microfluidics combined with scRNA-seq [5], to time-lapse microscopy of re-
62 porter genes expression [6, 7], to a dedicated procedure called MemorySeq
63 [8]. Those studies have been focused on the cellular memory of self-renewing
64 cells, such as mouse ES cells or melanoma cell line. In all cases, the authors
65 concluded to the existence of a transcriptional memory defined by the heri-
66 tability of gene expression levels in a gene-specific manner, extending up to
67 two or more generations. This transcriptional memory impacts subsets of
68 genes which expression is variable in a population of cells but correlated in
69 sisters or more generally related cells. Those genes are highly dependent of
70 the cell system used for the investigation. Beyond their actual function, the
71 fact that related cells present correlated expression for those genes is a read-
72 out for this transcriptional memory and demonstrates a constraint imposed
73 to the cells’ gene expression profile.

74 On the other hand, the transcriptional state of the original founder cell
75 is gradually reshaped divisions after divisions. All cellular processes are sub-
76 jected to stochastic molecular fluctuations which result in the decorrelation of
77 the sister-cells phenotypes and increase the transcriptional heterogeneity in
78 a clonal population of siblings. For example, relaxation experiments demon-
79 strated on various cell systems that after two weeks of culture under stable
80 conditions, the expression level of a specific gene in a selected homogeneous
81 cell clone becomes as heterogeneous as it was in the original population the

82 founder cell derived from [9]. Moreover, the capacity of a cell clone to re-
83 constitute the heterogeneity of the original population over time has been
84 observed in many instances in normal or pathological cell types [4, 8, 10].

85 The stochasticity of molecular interactions, involved in all cellular pro-
86 cesses but especially in gene transcription, has been proposed to play a key
87 role in the capacity of the cells to differentiate [11, 12]. A large range of
88 experimental studies have indeed demonstrated, that the first step in cell
89 differentiation is the rapid and transient increase of the variability in gene
90 expression in response to the stimuli inducing the differentiation, both *in*
91 *vitro* [13–21] and *in vivo* [22, 23].

92 Interestingly, fate decision is frequently thought to be related to cell di-
93 vision (see e.g. [24]). In fact, division is an ideal time to induce changes since
94 the whole cell organization will be altered by the division process [1]. There-
95 fore, the actual capacity of the cells to undergo a phenotypic change such
96 as differentiation [25] relies upon a precarious balance between the two op-
97 posing forces of the stabilizing cellular memory and change-inducing gene
98 expression fluctuations. However, very little is known about the extent to
99 which cell division alters this balance during a differentiation process.

100 Hence, considering the constraints applied by transcriptional memory and
101 the rise in gene expression variability, inherent to the initiation of the dif-
102 ferentiation process, we formulated 3 hypotheses on the possible behavior
103 of transcriptional memory upon differentiation induction (Figure 1). To il-
104 lustrate those hypotheses, cells in a self-renewing state are placed in a gene
105 expression space (grey sphere). Assuming the existence of transcriptional
106 memory in our self-renewing cells after mitosis, like in other cell models,
107 sister-cells start in roughly at the same position in that space (blue family
108 tree). Then, upon differentiation induction (red family tree), either:

- 109 • The transcriptional memory overrules the expression variability result-
110 ing in related cells following roughly the same path in the gene expres-
111 sion space toward the differentiated state (hypothesis 1 - Maintenance
112 of memory), or
- 113 • The memory is gradually erased, translated in our projection to differ-
114 entiating sister-cells starting to follow roughly the same path and pro-
115 gressively bifurcating from each other, and even more after one more
116 cell division (hypothesis 2 - Progressive erasure of memory), or
- 117 • The variability of gene expression pushes the balance and takes over the

118
119
120

transcriptional memory, leading each differentiating sister-cell to follow a completely different path from the beginning of the differentiation process (hypothesis 3 - Erasure of memory).

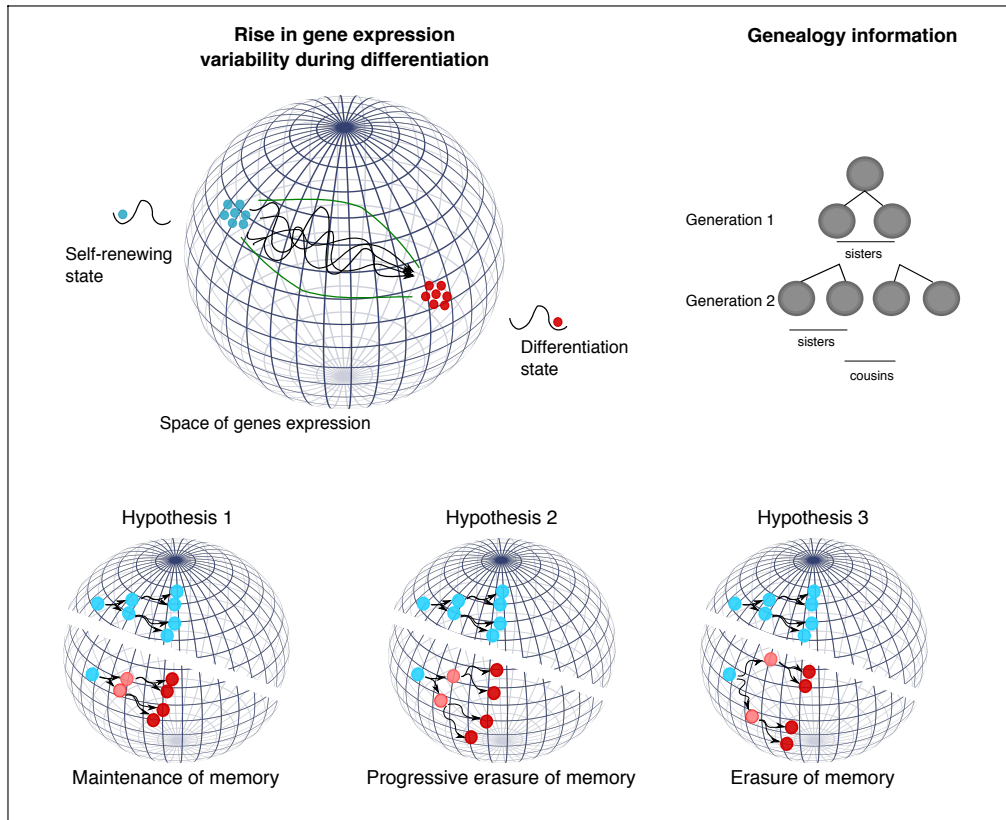


Figure 1: Hypotheses on transcriptional memory during a differentiation process.

Self-renewing cells (blue cells) are compared to differentiating cells (red cells) after one et two divisions.

121 In order to distinguish between those different scenarios, it is necessary
122 to be able to quantitatively evaluate, at the single-cell level, the similarity
123 of the gene expression profiles in sister-cells shortly after the division under
124 self-renewing versus under differentiation-promoting conditions.

125 Therefore, we developed two strategies to isolate cells while preserving
126 their precise lineage information after one (cells from generation 1) and two

127 (cells from generation 2) divisions, a manual one and a FACS oriented one.
128 Then, in order to assess the genericity and robustness of our findings, we
129 compared two different cell differentiation models (human CD34+ cells and
130 T2EC chicken primary erythrocytic progenitors) and for the T2EC model
131 two cellular states: self-renewing and differentiating. We used two different
132 single-cell transcriptomics methods: a highly sensitive targeted quantification
133 method, sc-RT-qPCR and a whole-transcriptome approach, scRNA-seq.

134 Through those experimental designs, we obtained qualitatively very sim-
135 ilar results using the two cell types and the two single-cell measurement
136 technologies. First, after one cell division (generation 1) in both models, and
137 in both states for the T2EC model, we observed a transcriptional memory
138 demonstrated by the sister-cells displaying more transcriptomic similarity
139 between each other than two randomly selected cells. Second, using the
140 T2EC model, which allows to compare sister-cells induce to differentiate to
141 sister-cells in self-renewing state, we also observed that this transcriptome
142 similarity decreased during the differentiation process as compared to the
143 self-renewing cells. Interestingly, this effect was even more pronounced one
144 division later (generation 2), when interrogating cousin-cells. Altogether our
145 results point toward a continuous gradual loss of transcriptional memory
146 during the differentiation sequence.

147 Results

148 Cellular models of differentiation

149 In order to consolidate our results we used two different cell differentiation
150 models. As a first model, we used primary human cord blood derived CD34+
151 cells. These cells are believed to be a mixture of so-called multipotent pro-
152 genitors and stem cells that retains the capacity to differentiate into various
153 cell types. Under *ex vivo* conditions, the CD34+ cells, unless stimulated, are
154 stopped in their cell cycle and survive only a few days. When stimulated
155 with a mixture of cytokines, they re-enter the cell cycle and will differentiate
156 into two different committed progenitors [17]. Briefly, by 24hrs after stimu-
157 lation, a burst in transcription produces a mixed transcription profile called
158 “multilineage primed” state [13] and by the end of the first cell cycle (be-
159 tween 40 and 60hrs), cells with two different transcription profiles emerge in
160 the population [17, 26]. However, this first fate-decision is a highly dynamic

161 and fluctuating process which is more complex than a simple binary switch
162 between 2 options [17]. In the present work, we investigated by sc-RT-qPCR
163 the transcriptional profile of couples of CD34+ sister-cells derived from the
164 first cell division after the cytokine stimulation.

165 As a second model, we used chicken primary erythrocytic progenitors
166 called T2EC [27]. Contrary to the human cord blood CD34+ cells, these
167 cells can be maintained in a self-renewing state *in vitro* under appropriate
168 culture conditions. They can be induced to differentiate at will into ma-
169 ture erythrocytes by a change of medium [28]. The T2EC cells undergo a
170 simple “switch”: they leave the self-renewing phase and enter a differenti-
171 ation trajectory without bifurcation toward different end point phenotypes.
172 This model allows a direct comparison of related cells in two different states:
173 self-renewing and during differentiation. Furthermore, a previous study on
174 this model had highlighted a critic point of cell commitment, 24hrs post-
175 differentiation induction characterized by the rise in gene expression vari-
176 ability, measured with entropy [29]. Thus, we focused on the first steps of
177 T2EC differentiation and investigated the transcriptional profile of couples
178 of generation 1 sister-cells compared in both cellular states and families of
179 generation 2 sisters and cousin-cells compared in both state by a scRNA-seq
180 approach [30].

181 Cells isolation

182 Isolation of first generation cells

183 We achieved the technical challenge to isolate related cells following their
184 first and second division (generation 1 sister-cells and generation 2 sisters
185 and cousin-cells). The usual molecular tagging or barcoding lineage trac-
186 ing approaches could not be used in our case since these approaches allow
187 retrieval and analysis of cells belonging to the same clones at later stages,
188 but not at the scale of one cell division [31]. For our investigation, a direct
189 observation of the dividing cells and individual isolation of the generation 1
190 sister-cells, and generation 2 sisters and cousin-cells were necessary. Further-
191 more the use of primary cells, with a short life span, precluded the possibility
192 to genetically engineer reporter systems.

193 We first developed two different methods to recover generation 1 sister-
194 cells, depending upon the cellular model at hand: a manual one and a
195 cytometry-based method. Those original strategies are presented below and

196 in Figure 2. The technical details are explained in the methods.

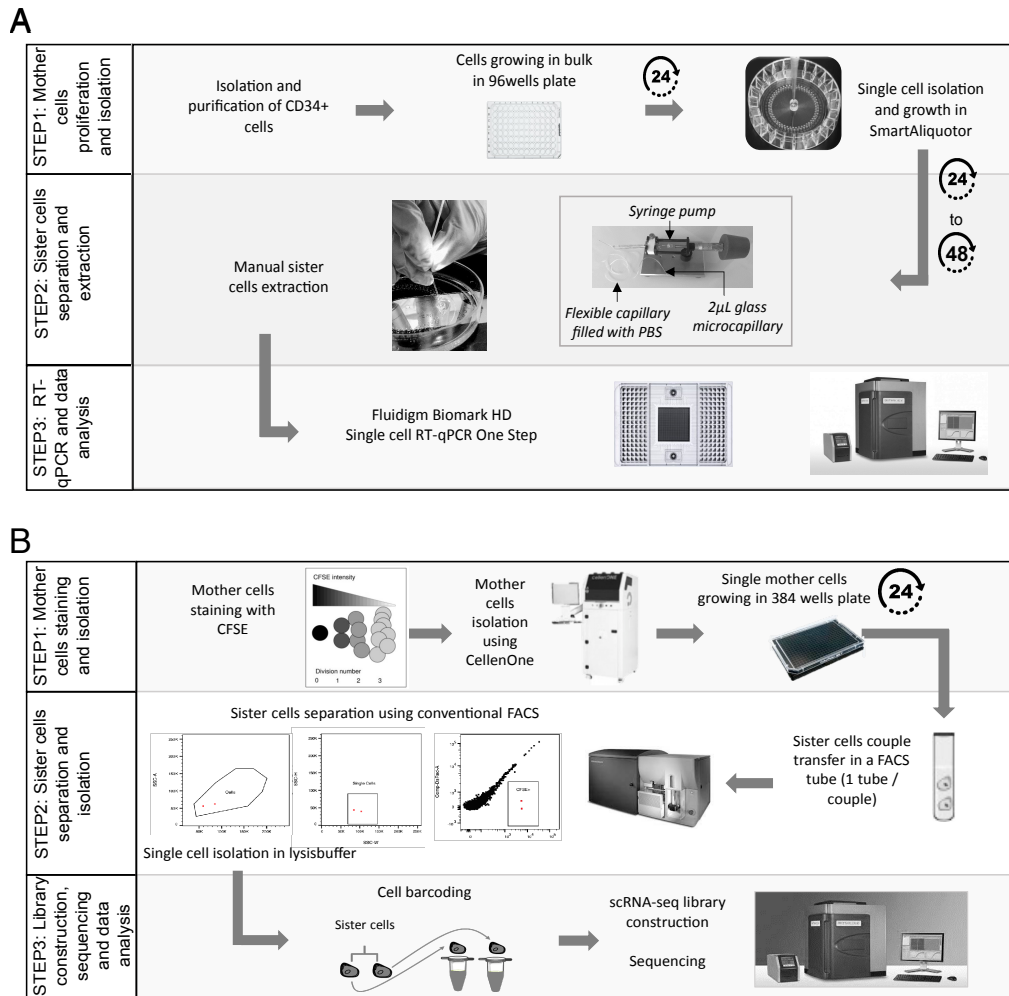


Figure 2: General workflows developed to generate, follow and separate generation 1 sister-cells from CD34+ (A - manual strategy) or T2EC (B - cytometry-based strategy) mother cells. See text and methods for details.

197 Human CD34+ cells were grown during 24hrs in a standard 96-well plate
 198 before being isolated into single cells, using a Smart Aliquotator in which
 199 individual cells still share the same medium. Isolated mother-cells were then
 200 cultured for 24 to 48hrs in the device to allow one cell division. The wells were
 201 regularly inspected to detect this first division. Then, the resulting sister-

202 cells were isolated manually under a microscope using a pressure controlled
203 microcapillary and recovered in lysis buffer for further processing. The cells
204 transcriptomes were analysed by single-cell quantitative RT-PCR using the
205 Fluidigm system as described here [17].

206 T2EC mother-cells were isolated after CFSE - carboxifluorescein diac-
207 etate succinimidyl ester - staining using CellenOne [®]low-pressure cell sorter
208 and plated in a 384-well plate. Cell doublets, resulting from the first division,
209 were identified using an inverted microscope. The two cells were then isolated
210 using an Aria FACS cytometer and recovered directly in tubes containing ly-
211 sis buffer and scRNA-seq primers, for which the cell barcodes sequences were
212 known in advance. scRNA-seq libraries were then constructed as previously
213 described here [32] and sequenced.

214 Successfully recovering the two sister-cells using FACS is *per se* a remark-
215 able achievement, as this method usually requires hundreds of cells to start
216 with, whereas the initial population here consisted of two cells. To achieve
217 this, we first used the CFSE fluorescence intensity to ensure that the objects
218 isolated were indeed cells (Figure S1 A-B for self-renewing medium and C-D
219 for differentiating medium). CFSE stably binds to the amine groups present
220 in cytoplasmic proteins, conferring stable fluorescence intensity to the cell.
221 As total protein content is supposed to be relatively equally distributed be-
222 tween sister-cells during cell division, so is the fluorescence intensity [33, 34].
223 We used this specification to validate that the two cells isolated were actu-
224 ally sister-cells. We evaluated the CFSE intensity correlation between pairs
225 of sister-cells, and compared it to intensity correlation values of randomly
226 paired cells from the same dataset (Figure S1 E-G for self-renewing cells and
227 F-H for differentiating cells). Outstandingly, CFSE correlation values be-
228 tween self-renewing sister-cells and differentiating sister-cells were extremely
229 high (0.91 and 0.95 Figure S1E and F, respectively), whereas for randomly
230 paired-cells, CFSE correlation values dropped between -0.07 for self-renewing
231 cells and 0.18 for differentiating cells (Figure S1G and H, respectively) indi-
232 cating no correlation. Those results validated that our general strategy did
233 allow to retrieve accurately generation 1 sister-cells. The same procedure
234 was applied to generation 1 T2EC cells in proliferating phase and in differ-
235 entiation by sorting the mother-cells either in self-renewing medium or in
236 differentiation-promoting medium.

237
238 We further analysed the T2EC scRNA-seq data quality and reproducibil-
239 ity by characterizing the observed biological process applying UMAP dimen-

240 sional reduction and projection method (see methods). As expected, the cells
241 separated based on their differentiation state (Figure S2A). This observation
242 was validated by a differential expression analysis between the two groups
243 (self-renewing and differentiating cells - Figure S2B). Genes involved in early
244 erythrocytes maturation, inhibition of differentiation such as *ID2* known to
245 be an erythropoiesis inhibitor in mice [35], *FTH1* and *TMSB4X* known to
246 be expressed in human erythroid progenitors [36] were up-regulated in self-
247 renewing cells while *HBBA*, *HBAD*, *HBA1*, genes involved in hemoglobin
248 complex and *TAL1*, erythroid differentiation factor, were up-regulated in
249 differentiating cells, as previously described [32].

250 **Isolation of second generation cells**

251 Using the T2EC model, we then developed another FACS sorting methodol-
252 ogy to retrieve generation 2 sisters and cousin-cells, that is to say the 4 cells
253 resulting from two divisions, both in self-renewing state or in differentiation
254 state. To record cells genealogies, we used different cells-tracers to achieve
255 fluorescent barcoding of cell's families and we stained the cells sequentially to
256 retrieve both cousins relationships and sisters relationships within different
257 families (Figure 3).

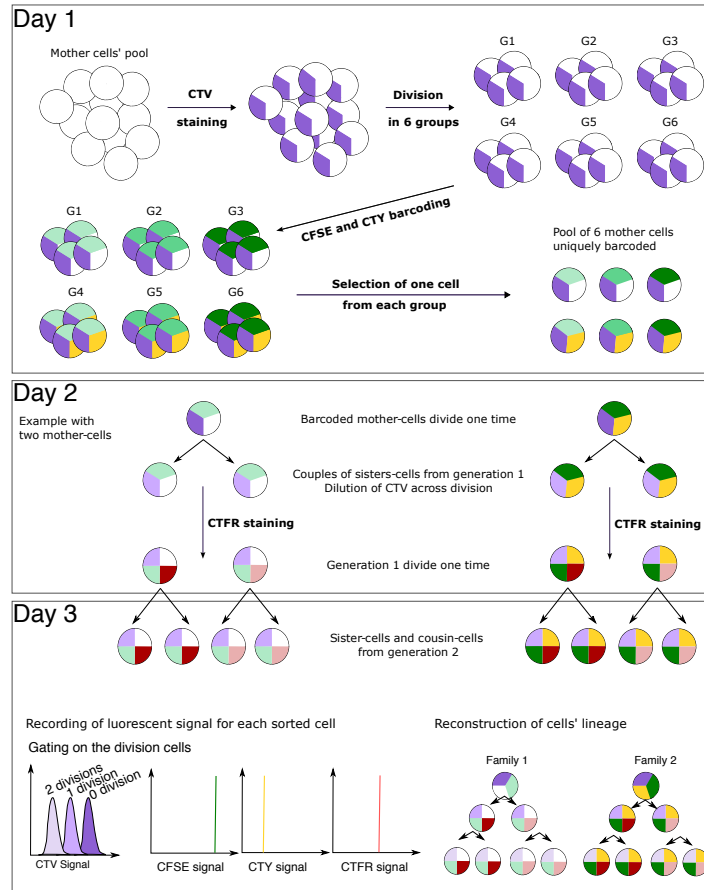


Figure 3: General labelling strategy for generation 2 T2EC cells identification.

On day 1, a population of mother cells was stained using CTV. The CTV positive population was split into 6 subgroups, each group was barcoded with a unique combination of CFSE and CTY concentration to achieve fluorescent barcoding (6 different barcodes). One mother cell from each group was then recovered and pooled together in a well to be cultured for around 24hrs (6 mother cells with a unique fluorescent barcode). At day 2, after the first division, a fourth dye, CTFR, was added to stain sister-cells with a different intensity in order to be able to discriminate the cells relationship after the next division. On day 3, cells which underwent 2 divisions, determined by the intensity of CTV, were sorted into single-cells and fluorescent intensities were recorded for CTY, CFSE and CTFR signals. Finally, a dedicated script was used to infer the relationships of cells based on the fluorescent intensities.

258 Briefly, a small number of mother-cells was stained such as every mother-
259 cell carried a unique fluorescent barcode. Each fluorescent barcode consist
260 in a combination of CTY and CFSE at different intensities, leading to 6 dif-
261 ferent barcodes. This barcode is passed along to the mother cell’s progeny
262 over two cell generations to allow a good discrimination of cell families. One
263 mother cell from each barcode was isolated by FACS in a well of a culture
264 plate. After the first cell division, another cell-tracer was added to discrimi-
265 nate sister-cells within the cousin groups. After the second cell division, the
266 cells (generation 2) were sorted in lysis buffer containing scRNA-seq primers
267 of known sequence and the relationships between the cells were recovered
268 using a clustering algorithm developed in our team. Details of the methodol-
269 ogy are presented in figure 3 and in methods. Further viability analysis was
270 performed and showed that the staining strategy did not compromise cells
271 physiology (Figure S3).

272
273 Using first generation methodologies, we successfully collected 86 CD34+
274 cells, 60 self-renewing T2EC cells and 64 differentiating T2EC cells encom-
275 passing respectively 43, 30 and 32 couples of generation 1 sister-cells. With
276 the second-generation original fluorescent barcoding approach, we collected
277 8 families of generation 2 self-renewing cells (32 cells) and 5 families of gen-
278 eration 2 differentiating cells (20 cells).

279 **Strategy to evaluate transcriptomic similarities between** 280 **related cells**

281 We used Manhattan distance as a metric to evaluate transcriptomic similari-
282 ties between cells. Manhattan distance is a robust geometric distance and is
283 less sensitive to data sparsity, which is inherent to single-cell transcriptomics
284 data [37].

285 We anticipated how the distance comparisons would result for each of the
286 hypotheses developed in the introduction.

287 The first hypothesis, maintenance of memory, assumes that there will
288 be no more transcriptional differences between self-renewing than between
289 differentiating sister-cells. This hypothesis would imply that at the first
290 cell generation, differentiating sister-cells would present a similar distance
291 between each other compared to self-renewing sister-cells. And at the second
292 generation, there would be no difference either between differentiating sister-

293 cells compared to self-renewing sister-cells nor between differentiating cousin-
294 cells compared to self-renewing cousin-cells.

295 The second hypothesis assumes that there will be a continuous and grad-
296 ual increase in the sister-to-sister differences as differentiation proceeds. Mean-
297 ing, at the first generation, differentiating sister-cells would present a greater
298 distance compared to self-renewing sister-cells. At the second generation, this
299 distance would increase and would be supported by (1) second-generation
300 differentiating sister-cells presenting a greater distance compared to second
301 generation self renewing sister-cells and (2) second generation differentiating
302 cousin-cells presenting a greater distance compared to self renewing cousin-
303 cells.

304 The last hypothesis assumes that there will be very strong transcrip-
305 tional differences between self-renewing and differentiating sister-cells at the
306 beginning of the differentiation process, with no evolution of those differ-
307 ences thereafter. That is, at the first generation, differentiating sister-cells
308 would present an substantial greater distance between each other compared
309 self-renewing sister-cells. At the second generation, differentiating sister-cells
310 cells would display a similar or smaller distance compared to self-renewing
311 sister-cells and differentiating cousin-cells would present a similar or slightly
312 greater distance compared to self renewing cousin-cells.

313 **Transcriptomic similarities between generation 1 sister-** 314 **cells after one division**

315 We started by assessing whether or not generation 1 sister-cells displayed
316 more similar global genes expression levels compared to non related cells.
317 Here non related cells correspond to cells which don't originate from a com-
318 mon mother-cell. Geometric distances were computed between the gene ex-
319 pression vectors of each cell. Gene expression vectors for the 43 couples of
320 CD34+ sister-cells were composed of 83 genes after quality control and data
321 filtering (see Methods). Those genes were either selected for their known
322 function in the early differentiation of hematopoietic cells (64% of them) or
323 randomly chosen (36%) to provide an assessment of the overall transcrip-
324 tional state of the genome. For the 62 couples of T2EC sister-cells gene
325 expression vectors, we retained 1177 genes after data filtering and normal-
326 ization of scRNA-seq data (see Methods). We performed the analysis by
327 computing distances between generation 1 sisters and randomly selected non

328 related cell pairs from the same pool of cells (Figure 4 A and B).

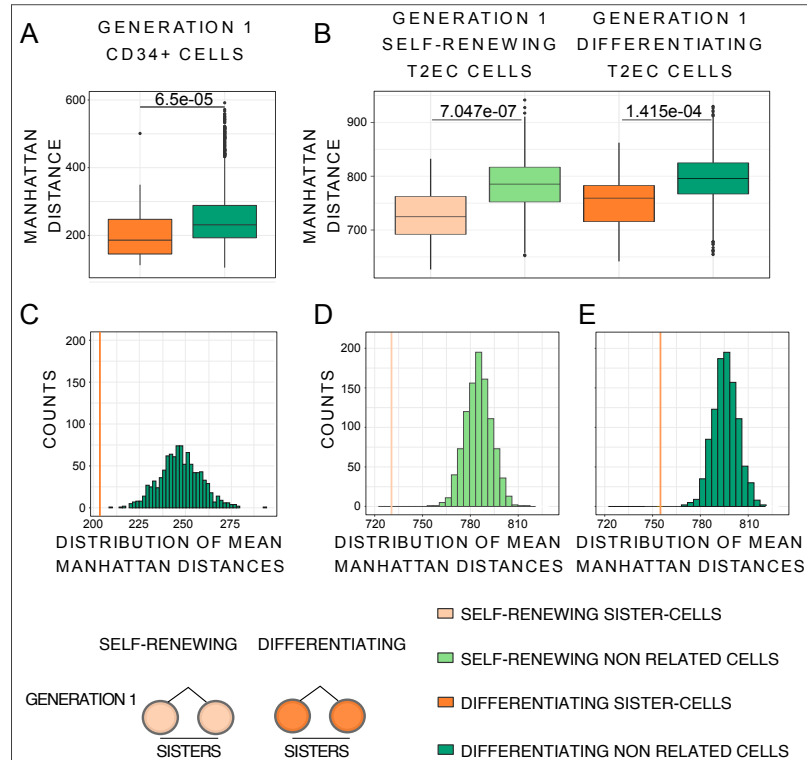


Figure 4: Manhattan distances comparison between generation 1 sister-cells and non related cells.

(A) Boxplot of Manhattan distances between the generation 1 CD34+. CD34+ sister-cells (43 couples) are in orange and CD34+ non related cells (3612 couples) in green. Manhattan distances were computed using all the 83 selected genes. Statistical comparison was performed using Wilcoxon test. (B) Boxplot of Manhattan distances between generation 1 T2EC sisters and non related cells. Manhattan distances were computed between all cells from the same biological conditions using all the 1177 selected genes. Self-renewing sister-cells (30 couples) are in light orange and self-renewing non related cells (1740 couples) in light green, differentiating sister-cells (32 couples) are in orange and differentiating non related cells (1984 couples) in green. Statistical comparison was performed using Student t-test. (C) Histograms of mean Manhattan distances of 1000 random straw of 43 CD34+ non related cell pairs distances (green), compared to mean distances of the 43 CD34+ generation 1 sister-cells pairs (orange line). (D) Histograms of mean Manhattan distances of 1000 random straw of 30 T2EC self-renewing non related cell pairs distances (light green histogram), compare to mean distances of the 30 T2EC self-renewing generation 1 sister-cells pairs (light orange line). (E) Histograms of mean Manhattan distances of 1000 random straw of 32 T2EC differentiating non related cell pairs distances (Green histogram), compare to mean distances of T2EC differentiating the 32 generation 1 sister-cells pairs (orange line).

329 Mean distances were then compared between the two groups (generation
330 1 sisters and non related cells) for both CD34+ and T2EC cells. For the
331 latter, both self-renewing and differentiating cells were analyzed separately.
332 For both models and in both biological conditions mean Manhattan distances
333 between generation 1 sister-cells were always significantly smaller than mean
334 distances between non related cells (Figure 4 A and B - Wilcoxon test for
335 CD34+ cells $pvalue = 6.5.10^{-5}$, Student t-test for self-renewing T2EC cells
336 $pvalue = 7.047.10^{-7}$ and for differentiating T2EC cells $pvalue = 1.415.10^{-4}$).

337 To ensure that the difference in mean distance observed between gener-
338 ation 1 sisters and non related cells was not an artefact due to difference in
339 sample size, we performed a randomization experiment by bootstrap. Briefly,
340 43 non related CD34+ cell pairs, 30 non related T2EC cell pairs and 32 non
341 related differentiating T2EC cell pairs were randomly drawn from the corre-
342 sponding groups 1000 times. The mean distance was calculated for each pair
343 and plotted on the histograms shown on figure 4 C, D and E. For both mod-
344 els, and for T2EC in both biological conditions, the mean distance between
345 generation 1 sister-cells was never part of the non related cells mean distance
346 distribution. Those results strongly suggest that the observed difference was
347 genuine and not due to sampling bias. This is a clear indication that the
348 gene transcription profiles of generation 1 sister-cells in both experimental
349 models are more similar to each-other than to those of non related cells of the
350 same type sharing the same environment and undergoing similar biological
351 processes.

352 Those results highlight that differentiating sister-cells from generation
353 1 display a form of transcriptional memory, which complements previous
354 studies demonstrating a transcriptional memory in self-renewing sister-cells.
355 Focusing on the T2EC model, for which we compared related cells in two cel-
356 lular states (self-renewing and differentiating), although the difference was
357 borderline non statistically significant ($pvalue = 0.06$), our results point to-
358 ward a decrease in transcriptome similarity during differentiation as shown
359 by a higher mean distance value for generation 1 differentiating T2EC sister-
360 cells compared to self-renewing T2EC sister-cells. We wondered whether or
361 not the sister-to-sister cell distance will continue to increase as the differen-
362 tiation proceeds in the T2EC cells, one generation later.

363 **Generation 2 cells transcriptomes continue to diverge**
364 **during differentiation**

365 We generated a second dataset consisting of generation 2 T2EC sisters and
366 cousin-cells (after two cell divisions) using the methodology described above.
367 As scRNA-seq requires the lysis of the cell under investigation, generation 1
368 data and generation 2 data consist of different cell families and thus cannot be
369 compared to each other. Both dataset were treated and analysed separately
370 (see methods).

371 The second generation dataset was composed of 4 cousin-cells per family
372 (8 families of cells in self-renewing and 5 families of cells in differentiation con-
373 dition), and within the 4 cousins, they consisted of two couples of sister-cells.
374 After data filtering and normalization, we retained 983 genes for subsequent
375 analysis.

376 Comparison of mean Manhattan distances from those data showed that
377 when comparing conditions, in line with previous results described after one
378 cell generation in figure 4, generation 2 differentiating sister-cells were less
379 close to each other than generation 2 self-renewing sister-cells, although not
380 significantly so (Figure 5).

381 Interestingly, generation 2 differentiating cousin-cells were statistically
382 further apart from the generation 2 self-renewing cousin-cells. Indeed, the
383 average Manhattan distance between generation 2 differentiating cousin-cells
384 was statistically greater than that of generation 2 self-renewing cousin-cells
385 further confirming a decrease in transcriptome similarity during the differen-
386 tiation process (Student t-test pvalue = 0.002218).

387 Finally, generation 2 sister-cells, regardless of their biological condition
388 (self-renewing or differentiating for 48hrs), were always closer to each other
389 than randomly paired cells (Figure 5 - Student t-test for self-renewing T2EC
390 cells pvalue = 0.0146 and for differentiating T2EC cells pvalue = 0.003503).
391 Furthermore, the mean Manhattan distances of the generation 2 cousin-cells
392 were also statistically smaller than those of non related cells for both biologi-
393 cal conditions, indicating a proximity of transcriptomes which persisted after
394 one more cell generation in both conditions, observed separately (Student t-
395 test for self-renewing T2EC cells pvalue = 0.00002313 and for differentiating
396 T2EC cells pvalue = 0.003912).

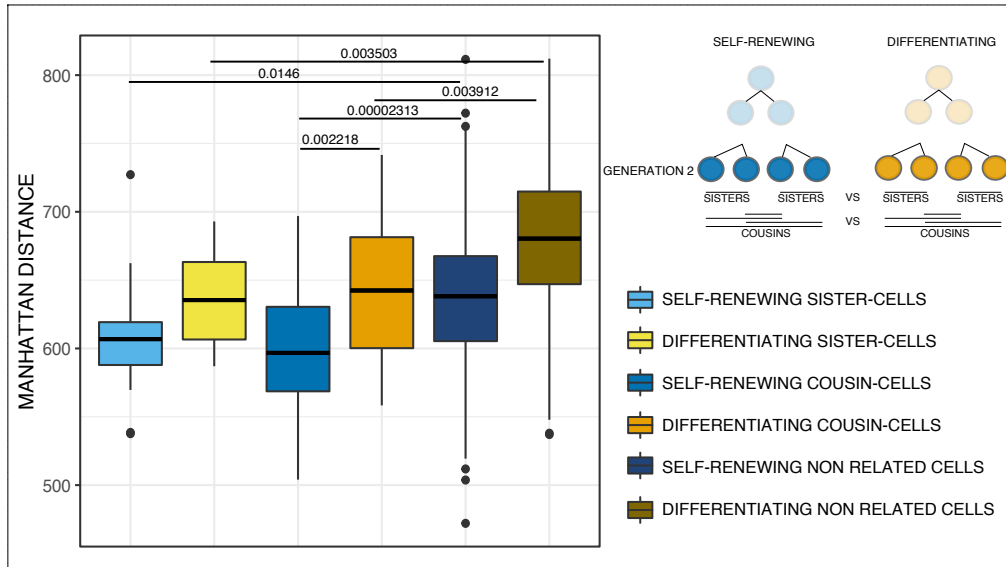


Figure 5: Manhattan distances comparison between generation 2 sisters, cousins and non related T2EC cells.

Boxplot of Manhattan distances between generation 2 sisters, cousins and non related T2EC cells. Manhattan distances were computed between all cells (32 self-renewing and 20 differentiating cells) from the same biological condition using the 983 selected genes. Self-renewing generation 2 sister-cells (16 pairs) are presented in light blue, self-renewing generation 2 cousin-cells (32 pairs) are in medium blue and self-renewing non related cells (448 pairs) are in dark blue. Differentiating generation 2 sister-cells (10 pairs) are in yellow, differentiating generation 2 cousin-cells (20 pairs) are in orange and differentiating non related cells (160 pairs) are in brown. Statistical comparisons were performed using Student t-test.

397 **Identification of genes subject to transcriptional mem-**
 398 **ory**

399 We expected that the transcriptomic similarities observed may concern a
 400 subset of genes, the "memory genes", which expression would be variable
 401 between couples of cells but correlated within them. Thus, we applied a
 402 "gene-wise" approach to identify genes subjected to transcriptional memory
 403 using a linear model with random variable and a mixed effects model. For

404 CD34+ cells, memory genes were identified using a linear model with a sister-
 405 hood random variable to capture sister-cells correlation. For T2EC cells, the
 406 expression of each gene was modeled by an additive model combining a fixed
 407 condition effect (differentiating or not) to account for difference in expres-
 408 sion level and a sisterhood random effect capturing sister-cells correlation.
 409 Memory genes were selected by testing the random effect using a likelihood
 410 ratio test comparing the model with and without the sisterhood effect. The
 411 test was performed on each gene followed by a Benjamini-Hochberg p-value
 412 adjustment for multiple testing. As a negative control, we performed the
 413 same test on randomly paired cells that detected no memory gene (Figure
 414 6).

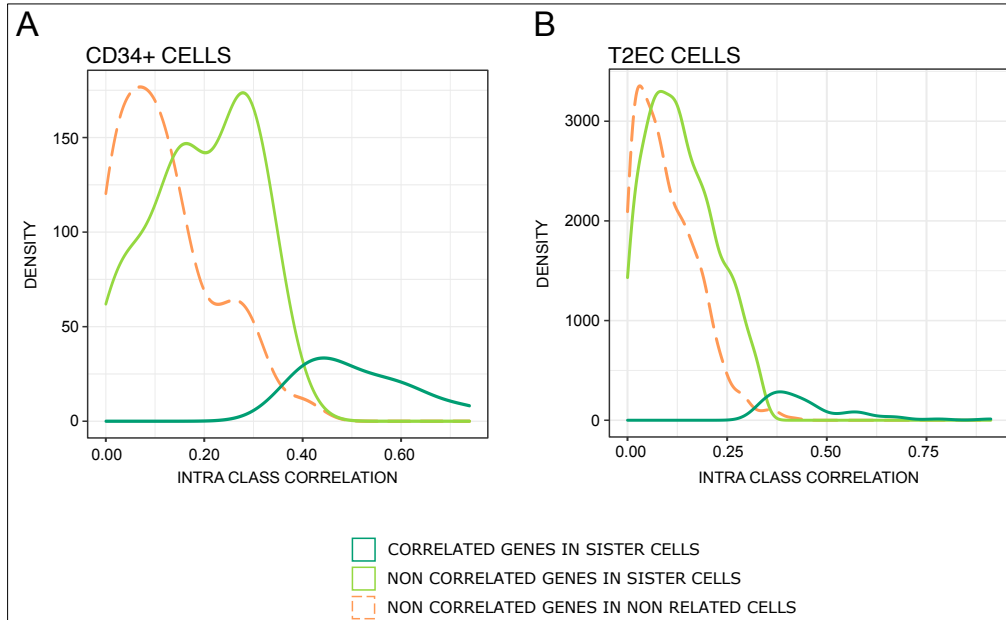


Figure 6: Density plot of genes correlation in generation 1 sister-cells and randomly paired CD34+ cells (A) and T2EC cells (B).

Identification of memory genes using a linear model with random variable (CD34+) and mixed effect model (T2E). Memory genes are in dark green (11 genes for the 86 CD34+ cells, 55 genes for the 104 T2EC cells), uncorrelated genes are in light green (72 for CD34+ cells, 1022 for T2EC cells), no memory genes were identified when cells were randomly paired (orange curve).

415 We detected 10 genes with significant between-sisters correlation in CD34+

416 cells and 55 genes in T2EC cells (cf. Supplement Table S1 for CD34+ and
417 for T2EC). In CD34+ cells, memory genes were involved in diverse functions,
418 including stemness (*GATA1*, *CD38*, *CD133*), differentiation and proliferation
419 (CD74, ERG, KIT), metabolism (*BCAT1*, *HK1*), cytoskeleton (*ACTB*) and
420 tRNA splicing (*C22orf28*). In T2EC, memory genes were involved in ery-
421 thropoietic differentiation (*HBBA*, *HBA1*, *HBAD*, which are hemoglobin sub-
422 units, or *RHAG* membrane channel component involved in carbon dioxide
423 transport), chromosome structure (*SMC2*, *H2AFZ*), ribosomes and trans-
424 lation (*RPS13*, *RPL22L1*, *UBA52*, *EEF1A1*) and metabolism (*GAPDH*,
425 *LDHA*). One should note that *LDHA* was previously found to also be in-
426 volved in the erythroid differentiation process [29].

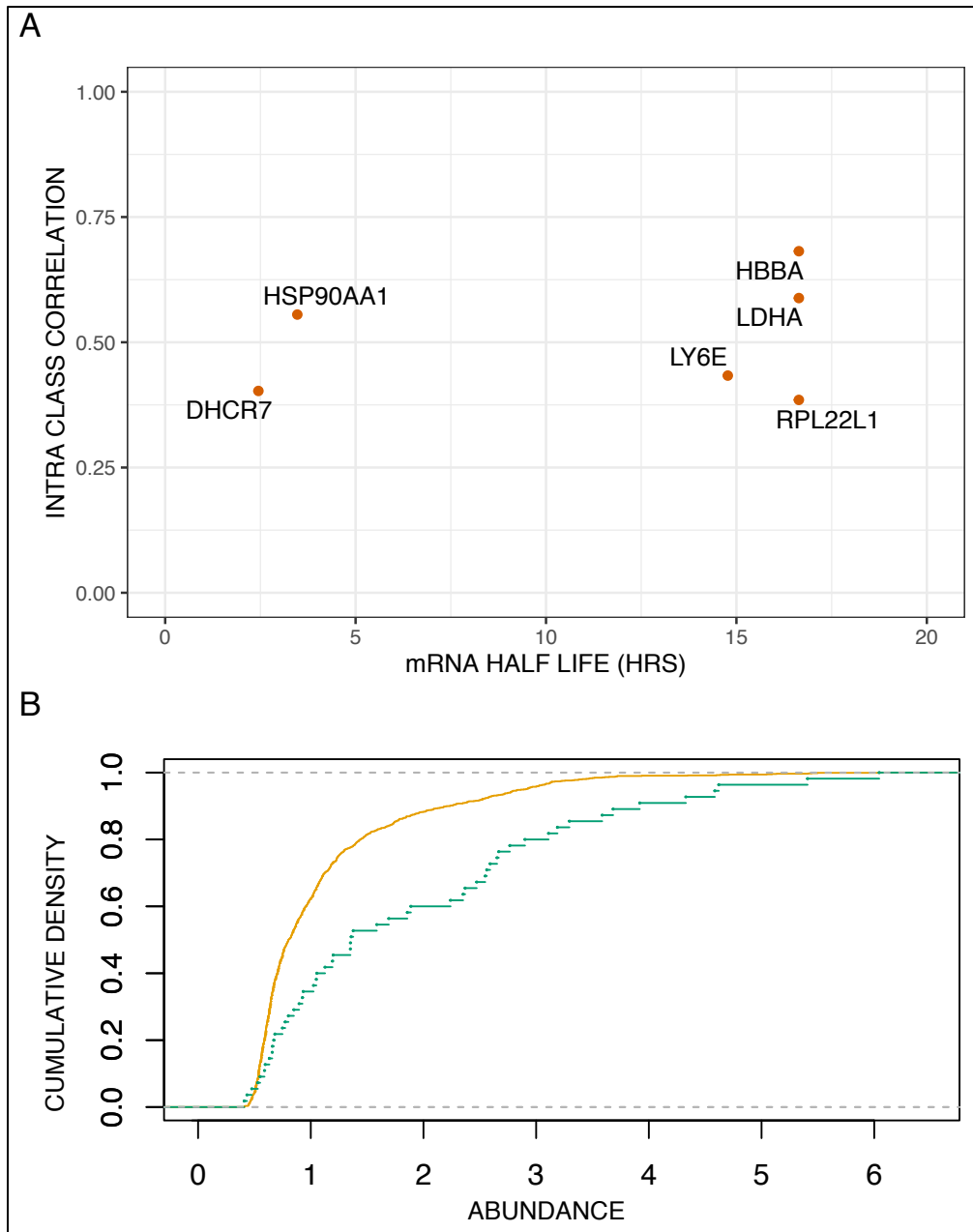


Figure 7: T2EC Memory genes characteristics.

(A) mRNA half-life of memory genes evaluated at 24hrs post differentiation induction in hours ([38]) vs their Intra Class Correlation value extracted from the mixed effects model. (B) Cumulative empirical distribution graph of abundance of the 55 Memory genes (green) compared to total genes (1177) of scRNA-seq data (yellow).

427 At that stage, we wondered if this memory could be explained by long
428 mRNA half-life of the identified genes. We crossed our gene list to a pre-
429 viously published dataset which evaluated half-life duration of genes during
430 T2EC differentiation using RT-qPCR [38]. We were able to compare the half-
431 life duration of 6 memory genes and found that 4 of them have a relatively
432 long half-life but 2 of them have a quite short half-life (Figure 7A). Further-
433 more, other genes with longer half-life were not identified by the model as
434 memory genes. Thus, half-life duration could not be the only cause of mem-
435 ory. We then questioned if there could be a relationship between the level
436 of expression of a gene and their belonging to the memory genes class. 1000
437 bootstrap distribution analysis of the abundance of the 55 memory genes
438 compared to the abundance of 55 randomly drawn genes showed an enrich-
439 ment for higher abundance of the 55 memory genes (Figure 7B - kolmogorov-
440 Smirnov test pvalue = 0.01672). We therefore can not exclude that part of
441 the memory is due to high level expression for at least some memory genes
442 and could be related to synthesis and degradation dynamics. However, this
443 result was expected because to prevent false correlation that would be due
444 to high numbers of zeros in expression value of lowly expressed genes be-
445 tween sister-cells, we selected genes with mid to high-level of expression in
446 our scRNA-seq data set (see methods). Finally, we didn't regress cell-cycle
447 effects on our data, due to the fact that cell-cycle is not as well described in
448 chicken cells as it is in mammalian cells, and thus cannot exclude that the
449 sister-to-sister resemblance may, in part, be a consequence of the sister-cells
450 being at similar state in the cell-cycle. However, while we found a GO term
451 "cell-cycle" enrichment in the 1177 selected genes, no cell-cycle related genes
452 were identified as memory genes, leading us to believe that cell-cycle is not
453 the main driver of this transcriptional memory.

454 Conclusion

455 In the present study, we addressed the interplay between transcriptional
456 memory and gene expression variability which characterizes differentiation
457 processes.

458 We developed two methods to recover sister-cells (Generation 1) and one
459 method to recover cousin-cells (Generation 2) transcriptomes while preserv-
460 ing the information about their lineage at the resolution of the cell division.
461 We analyzed the transcriptome of related cells from two different cell differ-
462 entiation systems using two different single-cell transcriptomics technologies.

463 Comparison of global transcriptomic state using Manhattan distances
464 showed that differentiating generation 1 sister-cells (both CD34+ cells and
465 T2EC cells) transcriptomes are globally significantly more similar between
466 each other than between non related cells.

467 In our controlled differentiation model (T2EC cells), we observed after
468 one cell division (generation 1), a greater mean distance for differentiating
469 sister-cells compared to self-renewing sister-cells. Moreover, the difference
470 becomes significant after a second division (generation 2), showed by dif-
471 ferentiating cousin-cells presenting a significantly higher distance than self-
472 renewing cousin-cells. Those results show that during cell differentiation,
473 related cells deviates faster from each other than during self-renewing divi-
474 sions.

475 Mixed effect models and linear models with random variable analysis fur-
476 ther highlighted that some genes have their expression statistically correlated
477 between sister-cells while none were found between non related cells. We
478 qualified those genes as "memory genes" and suspect that they may weight
479 out the transcriptomic resemblance observed between sister and cousin-cells.
480 However, the mechanisms leading to a more correlated expression between
481 related cells for those genes remain to be investigated.

482 In the introduction, we formulated 3 hypothesis on the possible behavior
483 of transcriptional memory behavior upon differentiation induction (Figure
484 1). Our results therefore support the second hypothesis: upon differentia-
485 tion induction, transcriptional memory is continuously and gradually erased
486 eventually reconstituting, at the clonal scale, the variability observed in the
487 initial population.

488 Discussion

489 At the single cell level, gene expression is in essence a probabilistic process
490 that is characterized by a given burst frequency and burst size [39]. The
491 mechanisms regulating this bursting process are still a matter of debate [40,
492 41], but are usually thought to involve: 1) the state of the underlying Gene
493 Regulatory Network (GRN) [42]; 2) the state of the chromatin, a.k.a. the
494 epigenetic marks [7, 8], and 3) the genomic 3D state [43]. Of course none of
495 these mechanisms operate in isolation and more integrated mechanisms, like
496 the metabolism, are also key players in the burst properties of transcription
497 (see e.g. [44]).

498 In order to explain the existence of memory genes as we (this work) and
499 others [5–8] have described, one need to assume that a significant fraction of
500 those mechanisms must “survive” the mitosis, i.e. be transmitted through the
501 dramatic epigenomic and cellular rearrangements involved in the cell division
502 process. If one assumes that the GRN state is essentially characterized by
503 protein quantities, then it is easy to see that it will be pass through, at least
504 for the proteins with a sufficiently long half life [17]. Reestablishment of the
505 epigenetic marks [45] and of genomic structure [46] after a division process
506 have also been documented.

507 It has recently been described that the persistence of a low level of tran-
508 scription throughout the mitosis might at least partly explain how transcrip-
509 tional memory can be maintained. It would be interesting in that regard, to
510 assess the overlap between our memory genes and these genes for which the
511 mitotic transcription can be detected using UEseq in mitotic chromosomes
512 [47].

513 Differentiating division is a specific challenge since at each division a
514 subtle combination of changes and stability must be imposed. In this respect
515 one can see the bookmarking process [48] as a stabilizing process, whereas
516 the increase in gene expression variability [13–21] will affect the GRN state
517 and therefore will tend to modify gene expression burst parameters.

518 It is interesting to note that our two model systems do behave quite
519 differently in regard to the division process. The initial stages of T2EC
520 erythrocytic differentiation have been shown to result in an increase of the
521 proliferation rate due to a shortening of the G1 period [27]. This is in sharp
522 contrast with the observation that the CD34+ first division occurs after an
523 unusually long cell cycle that lasts on average more than 55 hrs [17]. It
524 could therefore be that the molecular mechanisms linking cell division and

525 differentiation might be quite different in the two cell types, although the end
526 result will be similar: cellular memory will show a high level of robustness in
527 front of the cellular state change associated with the differentiation process.

528 Finally, it is tempting to speculate that the observed burst in entropy
529 at the beginning of the differentiation sequence is helping the differentiating
530 cells to overcome a memory process that is meant to prevent changes in
531 cellular identity.

532 **Material and methods**

533 **Cell culture**

534 Human hematopoietic CD34+ cells were purified from umbilical cord blood
535 from three anonymous healthy donors. First, mononuclear cells were isolated
536 by density centrifugation using Ficoll (Biocoll, Merck Millipore). CD34+
537 cells were then enriched by immunomagnetic beads using the AutoMACSpro
538 (Miltenyi Biotec). Cells were frozen in 90% fetal bovine serum (Eurobio)
539 10% dimethylsulfoxide (Sigma) and stored in liquid nitrogen. After thawing,
540 cells were grown in prestimulation medium made of Xvivo (Lonza) supple-
541 mented with penicillin/streptomycin (respectively 100U/mL and 100 μ g/mL
542 - Gibco, Thermo Scientific), 50 ng/ml h-FLT3-ligand, 25 ng/ml h-SCF, 25
543 ng/ml h-TPO, 10 ng/ml h-IL3 (Miltenyi) final concentration as previously
544 described [17]. Cells were cultured in a 96-well plate at 185 000 cells/mL
545 during 24hrs in a humidified 5% CO₂ incubator at 37°C before proceeding
546 to mother cells isolation.

547

548 T2EC cells were extracted from 19-days-old SPAFAS white leghorn chicken's
549 embryos' bone marrow (INRA, Tours, France). Cells were grown in LM1
550 medium (α -MEM, 10% Fetal bovine serum (FBS), 1 mM HEPES, 100 nM
551 β -mercaptoethanol, 100 U/ mL penicillin and streptomycin, 5 ng/mL TGF-
552 α , 1 ng/mL TGF- β and 1 mM dexamethasone) as previously described [27].
553 T2EC cells differentiation was induced by removing LM1 medium and placing
554 the cells into DM17 medium (α -MEM, 10% fetal bovine serum (FBS), 1 mM
555 Hepes, 100 nM β -mercaptoethanol, 100 U/mL penicillin and streptomycin,
556 10 ng/mL insulin and 5% anemic chicken serum [28]).

557 **Manual strategy for CD34+ sister-cells isolation**

558 Mother cells were isolated using a SmartAliquoter (iBioChips). It consists
559 of a polydimethylsiloxane chip divided into 100 wells (2 μ L per well, 1.8mm
560 of diameter) connected by microchannels to an insertion hole in the center.
561 This system allows to physically isolate cells while sharing the same medium.
562 200 μ L of cell suspension at 1000 cells/mL were injected in the chip through
563 the injection plug and cells were randomly divided into the wells. Air bubbles
564 were removed with sterile tips. Using a standard confocal microscope, wells
565 containing lonely cells were listed. 20mL of prestimulation medium (see Cell

566 culture part for composition) were added to avoid evaporation and cells were
567 incubated at 37°C in a humidified 5% atmosphere during 24 to 48hrs. Listed
568 wells were regularly checked with standard confocal microscope to identify
569 cell division. Sister-cells were manually collected under biological safety cab-
570 inet to keep sterile conditions and avoid impurities to fall in the culture dish.
571 A micromanipulator connected to a flexible microfluidic capillary filled with
572 PBS and ending in a 2 μ L glass microcapillary was used. Individual collected
573 cells were immediately inserted into 5 μ L of lysis buffer (Triton 4% (Sigma),
574 RNaseOUT Recombinant Ribonuclease Inhibitor 0.4U/ μ L (Thermo Scien-
575 tific), Nuclease free water (Thermo Scientific), Spikes 1 and 4 (Fluidigm C1
576 Standard RNA Assays)) and kept on dry ice to preserve RNA. Particular
577 attention has been given to preserve cells integrity. Samples were kept at
578 -20°C until further sc-RT-qPCR analysis.

579 **FACS-oriented strategy for T2EC sister-cells isolation**

580 Mother cells were stained using CFSE (Cell Trace CFSE Cell Proliferation
581 kit ThermoFisher), 5x10⁵ cells were placed in a 60mm plate in 5mL of culture
582 medium mixed with 5 μ L of CFSE at 5 mM (final concentration 5 μ M) and
583 incubated at 37°C for 30min. Cells were then centrifuged at 20°C, 1500rpm
584 for 5min. Medium was discarded and cells were resuspended in 5mL fresh
585 medium. CFSE stained mother cells were then isolated using the CellenONE
586 X1 (CELLENION) at CELLENION core facility (Lyon, France). A gating
587 based only on morphological criteria (diameter, elongation and circularity)
588 was performed to select single living cells. Selected single cells were sorted
589 in a 384-well plate containing 10 μ L of culture medium (either self-renewing
590 medium LM1 or differentiation-inducing medium DM17). The plate was then
591 kept in an incubator under 5% CO₂, 37°C for at least 20hrs to allow one cell
592 division. Each well of the 384-well plate was manually checked under a regu-
593 lar inverted microscope to identify cells that had undergone one cell division
594 (presence of cell doublets). Each doublet was then harvested and placed
595 in a FACS polypropylene tube containing 80 μ L of warm culture medium.
596 Tubes containing cell doublets were kept at room temperature throughout
597 the sorting process and were briefly vortex immediately before loading into
598 the sorter. Prior settings consisted in analysing the CFSE positive popula-
599 tion, the CFSE negative population and the culture medium. No fluorescent
600 signal was ever detected in medium or in negative population (Figure S1 A-B
601 self-renewing medium and C-D differentiation medium) indicating that only

602 cells of interest ever gave CFSE positive signal. Cells were sorted at 20 PSI
603 through a 100 μm nozzle on an AriaII FACS (BD). Gating was performed on
604 FSC-A/SSC-A to capture live cells, SSC-H /SSC-A to capture single cells,
605 and CFSE positive cells with yield, purity and phase mask of 32, 0, 0 respec-
606 tively. Those parameters were chosen because cell density being very low
607 (2 cells per tube), the probability of the two cells being in two consecutive
608 drops was extremely low. Furthermore, those parameters are very conserva-
609 tives and thus probability of the cell not being sorted is also very low. Cells
610 were isolated in 4 μL of lysis buffer in PCR tubes containing cell barcode
611 primers. Tubes were frozen in dry ice directly after sorting to prevent any
612 degradation of the samples.

613 **FACS-oriented strategy for T2EC cousin-cells isolation**

614 **Fluorescent barcoding for lineage tracing**

615 On the first day, 1×10^6 mother cells were labelled with 0.5 μM CTV (Cell
616 Trace Violet Cell Proliferation kit Thermofisher) for 20min at 37°C in PBS,
617 then 5mL of medium was added for 5min to dilute the fluorescent molecules.
618 The cells were centrifuged for 5min at 1500rpm at 20°C, resuspended and then
619 separated into 6 tubes (2×10^5 cells per tube) and resuspended in 1mL per
620 tube. Each sample was labelled with a different concentration of CFSE (3-
621 point range of 5 μM , 2.187 μM and 0.312 μM) plus or minus CTY (10 μM - Cell
622 Trace Yellow Cell Proliferation kit Thermofisher) for 30min at 37°C in PBS.
623 Each condition was centrifuged for 5min at 1500rpm at 20°C and resuspended
624 in 1mL of fresh medium. The different concentrations and combinations were
625 optimised so that even after two cell divisions, the barcodes will be different
626 enough to differentiate the cell clones. Cells were plated in a 6-well plate
627 and kept in culture conditions until sorting (in an incubator 37°C, 5% CO₂).
628 Cells were were stored at 37°C throughout the sorting process and sorted
629 at 20 PSI through a 100 μm nozzle on an AriaII FACS (BD). The sorting
630 strategy was done using single-labelled cell populations (CFSE, CTY, CTV
631 and negative), then gating was performed on FSC-A/SSC-A to capture live
632 cells, SSC-H /SSC-A to capture single cells, and CTV positive cells. One
633 cell from each subgroup (6 cells total) was isolated in a well of a 96-well
634 plate which contained 500 non-labelled feeder cells in either self-renewing
635 medium or differentiating medium through a 100 μm nozzle with yield, purity
636 and phase mask of 0, 32, 16 respectively (single-cell mask). A well then

637 contained 6 mother cells, each one labelled with a unique fluorescent barcode
638 and the feeder cells. The plate was then put back in culture conditions (in
639 an incubator 37°C, 5% CO₂).

640 CTFR (Cell Trace Far Red Proliferation kit Thermofisher) labelling was
641 performed 20hrs after mother cells sorting, in the plate, so that the cells had
642 time to divide once. The staining was made as heterogeneous as possible,
643 thanks to the feeder cells but also by using very low concentrations of dye
644 and for a very short amount of time. Indeed, 0.37 μ M of CTFR (Cell Trace
645 Far Red Cell Proliferation kit Thermofisher) was added to each sample (in
646 approximately 50 μ L of medium), and then 100 μ L of medium was added to
647 dilute the dye. The plate was centrifuged for 5min at 200G, then 120 μ L
648 of medium was removed and 50 μ L of new medium added to each labelled
649 well. This heterogeneous CTFR staining will allow to discriminate the next
650 division meaning within the 4 cousin-cells, how they are paired two by two.
651 Indeed, each daughter-cell will receive a unique intensity of CTFR dye which
652 will be discriminating after one more cell division. Cells were kept in culture
653 conditions for an additional 20hrs (in an incubator 37°C, 5% CO₂).

654 On the third day, after the second division, the content of the wells con-
655 taining the cousin-cells were transferred into polypropylene FACS tubes and
656 briefly vortexed immediately before loading into the sorter. The sorting
657 strategy was done using single-labelled cell populations (CFSE, CTY, CTV,
658 CTFR and negative), then gating was performed on FSC-A/SSC-A to cap-
659 ture live cells, SSC-H /SSC-A to capture single cells, and CTV positive cor-
660 responding to the second division peak and exclude feeder cells. Cells were
661 sorted on a AriaII FACS (BD) at 20 PSI through a 100 μ m nozzle with yield,
662 purity and phase mask of 32, 16, 0 respectively, in PCR tubes containing ly-
663 sis buffer (0.2% Triton (Sigma Aldrich), 0.4 U/ μ L RNaseOUT (Thermofisher
664 Scientific), 400nM RT primers (Sigma Aldrich)) and scRNA-seq primers. The
665 fluorescent intensities for CFSE, CTY and CTFR were recorded for each cell
666 to further reconstruct relationships between the cells using our clustering
667 algorithm.

668 **Cousin-cells identification**

669 Clustering was performed using the R mclust package [49] (version 5.4.10).
670 This clustering script finds the genealogical relationships between cells in two
671 steps. First, cousin-cells are grouped together by their fluorescent barcode,
672 determined by the CTFE and CTY fluorescent intensity values. Thus, if two,

673 three or four cells have the same CFSE and CTY intensities levels they will
674 be considered as cousins. In a second step, if all 4 cousin-cells of the group
675 were sorted in the plate, the program identifies the two pairs of sisters within
676 the 4 cousins. To do this, the median CTFR intensity is calculated, then the
677 two cells that have intensity values higher than the median are matched, and
678 the other two that have lower intensity values are matched together. Finally,
679 when sorting, we used a index sorting option, which allows us to know in
680 which well of the plate each cell was sorted. With this position information,
681 our analysis program returns the position of the retained cells, i.e. the cells
682 belonging to the cousin groups for which the 4 cells were successfully isolated
683 in the lysis plate.

684 **sc-RT-qPCR data generation**

685 **sc-RT-qPCR one step**

686 Lysed cells were heated at 65°C during 3 minutes for hybridization with
687 RT primer and immediately transferred into ice. 7µL of RT-PCR mix (Su-
688 perscript III RT/platinum Taq 0,1µL (Invitrogen), Reverse and Forward
689 primers and spikes at 1,33µM final concentration and homemade 2X reaction
690 Mix (120mM Tris SO4 pH=9, 2.4 mM MGSO4, 36mM (NH4)2SO4, 0.4mM
691 dNTP)) were added to each well before launching of reverse transcription
692 and PCR run on thermocycler (Program : 50°C 15min - 95°C 2min - 20
693 cycles 95°C 15sec/60°C 4min - Hold 4°C). 3µL of exonuclease mix (Exonu-
694 clease I 1.6U/mL (NEB), Exonuclease buffer 1X (NEB), Nuclease free water
695 (Thermo Scientific)) were added and samples were incubated for a digestion
696 run on thermocycler (Program : 37°C 30min - 80°C 10min). Pre-amplified
697 samples were diluted five times in TE low EDTA (10mM Tris, 0.1mM EDTA,
698 pH=8) and kept at -20°C for one night before qPCR.

699 **qPCR with Fluidigm Biomark technology**

700 3,15µL of pre-amplified samples were distributed into a 96-well plate and
701 3,85µL of qPCR mix (Sso EvaGreen Supermix with Low ROX (Bio-Rad)+
702 20X DNA binding dye sample loading reagent) were added to each well.
703 Simultaneously, a 96-well plate with primer mix (forward and reverse primers
704 and spike at 2µM final concentration, 2X Assay Loading reagent, TE low
705 EDTA) was prepared. The microfluidigm chip was primed with injection oil

706 using the IFC Controller HX (Fluidigm). 5 μ L of primers and 5 μ L of samples
707 were loaded in the dedicated wells of the chip. Air bubbles were removed
708 with a needle. Samples and primers were mixed in the IFC Controller HX
709 (Fluidigm) with the loading program. The chip was then transferred in
710 the Biomark HD system (Fluidigm) for qPCR with "HE 96x96 PCR+Melt
711 v2.pcl" thermal cycling protocol with auto exposure.

712 **Quality control and Normalization**

713 Ct values obtained from the Biomark HD System (Fluidigm) were exported
714 as excel files and quality control was manually done. For each gene, "failed"
715 quality control readings identified by the Fluidigm software were removed.
716 Four negative controls (mix of water and lysis buffer) were used to detect
717 unwanted amplification and the associated genes were also removed. Finally,
718 two externally added controls (spike 1 and spike 4, Fluidigm) were used to
719 control amplification consistency. Filtered data frame was then imported into
720 R (version 4.2.0) for normalization to remove amplification bias. For each
721 cell, expression values were calculated by subtracting the gene Ct value from
722 the geometric mean of Ct values from spike 1 and spike 4 of the corresponding
723 well. Then, an arbitrary differential cycle threshold value of -22 for null signal
724 (corresponding to a Ct value of 30) was assigned for all genes with a Ct value
725 less than -22.

726 **scRNA-seq data generation**

727 **scRNA-seq libraries preparation**

728 Subsequently to sister or cousin-cells isolation, we performed single cell RNA
729 sequencing (scRNA-seq) using a modified version of the Mars-seq protocol
730 [30] published here [32]. This specific protocol of scRNA-seq allowed us to
731 know in advance which cell barcode would be carried by each cell and thus
732 preserving the genealogy information of the cells. Briefly, Reverse Transcrip-
733 tion (RT) was performed so every mRNA of the cells was tagged with a
734 combination of unique cell barcode and a 8pb random UMIs sequence for
735 further demultiplexing. After barcoding, all mRNA were pooled and second
736 DNA strand were synthesized. Amplification was done over night using In
737 Vitro Transcription (IVT) to obtain a more linear amplification. A second
738 barcode was added by RT to identify plates. Libraries were amplified by

739 PCR and Illumina primers were added.

740 **Sequencing**

741 Libraries were sequenced on a Next500 sequencer (Illumina) with a custom
742 paired-end protocol (130pb on read1 and 20pb on read2) and a depth of 200
743 000 raw reads per cell.

744 **Data preprocessing**

745 Fastq files were pre-processed through a bio-informatics pipeline developed
746 in our team on the Nextflow platform [50] and published here [32]. Briefly,
747 the first step removed Illumina adaptors. The second step de-multiplexed
748 the sequences according to their plate barcodes. Then, all sequences con-
749 taining at least 4T following the cell barcode sequence and UMI sequence
750 were kept. Using UMIttools whitelist, the cell barcodes and UMI sequences
751 were extracted from the reads. The cDNA sequences were then mapped on
752 the reference transcriptome (Gallus GallusGRCG6A.95 from Ensembl) and
753 UMIs were counted. Finally, a count matrix was generated for each plate.

754 **Quality control and data filtering**

755 All analysis were carried out using R software (version 4.1.2; [51]). For the
756 sister-cells dataset, cells were filtered based on several criteria: reads number,
757 genes number, counts number and ERCC content. For each criteria the cut
758 off values were determined based on SCONE [52] pipeline and were calculated
759 as follows:

760 $\text{Mean}(\text{parameter}) - 3 * \text{sd}(\text{parameter})$

761 We then removed orphan cells, meaning cells which sister was not present in
762 the dataset. After filtering, we kept 60 undifferentiated cells (30 couples) and
763 64 differentiating cells (32 couples). For the cousin-cells dataset we performed
764 the same filtering strategy and kept only cell groups which contained the 4
765 cousin-cells. After filtering we kept 32 undifferentiated cells (8 groups of
766 cousins) and 20 differentiating cells (5 groups of cousins). Based on [53]
767 work, genes were kept in the data set if in mean present in every cell. After
768 applying this filter, we kept 1177 and 983 genes for the sister-cells dataset
769 and the cousin-cells dataset respectively.

770 **Normalization**

771 Filtered matrix were normalized using SCTransform from Seurat package
772 (version 1.6 [54]) and were corrected for batch effect, day of isolation effect,
773 medium effect and sequencing depth effect. Both datasets (sister-cells and
774 cousin-cells) were processed independently.

775 **Bioinformatics analysis on R**

776 All analysis were carried out using R software [51] (version 4.1.2 for T2EC
777 and version 4.2.0 for CD34+). Plots were performed ggplot2 package (version
778 3.3.6).

779 **Dimensional reduction**

780 UMAP dimension-reduction and visualization were performed using UMAP
781 package (version 0.2.8.0; [55]).

782 **Manhattan distance computation**

783 Distances were computed on normalized matrix between all cells using dist
784 function from R software. Distances between sister-cells were extracted and
785 compare to the same number of randomly chosen distances of non related
786 cells. 1000 bootstraps were performed this way. Mean comparison was per-
787 formed using Student t-test or Wilcoxon test when Student t-test was not
788 applicable.

789 **Linear model with random variable and Mixed effects model**

790 Linear model with random variable and Mixed effects model analysis were
791 performed using lme4 R package (version 1.1-29). The models were defined
792 as followed:

793 Mixed effect Model definition :

$$Y = p1 + p2 + e$$

794 Linear Model with random variable definition :

$$Y = p2 + e$$

795

796 where Y is the mean expression of each gene, p_1 is the fixed effect and p_2 is
797 the random effect. Here, p_1 corresponds to the biological condition and can
798 take two values (undifferentiated and differentiating) and p_2 is the sorority
799 effect. Two sister-cells have the same discrete value. And e is the error of
800 the model. Null models are the above model but without the random effect
801 e.g. the sorority effect. Genes were selected based on a significant adjusted
802 BH p-value after performing a likelihood ratio test between the model and
803 the null model.

804 **Abbreviations**

805 **Declarations**

806 **Ethics approval and consent to participate**

807 Human cord blood (UCB) was collected from placentas and/or umbilical
808 cords obtained from AP-HP, Hôpital Saint-Louis, Unité de Thérapie Cellu-
809 laire, CRB-Banque de Sang de Cordon, Paris, France (Authorization num-
810 ber: AC-2016-2759) or from Centre Hospitalier Sud Francilien, Evry, France
811 in accordance with international ethical principles and French national law
812 (bioethics law n°2011-814) under declaration N° DC-201-1655 to the French
813 Ministry of Research and Higher Studies.

814 **Consent for publication**

815 Not applicable

816 **Availability of data and materials**

817 Data tables are supplied as supplements files. scRNA-seq data are avail-
818 able in SRA repository under the BioProject accession PRJNA882056 and
819 BioSample accessions SAMN30926136 and SAMN30926137. Embargo will
820 be released upon publication.

821 R codes are available at : <https://gitbio.ens-lyon.fr/LBMC/sbdm/sister-cells>
822 Embargo will be released upon publication.

823 **Competing interests**

824 The authors declare that they have no competing interests.

825 **Funding**

826 This work was supported by funding from the French agency ANR (SinCity;
827 ANR-17-CE12-0031).

828 **Authors' contributions**

829 CF and LR contributed equally to the conceptualization of the study, data
830 generation, data analysis, statistical analysis, and interpretation, as well as
831 writing the manuscript. CK and SD participated in data generation and
832 methods optimization. EV participated in data generation. AM RP DS FB
833 participated in the methods optimization. LM provided support for data
834 analysis and developed the clustering R script. FP provided support for
835 statistical analysis. OG obtained the funding. AP, OG and SG participated
836 to the conceptualization of the study, the project administration, the project
837 supervision, and writing of the manuscript. All authors read and approved
838 the final manuscript.

839 **Acknowledgements**

840 We gratefully thank all members of SBDM team for very fruitfull discus-
841 sions, suggestions and commentaries on our project. We thank the com-
842 putational center of IN2P3 (Villeurbanne/France) and Pôle Scientifique de
843 Modélisation Numérique (PSMN, Ecole Normale Supérieure de Lyon) where
844 computations were performed. We thank the BioSyL Federation and the
845 LabEx Ecofect (ANR-11-LABX-0048) of the University of Lyon for inspir-
846 ing scientific events. We acknowledge the contributions of the CELPHEDIA
847 Infrastructure (<http://www.celphedia.eu/>), especially the center AniRA in
848 Lyon.

849 **Supplementary**

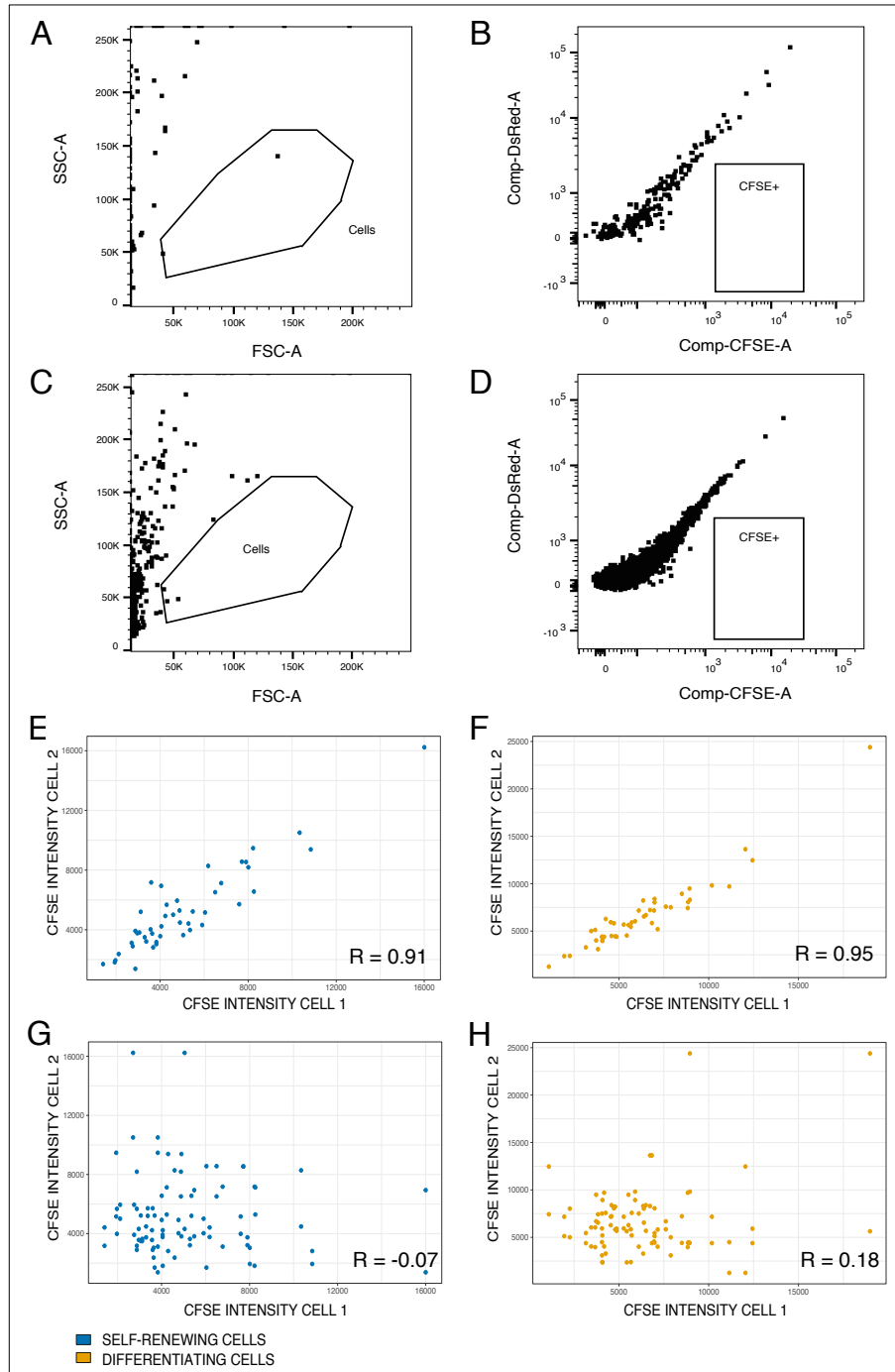


Figure S1: Technical validation of sister-cells isolation method using CFSE intensity data and evaluation of background noise for self-renewing and differentiating cells.

Figure S1: (A) Artefact detection in self-renewal medium. A few events are detected in the cell gate. (B) Artefact detection in self-renewal medium using CFSE signal (488nm, emission 530/30nm) versus auto-fluorescence (488nm, emission 585/42nm). No events is detected in the CFSE positive gate. For graphs A and B all events are displayed. (C) Artefact detection in differentiation medium. A few events are detected in the cell gate. (D) Artefact detection in differentiation medium using CFSE signal (488nm, emission 530/30nm) versus auto-fluorescence (488nm, emission 585/42nm). No events is detected in the CFSE positive gate. For graphs A and B all events are displayed. (E) Analysis of CFSE intensity correlation between self-renewing sister-cells (Spearman $R = 0.91$). (F) Analysis of CFSE intensity correlation between differentiating sister-cells (Spearman $R = 0.95$). (G) CFSE intensity correlation of randomly paired self-renewing cells (Spearman $R = -0.07$). (H) CFSE intensity correlation of randomly paired differentiating cells (Spearman $R = 0.18$).

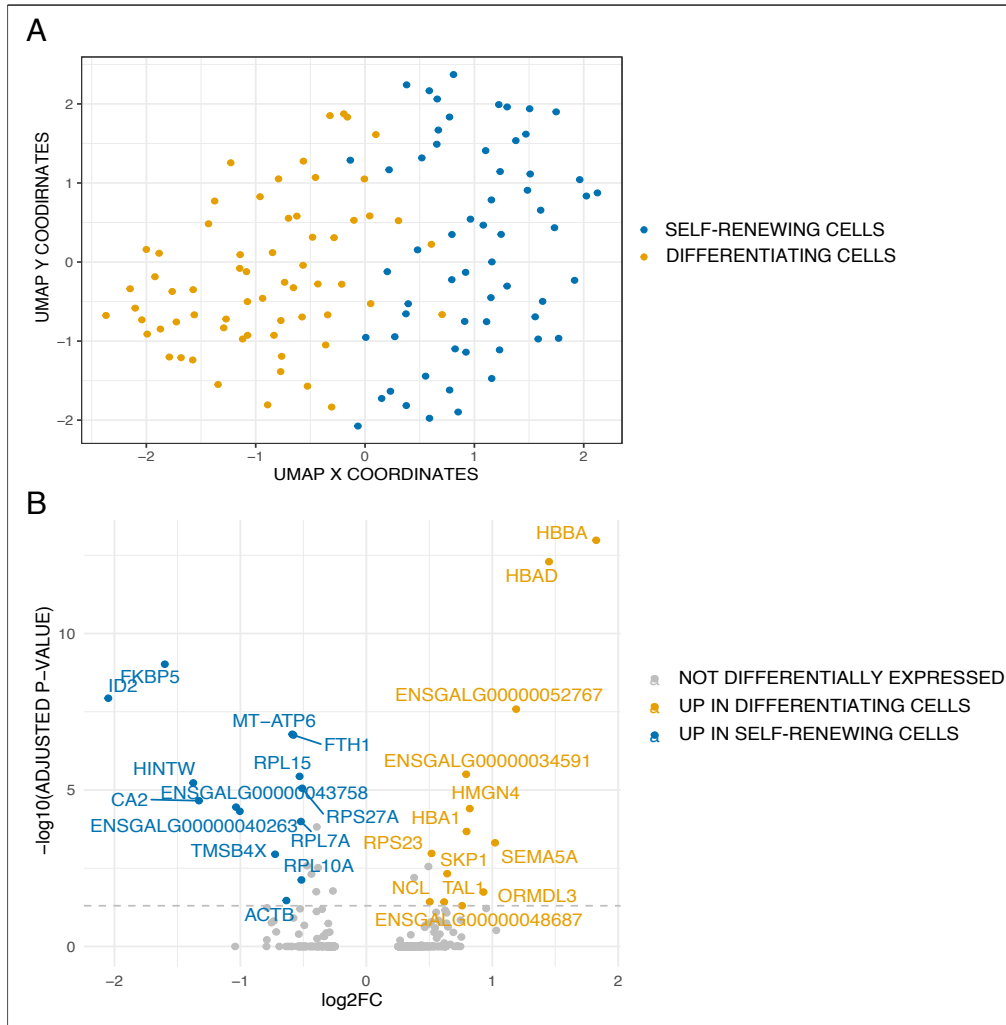


Figure S2: General structure of the data and characterization of the differentiation process.

(A) Dimensional reduction and projection with UMAP of the scRNA-seq data on T2EC cells, cells in self-renewal are in blue and differentiating cells are in yellow. (B) Volcano plot of genes differentially expressed between the two conditions. Genes are considered significantly differentially expressed when the fold change is equal or above 0,5 and adjusted p-value is below 0.05 (grey dotted line). Blue dots represent significantly up-regulated genes in self-renewing cells and yellow dots represent significantly up-regulated genes in differentiating cells.

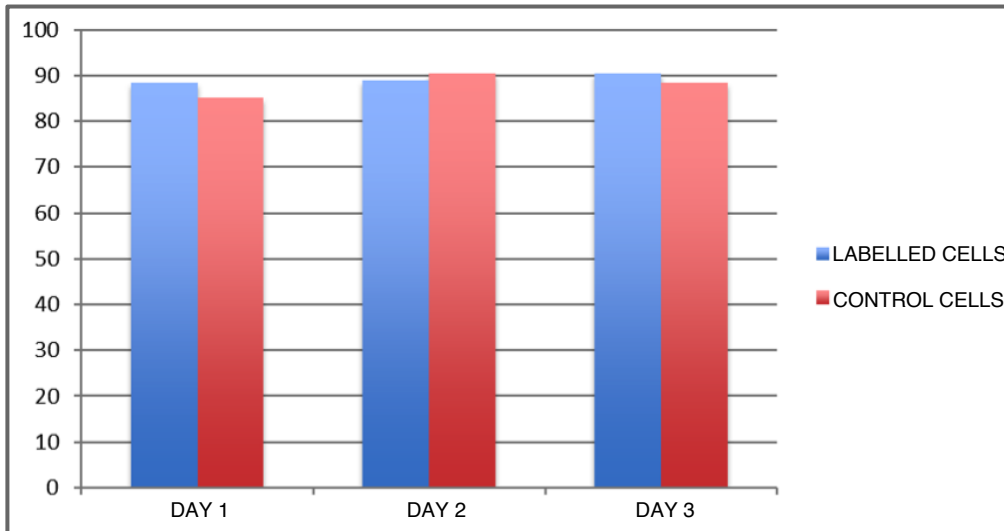


Figure S3: Histograms of cells viability. Histograms of the percentage of viability during the cells staining (after day 1, 2 and 3) of the fluorescently labelled cells compared to negative control cells. T-test showed no significant differences.

T2EC MEMORY GENES	
Gene_name	Ensembl_Gene_ID
ACTB	ENSGALG00000009621
ATP5G3	ENSGALG00000009286
ATP6V0C	ENSGALG00000009229
B2M	ENSGALG00000002160
CCNG1	ENSGALG00000001718
CD99	ENSGALG00000024488
CLTA	ENSGALG00000015326
DHRS7	ENSGALG00000011921
EEF1A1	ENSGALG00000015917
ENSGALG000	ENSGALG00000040263
ENSGALG000	ENSGALG00000043758
ENSGALG000	ENSGALG00000050548
ENSGALG000	ENSGALG00000052767
ENSGALG000	ENSGALG00000053077
ENSGALG000	ENSGALG00000053765
ESF1	ENSGALG00000034768
GAPDH	ENSGALG00000014442
H2AFZ	ENSGALG00000014023
HBA1	ENSGALG00000043234
HBAD	ENSGALG00000031597
HBBA	ENSGALG00000047152
HINTW	ENSGALG00000035998
HMGB2	ENSGALG00000010745
HSP90AA1	ENSGALG00000033212
ID2	ENSGALG00000035016
KPNA2	ENSGALG00000003584
LBR	ENSGALG00000009305
LDHA	ENSGALG00000006300
LY6E	ENSGALG00000041621
MLANA	ENSGALG00000019756
MRPS28	ENSGALG00000036749
MT-ATP6	ENSGALG00000041091
MT-COX3	ENSGALG00000035334
MT-ND2	ENSGALG00000043768
PLK1	ENSGALG00000006110
PPIA	ENSGALG00000028600
RHAG	ENSGALG00000016684
RPL13	ENSGALG00000006179
RPL22L1	ENSGALG00000009312
RPL37	ENSGALG00000014833
RPS23	ENSGALG00000015617
RTFDC1	ENSGALG00000007709
SAT1	ENSGALG00000016348
SEMA5A	ENSGALG00000028685
SH3BGRL3	ENSGALG00000038536
SMC2	ENSGALG00000015691
SOD1	ENSGALG00000015844
SPARC	ENSGALG00000004184
ST13P5	ENSGALG00000012007
TFRC	ENSGALG00000007485
TPD52	ENSGALG00000040167
TPX2	ENSGALG00000006267
TUBA1B	ENSGALG00000052192
UBA52	ENSGALG00000037716
UBE2I	ENSGALG00000006428

CD34+ MEMORY GENES	
Gene_name	Ensembl_Gene_ID
BCAT1	ENSG00000060982
GATA1	ENSG00000102145
HK1	ENSG00000156515
ACTB	ENSG00000075624
KIT	ENSG00000157404
CD38	ENSG00000044468
C2orf28	ENSG00000100220
ERG	ENSG00000157554
CD133	ENSG00000007062
CD74	ENSG00000019582

Table S1: Memory genes list with gene names and ENSEMBL gene ID.

850 References

- 851 1. Miura, H. & Hiratani, I. Cell cycle dynamics and developmental dy-
852 namics of the 3D genome: toward linking the two timescales. *Curr Opin*
853 *Genet Dev* **73**, 101898. ISSN: 1879-0380 (Electronic) 0959-437X (Link-
854 ing) (2022).
- 855 2. Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y.,
856 Rosenfeld, N., Danon, T., Perzov, N. & Alon, U. Variability and memory
857 of protein levels in human cells. *Nature* **444**, 643–646. ISSN: 0028-0836,
858 1476-4687 (Nov. 2006).
- 859 3. Schwanhäusser, B., Wolf, J., Selbach, M. & Busse, D. Synthesis and
860 degradation jointly determine the responsiveness of the cellular pro-
861 teome: Insights & Perspectives. *BioEssays* **35**, 597–601. ISSN: 02659247
862 (July 2013).
- 863 4. Corre, G., Stockholm, D., Arnaud, O., Kaneko, G., Viñuelas, J., Yam-
864 agata, Y., Neildez-Nguyen, T. M. A., Kupiec, J.-J., Beslon, G., Gan-
865 drillon, O. & Paldi, A. Stochastic Fluctuations and Distributed Con-
866 trol of Gene Expression Impact Cellular Memory. *PLoS ONE* **9** (ed
867 MacArthur, B. D.) e115574. ISSN: 1932-6203 (Dec. 22, 2014).
- 868 5. Kimmerling, R. J., Lee Szeto, G., Li, J. W., Genshaft, A. S., Kazer,
869 S. W., Payer, K. R., de Riba Borrajo, J., Blainey, P. C., Irvine, D. J.,
870 Shalek, A. K. & Manalis, S. R. A microfluidic platform enabling single-
871 cell RNA-seq of multigenerational lineages. *Nat Commun* **7**, 10220 (2016).
- 872 6. Phillips, N. E., Mandic, A., Omid, S., Naef, F. & Suter, D. M. Memory
873 and relatedness of transcriptional activity in mammalian cell lineages.
874 *Nature Communications* **10**, 1208. ISSN: 2041-1723 (2019).
- 875 7. Muramoto, T., Muller, I., Thomas, G., Melvin, A. & Chubb, J. R.
876 Methylation of H3K4 Is required for inheritance of active transcriptional
877 states. *Curr Biol* **20**, 397–406. ISSN: 1879-0445 (Electronic) 0960-9822
878 (Linking) (2010).
- 879 8. Shaffer, S. M., Emert, B. L., Reyes Hueros, R. A., Cote, C., Harmange,
880 G., Schaff, D. L., Sizemore, A. E., Gupte, R., Torre, E., Singh, A., Bas-
881 sett, D. S. & Raj, A. Memory Sequencing Reveals Heritable Single-Cell
882 Gene Expression Programs Associated with Distinct Cellular Behaviors.
883 en. *Cell* **182**. Number: 4 Reporter: Cell, 947–959.e17. ISSN: 00928674
884 (Aug. 2020).

- 885 9. Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E. & Huang, S.
886 Transcriptome-wide noise controls lineage choice in mammalian progen-
887 itor cells. en. *Nature* **453**, 544–547. ISSN: 0028-0836, 1476-4687 (May
888 2008).
- 889 10. Kalmar, T., Lim, C., Hayward, P., Munoz-Descalzo, S., Nichols, J.,
890 Garcia-Ojalvo, J. & Martinez Arias, A. Regulated fluctuations in nanog
891 expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol*
892 **7**, e1000149. ISSN: 1545-7885 (Electronic) (2009).
- 893 11. Kupiec, J. J. A Darwinian theory for the origin of cellular differentia-
894 tion. *Molecular & general genetics: MGG* **255**, 201–208. ISSN: 0026-8925
895 (June 1997).
- 896 12. Paldi, A. Stochastic gene expression during cell differentiation: order
897 from disorder? *Cellular and molecular life sciences: CMLS* **60**, 1775–
898 1778. ISSN: 1420-682X (Sept. 2003).
- 899 13. Hu, M., Krause, D., Greaves, M., Sharkis, S., Dexter, M., Heyworth, C.
900 & Enver, T. Multilineage gene expression precedes commitment in the
901 hemopoietic system. *Genes & Development* **11**, 774–785. ISSN: 0890-
902 9369 (Mar. 15, 1997).
- 903 14. Pina, C., Fugazza, C., Tipping, A. J., Brown, J., Soneji, S., Teles,
904 J., Peterson, C. & Enver, T. Inferring rules of lineage commitment
905 in haematopoiesis. *Nature Cell Biology* **14**, 287–294. ISSN: 1465-7392,
906 1476-4679 (Mar. 2012).
- 907 15. Mojtahedi, M., Skupin, A., Zhou, J., Castaño, I. G., Leong-Quong,
908 R. Y. Y., Chang, H., Trachana, K., Giuliani, A. & Huang, S. Cell Fate
909 Decision as High-Dimensional Critical State Transition. en. *PLOS Biol-*
910 *ogy* **14**. Number: 12 Reporter: PLOS Biology, e2000640. ISSN: 1545-7885
911 (Dec. 2016).
- 912 16. Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin,
913 E., Guillemin, A., Papili Gao, N., Gunawan, R., Cosette, J., Arnaud,
914 O., Kupiec, J.-J., Espinasse, T., Gonin-Giraud, S. & Gandrillon, O.
915 Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecu-
916 lar Variability Preceding Irreversible Commitment in a Differentiation
917 Process. en. *PLOS Biology* **14** (ed Teichmann, S. A.) e1002585. ISSN:
918 1545-7885 (Dec. 2016).

- 919 17. Moussy, A., Cosette, J., Parmentier, R., da Silva, C., Corre, G., Richard,
920 A., Gandrillon, O., Stockholm, D. & Páldi, A. Integrated time-lapse
921 and single-cell transcription studies highlight the variable and dynamic
922 nature of human hematopoietic cell fate commitment. en. *PLOS Biology*
923 **15** (ed Huang, S.) e2001867. ISSN: 1545-7885 (July 2017).
- 924 18. Gao, M., Ling, M., Tang, X., Wang, S., Xiao, X., Qiao, Y., Yang, W.
925 & Yu, R. *Comparison of High-Throughput Single-Cell RNA Sequencing*
926 *Data Processing Pipelines* en. preprint (Feb. 2020).
- 927 19. Moris, N., Edri, S., Seyres, D., Kulkarni, R., Domingues, A. F., Balayo,
928 T., Frontini, M. & Pina, C. Histone Acetyltransferase KAT2A Stabilizes
929 Pluripotency with Control of Transcriptional Heterogeneity. *Stem Cells*
930 **36**, 1828–1838. ISSN: 1066-5099 (Print) 1066-5099 (2018).
- 931 20. Guillemin, A., Duchesne, R., Crauste, F., Gonin-Giraud, S. & Gan-
932 drillon, O. Drugs modulating stochastic gene expression affect the ery-
933 throid differentiation process. *PLOS ONE* **14**, e0225166 (2019).
- 934 21. Stumpf, P. S., Smith, R. C. G., Lenz, M., Schuppert, A., Müller, F.-J.,
935 Babtie, A., Chan, T. E., Stumpf, M. P., Please, C. P., Howison, S. D.,
936 Arai, F. & MacArthur, B. D. Stem Cell Differentiation as a Non-Markov
937 Stochastic Process. *Cell Systems* **5**, 268–282 (2017).
- 938 22. Dussiau, C., Boussaroque, A., Gaillard, M., Bravetti, C., Zaroili, L.,
939 Knosp, C., Friedrich, C., Asquier, P., Willems, L., Quint, L., Bouscary,
940 D., Fontenay, M., Espinasse, T., Plesa, A., Sujobert, P., Gandrillon,
941 O. & Kosmider, O. Hematopoietic differentiation is characterized by a
942 transient peak of entropy at a single-cell level. *BMC Biology* **20**, 60.
943 ISSN: 1741-7007 (2022).
- 944 23. Toh, K., Saunders, D., Verd, B. & Steventon, B. Zebrafish Neuromeso-
945 dermal Progenitors Undergo a Critical State Transition in vivo. *bioRxiv*,
946 2022.02.25.481986 (2022).
- 947 24. Dalton, S. Linking the Cell Cycle to Cell Fate Decisions. *Trends Cell*
948 *Biol* **25**, 592–600. ISSN: 1879-3088 (Electronic) 0962-8924 (Linking)
949 (2015).
- 950 25. Huang, S. Reprogramming cell fates: reconciling rarity with robustness.
951 en. *BioEssays* **31**, 546–560. ISSN: 02659247, 15211878 (May 2009).

- 952 26. Parmentier, R., Moussy, A., Chantalat, S., Racine, L., Sudharshan, R.,
953 Papili Gao, N., Stockholm, D., Corre, G., Fourel, G., Deleuze, J., Gu-
954 nawan, R. & Paldi, A. Global genome decompaction leads to stochastic
955 activation of gene expression as a first step toward fate commitment in
956 human hematopoietic stem cells. *bioRxiv* (2021).
- 957 27. Gandrillon, O., Schmidt, U., Beug, H. & Samarut, J. TGF-Beta cooper-
958 ates with TGF-Alpha to induce the self-renewal of normal erythrocytic
959 progenitors: evidence for an autocrine mechanism. *The EMBO Journal*
960 **18**, 2764–2781. ISSN: 0261-4189, 1460-2075 (May 17, 1999).
- 961 28. Gandrillon, O. & Samarut, J. Role of the different RAR isoforms in
962 controlling the erythrocytic differentiation sequence. Interference with
963 the v-erbA and p135gag-myb-ets nuclear oncogenes. *Oncogene* **16**, 563–
964 74 (1998).
- 965 29. Richard, A., Vallin, E., Romestaing, C., Roussel, D., Gandrillon, O. &
966 Gonin-Giraud, S. Erythroid differentiation displays a peak of energy
967 consumption concomitant with glycolytic metabolism rearrangements.
968 *PLoS One* **14**, e0221472. ISSN: 1932-6203 (Electronic) 1932-6203 (Link-
969 ing) (2019).
- 970 30. Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F.,
971 Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A. & Amit, I. Mas-
972 sively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of
973 Tissues into Cell Types. *Science* **343**. Number: 6172 Reporter: Science,
974 776–779. ISSN: 0036-8075, 1095-9203 (Feb. 14, 2014).
- 975 31. Woodworth, M. B., Girsakis, K. M. & Walsh, C. A. Building a lineage
976 from single cells: genetic techniques for cell lineage tracking. *Nature*
977 *Reviews Genetics* **18**, 230–244. ISSN: 1471-0056, 1471-0064 (Apr. 2017).
- 978 32. Zreika, S., Fourneaux, C., Vallin, E., Modolo, L., Seraphin, R., Moussy,
979 A., Ventre, E., Bouvier, M., Ozier-Lafontaine, A., Bonnaffoux, A., Pi-
980 card, F., Gandrillon, O. & Gonin-Giraud, S. Evidence for close molecu-
981 lar proximity between reverting and undifferentiated cells. *BMC Biology*
982 **20**, 155. ISSN: 1741-7007 (Dec. 2022).
- 983 33. Terrén, I., Orrantia, A., Vitallé, J., Zenarruzabeitia, O. & Borrego, F.
984 in *Methods in Enzymology* 239–255 (Elsevier, 2020). ISBN: 978-0-12-
985 818673-2.

- 986 34. Parish, C. R. Fluorescent dyes for lymphocyte migration and prolifera-
987 tion studies. *Immunology and Cell Biology* **77**, 499–508. ISSN: 08189641
988 (Dec. 1999).
- 989 35. Kim, W., Klarmann, K. D. & Keller, J. R. Gfi-1 regulates the ery-
990 throid transcription factor network through Id2 repression in murine
991 hematopoietic progenitor cells. *Blood* **124**, 1586–1596 (Sept. 4, 2014).
- 992 36. Da Cunha, A. F., Brugnerotto, A. F., Duarte, A. d. S. S., Lanaro, C.,
993 Costa, G. G. L., Saad, S. T. O. & Costa, F. F. Global gene expression
994 reveals a set of new genes involved in the modification of cells dur-
995 ing erythroid differentiation: Modification of cells during erythroid dif-
996 ferentiation. *Cell Proliferation* **43**, 297–309. ISSN: 09607722, 13652184
997 (Apr. 28, 2010).
- 998 37. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. *On the Surprising Behav-*
999 *ior of Distance Metrics in High Dimensional Space in Database Theory*
1000 *— ICDT 2001* (eds Van den Bussche, J. & Vianu, V.) (Springer Berlin
1001 Heidelberg, Berlin, Heidelberg, 2001), 420–434. ISBN: 978-3-540-44503-
1002 6.
- 1003 38. Bonnaïffoux, A., Herbach, U., Richard, A., Guillemin, A., Gonin-Giraud,
1004 S., Gros, P.-A. & Gandrillon, O. WASABI: a dynamic iterative frame-
1005 work for gene regulatory network inference. en. *BMC Bioinformatics*
1006 **20**, 220. ISSN: 1471-2105 (Dec. 2019).
- 1007 39. Suter, D. M., Molina, N., Gatfield, D., Schneider, K., Schibler, U. &
1008 Naef, F. Mammalian genes are transcribed with widely different burst-
1009 ing kinetics. *Science* **332**, 472–4. ISSN: 1095-9203 (Electronic) 0036-8075
1010 (Linking) (2011).
- 1011 40. Tunnacliffe, E. & Chubb, J. R. What Is a Transcriptional Burst? *Trends*
1012 *Genet* **36**, 288–297. ISSN: 0168-9525 (Print) 0168-9525 (Linking) (2020).
- 1013 41. Rodriguez, J. & Larson, D. R. Transcription in Living Cells: Molecular
1014 Mechanisms of Bursting. *Annu Rev Biochem* **89**, 189–212. ISSN: 1545-
1015 4509 (Electronic) 0066-4154 (Linking) (2020).
- 1016 42. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene net-
1017 works. *Science* **307**, 1965–9 (2005).
- 1018 43. Kim, S. & Shendure, J. Mechanisms of Interplay between Transcription
1019 Factors and the 3D Genome. *Mol Cell* **76**, 306–319. ISSN: 1097-4164
1020 (Electronic) 1097-2765 (Linking) (2019).

- 1021 44. Martin-Martin, N., Carracedo, A. & Torrano, V. Metabolism and Tran-
1022 scription in Cancer: Merging Two Classic Tales. *Front Cell Dev Biol* **5**,
1023 119. ISSN: 2296-634X (Print) 2296-634X (Linking) (2017).
- 1024 45. Wang, F. & Higgins, J. M. Histone modifications and mitosis: counter-
1025 marks, landmarks, and bookmarks. *Trends Cell Biol* **23**, 175–84. ISSN:
1026 1879-3088 (Electronic) 0962-8924 (Linking) (2013).
- 1027 46. Golloshi, R., Sanders, J. T. & McCord, R. P. Genome organization
1028 during the cell cycle: unity in division. *Wiley Interdiscip Rev Syst Biol*
1029 *Med* **9**. ISSN: 1939-005X (Electronic) 1939-005X (Linking) (2017).
- 1030 47. Palozola, K. C., Donahue, G. & Zaret, K. S. EU-RNA-seq for in vivo
1031 labeling and high throughput sequencing of nascent transcripts. *STAR*
1032 *Protoc* **2**, 100651. ISSN: 2666-1667 (Electronic) 2666-1667 (Linking)
1033 (2021).
- 1034 48. Kadauke, S., Udugama, M. I., Pawlicki, J. M., Achtman, J. C., Jain,
1035 D. P., Cheng, Y., Hardison, R. C. & Blobel, G. A. Tissue-specific mitotic
1036 bookmarking by hematopoietic transcription factor GATA1. *Cell* **150**,
1037 725–37. ISSN: 1097-4172 (Electronic) 0092-8674 (Linking) (2012).
- 1038 49. Scrucca, L., Fop, M., Murphy, T. B. & Raftery, A. E. mclust 5: cluster-
1039 ing, classification and density estimation using Gaussian finite mixture
1040 models. *The R Journal* **8**, 289–317 (2016).
- 1041 50. Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E.
1042 & Notredame, C. Nextflow enables reproducible computational work-
1043 flows. *Nature Biotechnology* **35**, 316–319. ISSN: 1087-0156, 1546-1696
1044 (Apr. 2017).
- 1045 51. R Core Team. *R: A Language and Environment for Statistical Comput-*
1046 *ing* R Foundation for Statistical Computing (Vienna, Austria, 2021).
- 1047 52. Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom,
1048 E., Dudoit, S. & Yosef, N. Performance Assessment and Selection of
1049 Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems* **8**.
1050 Number: 4 Reporter: Cell Systems, 315–328.e8. ISSN: 24054712 (Apr.
1051 2019).
- 1052 53. Breda, J., Zavolan, M. & van Nimwegen, E. Bayesian inference of gene
1053 expression states from single-cell RNA-seq data. *Nature Biotechnology*
1054 **39**, 1008–1016. ISSN: 1087-0156, 1546-1696 (Aug. 2021).

- 1055 54. Hafemeister, C. & Satija, R. Normalization and variance stabilization of
1056 single-cell RNA-seq data using regularized negative binomial regression.
1057 *bioRxiv* (Mar. 18, 2019).
- 1058 55. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng,
1059 L. G., Ginhoux, F. & Newell, E. W. Dimensionality reduction for visual-
1060 izing single-cell data using UMAP. *Nature Biotechnology* **37**. Number:
1061 1 Reporter: Nature Biotechnology, 38–44. ISSN: 1087-0156, 1546-1696
1062 (Jan. 2019).

Chapitre 3

Isolement de cellules apparentées par puce microfluidique pour analyse transcriptomique

3.1 Introduction

Dans le cadre d'une ANR portée par notre équipe et en collaboration avec le laboratoire d'Andrew DeMellow à l'ETH Zurich, nous avons cherché à développer une puce microfluidique permettant également d'étudier la mémoire transcriptionnelle au cours de plusieurs générations cellulaires.

La puce envisagée devait permettre d'isoler des cellules uniques dans des chambres individuelles (une cellule par chambre), les laisser se diviser une première fois, re-localiser les cellules soeurs issues de cette mitose dans de nouvelles chambres individuelles et les laisser se diviser à nouveau. L'idéal étant de pouvoir répéter l'opération sur un plus grand nombre de divisions cellulaires. Durant toutes ces étapes, la prolifération des cellules devait être suivie en microscopie time-lapse. La suite du protocole consistait à extraire les cellules issues des différentes divisions afin d'analyser leur transcriptome par sc-RNA-seq, et reconstruire l'arbre généalogique précis et résolutif à l'échelle de la division cellulaire grâce aux données de microscopie.

La preuve de concept de ce dispositif à l'échelle de la première division cellulaire fait l'objet d'un article en cours de rédaction.

3.2 Optimisations

Ce projet exploratoire a été mené en collaboration avec Kamil Aslan, étudiant en thèse dans l'équipe d'Andrew DeMello, qui a travaillé sur le développement d'une puce microfluidique dédiée à ce projet. Ensemble, en fonction des difficultés rencontrées, nous avons apporté différentes améliorations à cette puce. En effet, nous avons effectué conjointement, parfois à distance de part les circonstances particulières de ces 2 dernières années, un grand nombre d'optimisations relatives au design de la puce et aux différentes étapes de son utilisation que je vais rapidement détailler dans la figure ci-dessous (Figure 3.1).

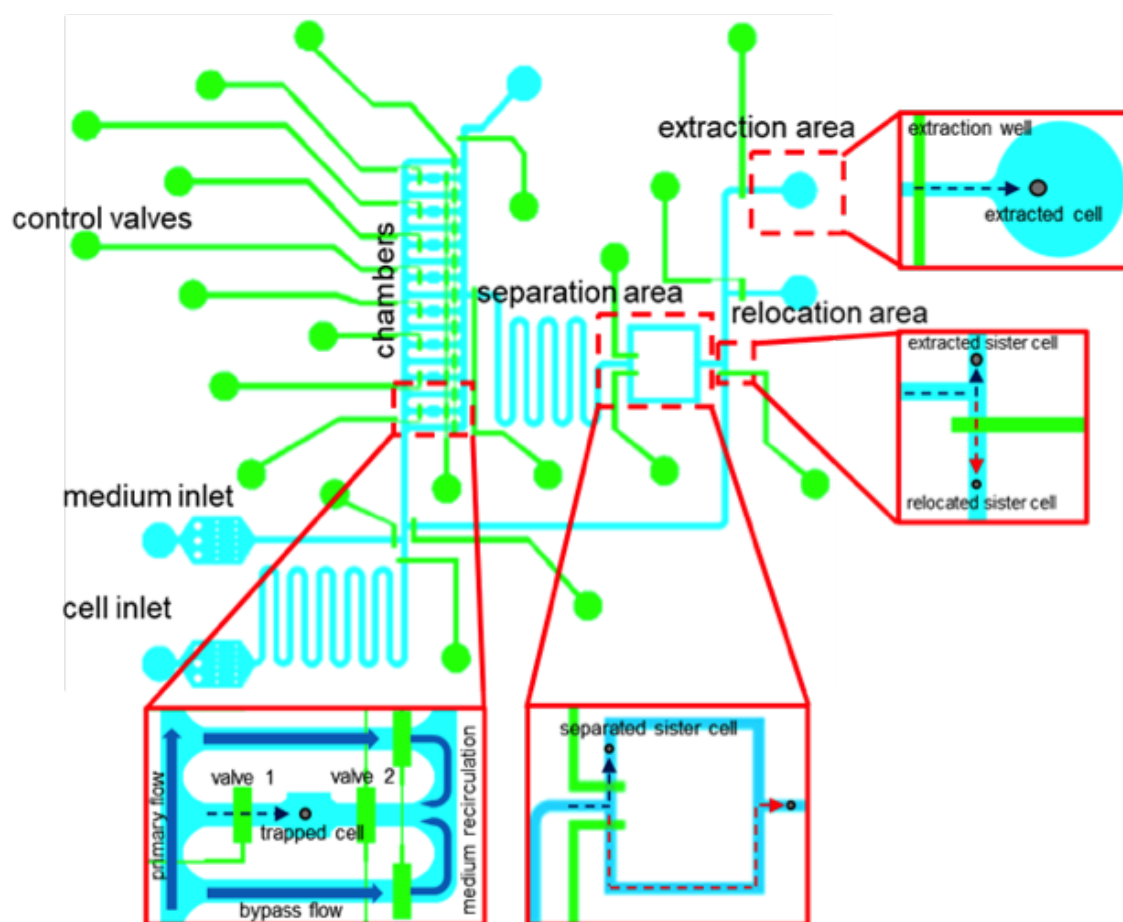


FIGURE 3.1 – Schéma de la puce microfluidique.

Le dispositif se compose de 8 chambres de prolifération individuelle, d'une jonction contrôlée par une valve pour la séparation des cellules sœurs après la division et d'un canal de retour qui permet de replacer les cellules sœurs après leur division dans deux nouvelles chambres individuelles.

Premièrement, les cellules sont capturées dans les chambres individuelles de la puce (Figure 3.1 premier carré rouge à gauche « chambers »). Nous avons optimisé le design de la zone d'entrée des cellules en créant des pièges afin de limiter le passage de débris fibreux (provenant du milieu)

et de bulles d'air dans la puce (Figure 3.1 « cell inlet »).

Deuxièmement, nous avons travaillé sur les conditions de culture des cellules uniques dans les chambres de la puce pour permettre leur division. La première difficulté majeure a consisté à maintenir le taux de CO₂ à 5% dans l'enceinte du microscope, nous avons pour cela travaillé avec du milieu saturé en CO₂. La deuxième difficulté rencontrée était des contaminations fongiques et bactériennes dans les valves et dans la puce en général. Nous avons donc mis en place un protocole strict de nettoyage à l'acide acétique et stérilisation aux UV de tous les éléments de la puce ainsi que l'enceinte du microscope afin de limiter au maximum les contaminations. Kamil a également créé des voies de re-circulation de milieu afin d'assurer un apport constant en nutriments aux cellules pendant toute la durée de leur culture permettant ainsi d'obtenir des taux de division comparables à ceux de cellules maintenues dans des conditions classiques de culture (Figure 3.1 « bypass flow »).

Troisièmement, nous avons optimisé la séparation des cellules soeurs post-mitose. En effet, comme expliqué dans le chapitre précédent, les cellules soeurs ont tendance à rester attachées l'une à l'autre même bien après la cytokinèse. Ainsi, post-mitose nous obtenions des doublets de cellules qu'il fallait donc dissocier. Pour cela, nous avons testé différentes approches. Dans un premier temps une séparation mécanique en générant des courants de liquide avec les valves contrôlant la puce. Cependant les liaisons entre les cellules étaient trop fortes. Dans un deuxième temps nous avons tenté une séparation à l'aide d'ultrasons. Toutes les optimisations de la séparation des cellules à l'aide d'ultrasons ont été réalisées en France, notamment par Aïcha Mahavory étudiante en Master1 que j'ai encadré pendant 7 semaines; nous avons fait particulièrement attention à réaliser ces optimisations dans des conditions les plus proches de celles de la puce qu'il nous a été possible de reproduire dans notre laboratoire. Cependant, le transfert de ces optimisations dans les vraies conditions expérimentales a été infructueux. Nous nous sommes finalement orientés sur une séparation enzymatique des cellules soeurs à l'aide d'Accutase. L'Accutase permet une dissociation douce des cellules et à l'avantage de s'auto-inactiver à 37°C après 30-45min. Pour dissocier les doublets de cellules, nous faisons circuler de l'Accutase pendant 30min dans les chambres des cellules ce qui nous a permis de séparer et d'extraire les deux cellules soeurs indépendamment.

Enfin, nous avons travaillé sur la quatrième et dernière étape qui consiste en l'extraction des cellules de la puce. Afin de contrôler au maximum le volume de milieu extrait avec les cellules,

conditionnant la réussite ultérieure du scRNA-seq, nous avons d'abord utilisé des petits tubes de longueur et de diamètre connus. Il est d'usage en microfluidique pour extraire un objet, d'extraire 3 volumes morts. Ainsi si l'objet a une vitesse plus importante que le liquide il passera dans le premier volume mort, si il a la même vitesse dans le deuxième et si il a une vitesse moins importante dans le troisième volume mort. Cependant, même en réduisant la taille et le diamètre des tubes d'extraction, les volumes minimaux d'extraction (correspondant donc à 3 volumes morts) étaient encore trop importants, et nos tests ont montré que la méthode dans son ensemble était peu reproductible. Nous avons finalement tiré profit de notre expérience d'utilisation de capillaires en verre (optimisations décrites dans le chapitre 2). Pour cela, les cellules ont été dirigées dans un puits d'extraction dans lequel on a inséré un capillaire en verre gradué afin de surveiller le volume ; par capillarité la cellule est aspirée dans le capillaire et est ensuite transférée dans un tube contenant du tampon de lyse et les amorces de sc-RNA-seq. Cette approche a permis de réduire considérablement le volume de récupération des cellules.

Pour réaliser toutes ces optimisations, nous avons utilisé soit des progéniteurs érythrocytaires aviaires primaires (T2EC) soit des cellules issues d'une lignée appelées 6C2 (progéniteurs aviaires transformés par l'oncogène *verb-A* et exprimant constitutivement la protéine fluorescente m-Cherry) dont la viabilité est plus facile à préserver et dont la fluorescence nous a permis de contrôler l'extraction efficace des cellules.

3.3 Résumé des résultats

Toutes nos optimisations nous a permis de réaliser une expérience pilote avec une puce qui permet : 1) d'isoler des cellules mères 6C2 dans des chambres individuelles ; 2) la division de chaque cellule mère en deux cellules filles ; 3) l'extraction des deux cellules filles à l'aide d'un capillaire en verre, validée grâce à l'observation du signal fluorescent m-Cherry des cellules ; 4) l'analyse de leur transcriptome par sc-RNA-seq.

Nous avons ainsi obtenu et analysé avec succès le transcriptome de 2 couples de cellules soeurs. Nous avons comparé la qualité des données obtenues, notamment le nombre de gènes cellulaires détectés, sur des cellules soit isolées *via* la puce microfluidique soit isolées au FACS (méthode de référence) et les résultats obtenus sont très satisfaisants. En effet, les cellules isolées par microfluidique ont la même quantité de gènes détectés en moyenne que les cellules triées au FACS, et la réduction de dimension ne montre pas de séparation entre les cellules isolées avec les deux

méthodes différentes.

3.4 Principales conclusions

Notre méthode utilisant une puce microfluidique permet la capture de cellules non-adhérentes et leur isolement dans une chambre de culture, leur division, qui peut être suivie par microscopie time-lapse et enfin l'extraction des cellules pour réaliser des analyses de sc-RNA-seq tout en conservant les informations généalogiques entre elles. Cette approche microfluidique permet d'obtenir des données de sc-RNA-seq de la même qualité que celles obtenues lorsque les cellules sont triées au FACS.

3.5 Publication - article 2

1 **An Image-Guided Microfluidic System for Single Cell Lineage Tracking**

2 Aslan Kamil Mahmut¹, Fourneaux Camille², Yilmaz Alperen³, Parmentier Romuald⁴, Gonin-Giraud
3 Sandrine², Paldi Andras⁴, DeMello J Andrew¹, Gandrillon Olivier^{2,5}

4 ¹ Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences,
5 ETH Zürich, Wolfgang-Pauli-Strasse 10, CH-8093 Zürich

6 ² ENS de Lyon, Université Claude Bernard, CNRS UMR 5239, INSERM U1210, Laboratory of
7 Biology and Modelling of the Cell, Université de Lyon, 46 allée d'Italie Site Jacques Monod, F-
8 69007, Lyon, France.

9 ³ Faculty of Medicine, Koç University, 34450 Istanbul, Turkey.

10 ⁴ Ecole Pratique des Hautes Etudes, PSL Research University, St-Antoine Research Center,
11 Inserm U938, 34 rue Crozatier, 75012, Paris, France

12 ⁵ Inria Team Dracula, Inria Center Grenoble Rhône-Alpes, Montbonnot-Saint-Martin, France.

13 **Abstract:** Cell lineage tracking is a long-standing, open problem in biology. Microfluidic
14 technologies have the potential to address this problem, due to their ability to manipulate
15 single cells in a rapid and efficient manner. When coupled with traditional imaging approaches,
16 microfluidics systems offer the possibility to follow single-cell division over time. In this work,
17 we present a valve-based microfluidic system able to probe the decision-making process of
18 single cells, by tracking their lineage over generations. The system is able to trap single-cells
19 inside growth chambers, isolate sister-cells after one to a few divisions and extract them for
20 downstream transcriptome analysis. The setup incorporates cell manipulation steps, image
21 processing-based automation for cell loading and cell-growth monitoring, reagent addition and
22 device washing. 6C2 (chicken erythro leukemia cell line) and T2EC (primary chicken erythrocytic
23 progenitors) cells were tracked inside the microfluidic device over two generations, with a cell
24 viability rate in excess of 90% being achieved. Sister-cells were successfully isolated after
25 division and extracted from the device in a 500 nl volume, which is compatible with
26 downstream single-cell RNA sequencing analysis.

27 **Introduction**

28 One of the biggest challenges in quantitative biology is to understand the decision-making
29 processes of cells. Over the past 20 years, a change in the scale of investigation from cell
30 populations to single-cell level has already brought numerous insight on such processes ¹⁻³.
31 Indeed, the main power of the single-cell studies is to reveal the underlying transcriptional
32 heterogeneity of normal and pathological cells ^{4,5}. Furthermore, single-cell resolution studies
33 have provided evidences that gene expression variability also powers how cells make decisions
34 ^{3,6}.

35 Cellular differentiation is the process by which any pre-committed cell acquires its identity and
36 can be viewed as a dynamic process wired by the underlying gene regulatory network (GRN).
37 Cells can be visualized as "moving particles" in a landscape, the cell state space shaped by the
38 GRN state ⁷. In this space, steady states can be represented by attraction wells. In this view, cells
39 can escape their self-renewing steady state through a rise in gene expression variability and
40 then explore freely, to some extent, the landscape to finally reach a new attractor state, the
41 differentiated state ⁷. Accordingly, single-cell analysis of *in vitro* and *in vivo* differentiation
42 models have confirmed that this cellular process is indeed characterized by a global rise in gene
43 expression variability ⁸⁻¹².

44 However, the establishment of the gene expression variability across cell generations is still
45 poorly understood. Such a fundamental question might be of critical importance as it seems to
46 be a conserved phenomenon across biological systems and across species ¹³⁻¹⁶. Indeed, at the
47 organism scale, during differentiation, the cells must maintain their identity through mitosis in
48 order to eventually reach their differentiation state. Based on recent studies, the support for
49 this state memory is the inheritance of mRNA level from mother cells to daughter cells. This

50 transmission is, with high probability, supported by the inheritance of epigenetic modifications
51 allowing the maintenance of gene-specific transcription levels over cell divisions ^{16,17}. In recent
52 studies, it has been described that some genes which expression is variable amongst an isogenic
53 cell population have their expression correlated between genealogically related cells. For some
54 of those “memory genes” the correlation in expression may last for tenth of cell generations.
55 These data, gathered on self-renewing cells, imply a gene-specific transcriptional memory over
56 several cell generations ¹³.

57 In order to investigate how cells reconcile the constraints of transcriptional memory and the rise
58 in gene expression variability during the process of differentiation, one must recover single-cells
59 and their lineage information over a few cell divisions.

60 While some methods are available to perform cell-tracking over multiple cell generations
61 coupled with transcriptomics analysis, they require heavy genetic modifications and do not offer
62 the resolution of one cell division ^{18–20}.

63 As an alternative method, microfluidic tools are adept at performing single-cell manipulations
64 ²¹. Indeed, microfluidics has already offer high quality, easy access and automatized platforms to
65 study gene expression at the single-cell level ^{22,23}. Furthermore, microfluidic systems are well-
66 suited to controlling and varying environmental conditions in a precise manner, and since they
67 can be easily integrated with sensitive optical detection systems and imaging modalities, time-
68 course experiments for long-term tracking are also possible ²⁴. At a basic level, microfluidic cell
69 culture systems have many advantages over conventional cell culture methods, including low
70 reagent consumption, multiplexing capabilities and easy automation of the cell culture tasks ²⁵.
71 Accordingly, monitoring single-cell lineages and analyzing the differences between sister cells

72 upon division becomes possible without the need to perform genetic modifications of the
73 mother cells.

74 Recently, several microfluidic-based cell culture systems have been developed for tracking cell
75 lineage. For example, Kimmerling *et al.* developed a parallel micron-sized trap structure for
76 tracking the lineage of murine CD8⁺ T-cells and lymphocytic leukemia cell lines¹⁴. In this study,
77 cells trapped in individual hydrodynamic traps were grown in a serpentine-shaped parallel
78 microchannel network. Then, divided sister cells were separated from each other using a fluid
79 flow through the traps and extracted from the device outlet. This device allows multi-
80 generations cell lineage tracking. However, fluid flow conditions and the hydrodynamic traps
81 were optimized for only specific cell sizes; therefore, the device must be repeatedly redesigned
82 to assay different cell types. Another drawback of this system is that it is not possible to address
83 the divided cells independently; thus, extracting specific sister cells from the device is
84 challenging. There are other microfluidic approaches that can track and extract specific cells
85 from culture²⁶, but the extraction volume is often too large (a few microliters) to be compatible
86 with downstream analysis. Other methods, such as the Polaris system²⁷, allow single-cell
87 transcriptomics measurements, but it is impossible to track the cell lineage over multiple
88 generations. Therefore, there is still a pressing need for an automated microfluidic-based
89 platform that can combine cell lineage tracking and single-cell extraction at low volumes of less
90 than 1 μ L. To this end, we report the design, fabrication and development of an automated
91 image-based microfluidics platform for tracking non-adherent single-cell lineage. Essential
92 characteristics of the system include: (i) the integration of a microfluidic chamber array for
93 single-cell trapping, (ii) the ability to monitor of cell growth, (iii) the ability to separate sister

- 94 cells after division, (iv) easy reallocation of sister cells to allow monitoring of second and third
- 95 division events and (v) extraction of the cells for downstream analysis.

96 **Materials and Methods**

97 **Microscope Incubator Setup**

98 To monitor *in vitro* cell proliferation, environmental conditions must be mimicked under a
99 microscope placed inside an incubator. Therefore, an inverted microscope (Eclipse Ti-E, Nikon)
100 was enclosed in a custom-designed Makrolon incubation box (Life Imaging Services) to provide
101 the optimum proliferation conditions for the cells (5% CO₂, 95% humidity, 37°C). The box was
102 connected to an air-heating device with precise temperature control (Life Imaging Services). An
103 in-house CO₂ chamber connected to a 5% CO₂ mixture tank (PanGas) with electronic flow control
104 (Red-y, Vögtlin Instruments GmbH) was attached to the microfluidic device, which was placed
105 on a motorized xy translation stage (Mad City Labs GmbH). An optical shutter controlled by an
106 automation system (ProScan III, Prior Scientific Instruments GmbH) was used to regulate the
107 light exposure. A scientific complementary metal-oxide-semiconductor (sCMOS) camera (pco
108 edge, PCO GmbH) in conjunction with a 10X/0.3 NA objective (Plan Fluor, Nikon) was used to
109 image the cells for periods of between 24 and 48 hours. A pressure-based flow controller (Flow
110 EZ™, Fluigent) was used to drive the cells and the reagents into the microfluidic device. Solenoid
111 valves (MH1, Festo AG) were incorporated into the microfluidic device to manipulate the cells
112 and automate the whole experimental process via a custom-developed MATLAB® code.

113

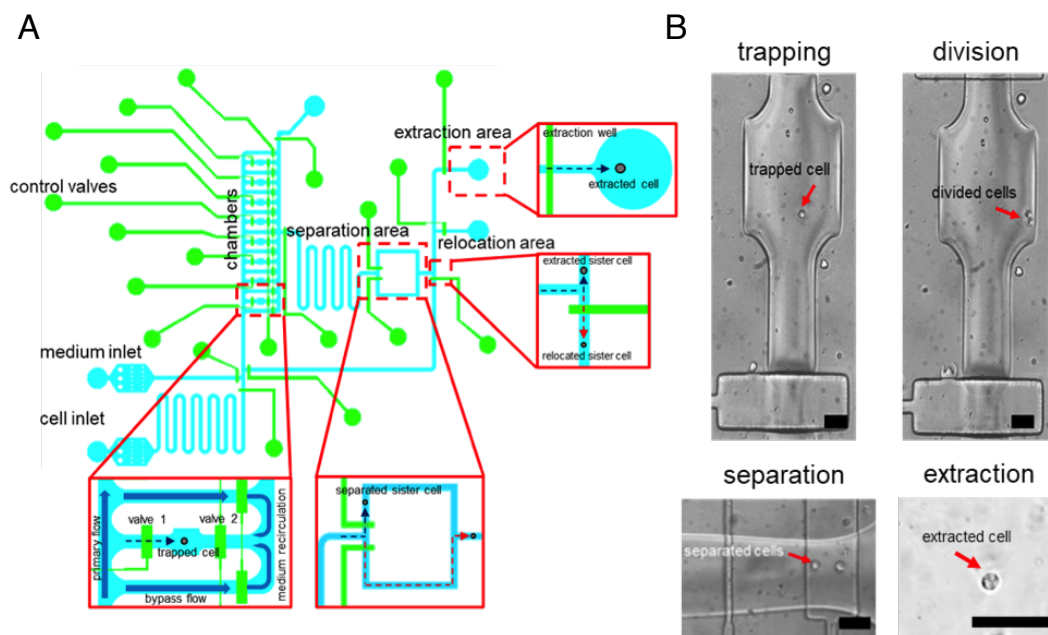
114 **Microfluidic Device Design and Operation**

115 A two-layer microfluidic device was designed to trap and proliferate cells in a controlled
116 manner. The core elements controlling fluid flow within the microfluidic device were
117 polydimethylsiloxane (PDMS) pneumatic microvalves similar to those introduced in 2000 by
118 Quake and co-workers²⁸. The microfluidic device consists of a control layer and a fluidic layer,

119 each consisting of a group of channels. The control layer is located above or underneath the
120 fluidic layer and can be deformed, with applied pressure resulting in the blockage of the fluidic
121 channel. Such "Quake valves" can be designed in a "push-up" or "push-down" type depending
122 on the location of the control layer with respect to the fluidic layer. Push-up valves are more
123 desirable for applications involving eukaryotic cell manipulation in deeper fluidic channels,
124 whereas push-down valves are more suitable, for example when molecules are patterned on a
125 glass slide ²⁹. In the current case, we used a push-up valve structure, since the device was
126 exclusively intended for culturing and manipulating eukaryotic cells.

127 A two-layer microfluidic device with eight chambers was designed and fabricated for long-term
128 monitoring and tracking of sister stem cells over two generations (Figure 1.A). To conduct a
129 single cell proliferation experiment, single cells located inside the trapping chambers must be in
130 contact with the medium (Figure 1.A). Single cells can be trapped inside these chambers via the
131 aid of control valve 1 (left valve on the trapping chamber), which, upon actuation, blocks fluid
132 flow entering the trapping region (Figure 1.A bottom left inset). In such a situation, providing
133 medium to the cells can be accomplished through the combination of bypass flow channels
134 added to both sides of each chamber and the opening of control valve 2 (the right valve in the
135 trapping chamber). This whole process starts with a primary medium flow, firstly divided into
136 many bypass flow paths. The fluid flow in these bypass channels is controlled using control
137 valves, which were kept open to maintain a constant circulation of fresh medium around the
138 cell trapping chambers. With control valve 2 being in an open state, medium, which is
139 constantly circulated in the bypass channels, can diffuse to the trapping chambers (Figure 1.A
140 bottom left inset). Separation, relocation and extraction areas were also integrated into the
141 device as shown in the middle and right part of Figure 1.A. More specifically, after division,

142 sister cells flow into the separation area that incorporates two control valves. Actuating one of
143 these valves will allow one of the cells to flow towards the extraction area, while the other cell
144 will remain trapped (Figure 1.A bottom right inset). The feedback channel allows relocation of
145 the sister cells after division from the separation area to the cell trapping chambers (Figure 1
146 top right inset). Sister cells separated after division flow through the feedback channel upon
147 actuation of the control valves (Figure 1 top right inset) and were subsequently placed in the
148 trapping chambers. Then, after division (second generation), they can be separated and either
149 extracted from the device or relocated back to the chambers for tracking the third generation.
150 The extraction area includes eight independently addressable, 1 mm diameter and 3 mm depth
151 open wells for the collection of sister cells (Figure 1.A right part). The device had eight chambers
152 allowing monitoring up to three generations of a single cell (from one parent cell to eight
153 daughter cells). Images in Figure 1.B depict the entire workflow of the single cell proliferation
154 process that takes place in the second-generation device, which comprises trapping of a single
155 cell inside a growth chamber, cell growth and division, separation of the sister cells after division
156 and extraction of the individual sister cells for downstream transcriptome analysis.



157

158 **Figure 1: Schematic and workflow of the microfluidic single-cell processing system.** (A) The
 159 device consists of 8 individually addressable proliferation chambers, a valve-based junction for
 160 separation of the sister cells after division and a feedback channel that allows relocation of the
 161 sister cells after division from the separation area to the cell trapping chambers. (B) The process
 162 involves parallel chambers where cells are trapped and proliferate overnight. Then, through use
 163 of the valve system, sister cells are separated from each other and then extracted from the
 164 device for downstream analysis. The scale bars are 50 μm .

165

166 Cell culture

167 6C2-C11 cells are chicken erythroblasts cell line transformed by the avian erythroblastosis virus
 168 (AEV) carrying a stably integrated mCHERRY transgene and were maintained in α Minimal
 169 Essential Medium (Gibco) complemented with 10% Fetal Bovine Serum, 1% Normal Chicken
 170 Serum³⁰, 1% penicillin and streptomycin 10000U/ml (Gibco), 0,1mM Beta-mercaptoethanol
 171 (Sigma), and kept at 37°C with 5% CO₂ in an incubator.

172 T2EC cells were extracted from the bone marrow of white leghorn chicken embryos (INRA,
 173 Tours, France)³¹. The cells were cultured in α MEM medium (Sigma-Aldrich), supplemented with
 174 1 mM HEPES, 10% (v/v) fetal bovine serum (FBS, Life Technologies), 1% (v/v) Penicillin-

175 Streptomycin (10,000 U/mL, Life Technologies), 100 nM β -mercaptoethanol, 1 mM
176 dexamethasone, 5 ng/mL transforming growth factor-alpha (TGF- α , Peprotech) and 1 ng/mL
177 transforming growth factor-beta (TGF- β , Peprotech) at 37°C, 5% CO₂.

178

179 **ScRNA-seq library preparation**

180 Single-cell RNA-sequencing was performed using an adapted version of MARS-seq protocol
181 (Massively parallel single-cell RNA sequencing)³² published here³³.

182

183 **Sequencing**

184 Library was sequenced on a Next500 sequencer (Illumina) with a custom paired-end protocol to
185 avoid a decrease of sequencing quality on read1 due to high number of T added during polyA
186 reading (130pb on read1 and 20pb on read2). We aimed for a depth of 200 000 raw reads per
187 cell.

188

189 **Data pre-processing**

190 Fastq files were pre-processed through a bio-informatics pipeline developed in our team on the
191 Nextflow platform³⁴ and published here³³. Briefly, the first step removed Illumina adaptors.
192 The second step de-multiplexed the sequences according to their plate barcodes. Then, all
193 sequences containing at least 4T following cell barcode and UMI were kept. Using UMItools
194 whitelist, the cell barcodes and UMI were extracted from the reads. The sequences were then
195 mapped on the reference transcriptome (Gallus GallusGRCG6A.95 from Ensembl) and UMI were
196 counted. Finally, a count matrix was generated.

197

198 **Quality control and data filtering**

199 All analysis were carried out using R software (4.0.5; ³⁵). Matrices from the two plates were
200 pooled together. Cells were filtered based on several criteria: reads number, genes number,
201 counts number and ERCC content. For each criteria the cut off values were determined based
202 on SCONE ³⁶ pipeline and were calculated as follows:

203 $\text{Mean}(\text{parameter}) - 3 * \text{sd}(\text{parameter})$

204 After filtering, remained 4 cells (2 couples), 3 orphan cells meaning cells which sister wasn't
205 present in the dataset, and 82 FACS sorted control cells. Based on Breda *et al.* work ³⁷, genes
206 were kept in the data set if in mean present in every cell.

207

208 **Normalization**

209 Filtered matrix was then normalized using SCTransform from Seurat package ³⁸ and was
210 corrected for sequencing depth.

211

212 **UMAP**

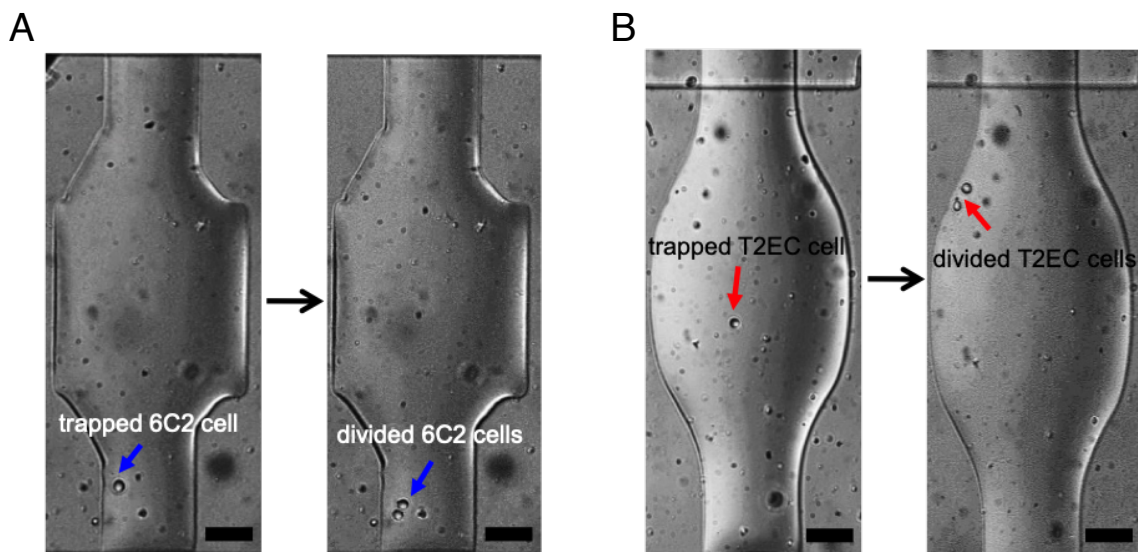
213 Dimensionality reduction and visualization was performed using UMAP ³⁹ default parameters.

214

215 **Results**

216 **Proliferation Experiments**

217 Significantly, the microfluidic device was used to process two models of non-adherent cells. 6C2
218 are a chicken erythro leukemia cell line and T2EC are primary chicken erythrocytic progenitors.
219 Cells at a concentration of 10^6 cells/ml were loaded in the device and single-cell trapping was
220 achieved in each chamber (Figure 2.A and B). The trapped mother cells were then monitored for
221 18 hours in mean, and cell divisions were observed for both 6C2 (Figure 2.A) and T2EC cells
222 (Figure 2.B), in independent experiments. The average cell proliferation rate (over 20 cells) was
223 defined as the rate of divided cells after 18 hours, as this time is the reference dividing time in
224 regular culture conditions. Accordingly, proliferation rates were measured as 93% and 86% for
225 6C2 and T2EC cells, respectively and matched proliferation times observed in conventional
226 culture conditions. In a second phase, sister cells were relocated after the first division in
227 individual chambers and we were able to observe a second cell division for both cell models
228 (around 36-40 hours of culture in the chip).



229

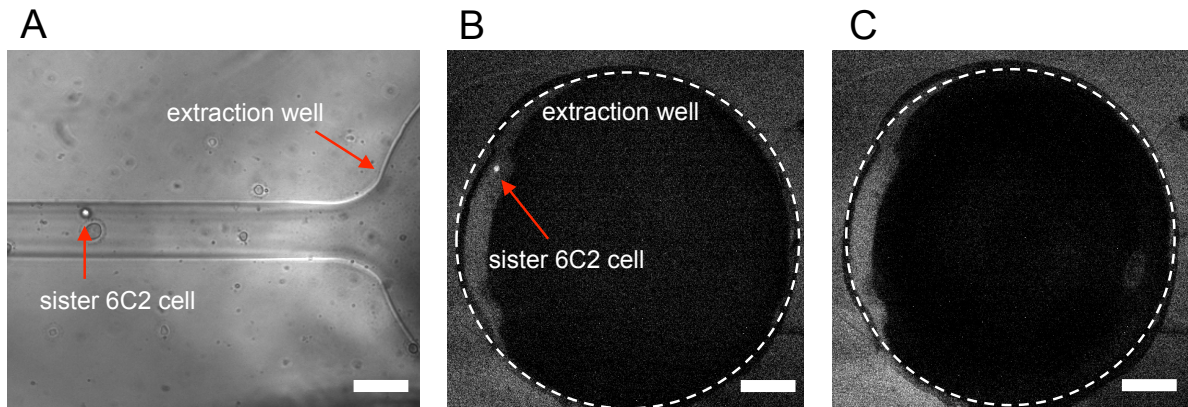
230 **Figure 2: Proliferation experiments.** Trapped 6C2 mother cell (A left panel) and T2EC mother
231 cell (B left panel) and 6C2 sister cells (A right panel) and T2EC sister cells (B right panel) after the
232 mother cell division in the proliferation chambers. The scale bar is 50 μm .
233

234 **Single Cell Extraction**

235 One of the most daunting tasks in the experimental process was the final collection of the
236 separated sister cells within a nanoliter volume (approximately 500 nl). This was required to
237 ensure compatibility with downstream scRNA-seq analysis. To generate a 500 nl extraction
238 volume, a 5 μl glass capillary tube (Sigma-Aldrich) was used. First, the middle part of the
239 capillary tube was heated up slowly in a Bunsen burner flame. Then, when the tube started
240 melting, it was pulled rapidly downward, splitting into two pieces and forming a thin capillary.
241 This capillary tube inserted in the 1 mm extraction wells allowed the extraction of single cells in
242 a 500 nl volume.

243 Extraction experiments were performed only on 6C2 cells, because they express a mCherry
244 transgene (640 nm emission), and thus fluorescence imaging could be used to follow the cell
245 extraction process. In addition, a range of different LED light intensities were assessed, with a
246 view to minimizing fluorescence photobleaching during continuous imaging 6C2 cells. The
247 extraction protocol of the 6C2 cells included tracking sister cells after division, using both
248 brightfield and fluorescence imaging. The sister cells were driven to the extraction well by
249 applying 10 mbar of pressure, resulting in a cellular velocity of 10 $\mu\text{m/s}$ (Figure 3.A).
250 Fluorescence imaging was used to track single cells subsequent to their delivery into the
251 extraction wells (Figure 3.B). Next, single cells were manually extracted from the device using
252 the thin capillary tube, with simultaneous monitoring the fluorescence signal of the cells to

253 ensure that only a single cell is collected (Figure 3.C). This method proved to work successfully
254 with an extraction volume of 500 nl, which was compatible with downstream analysis.

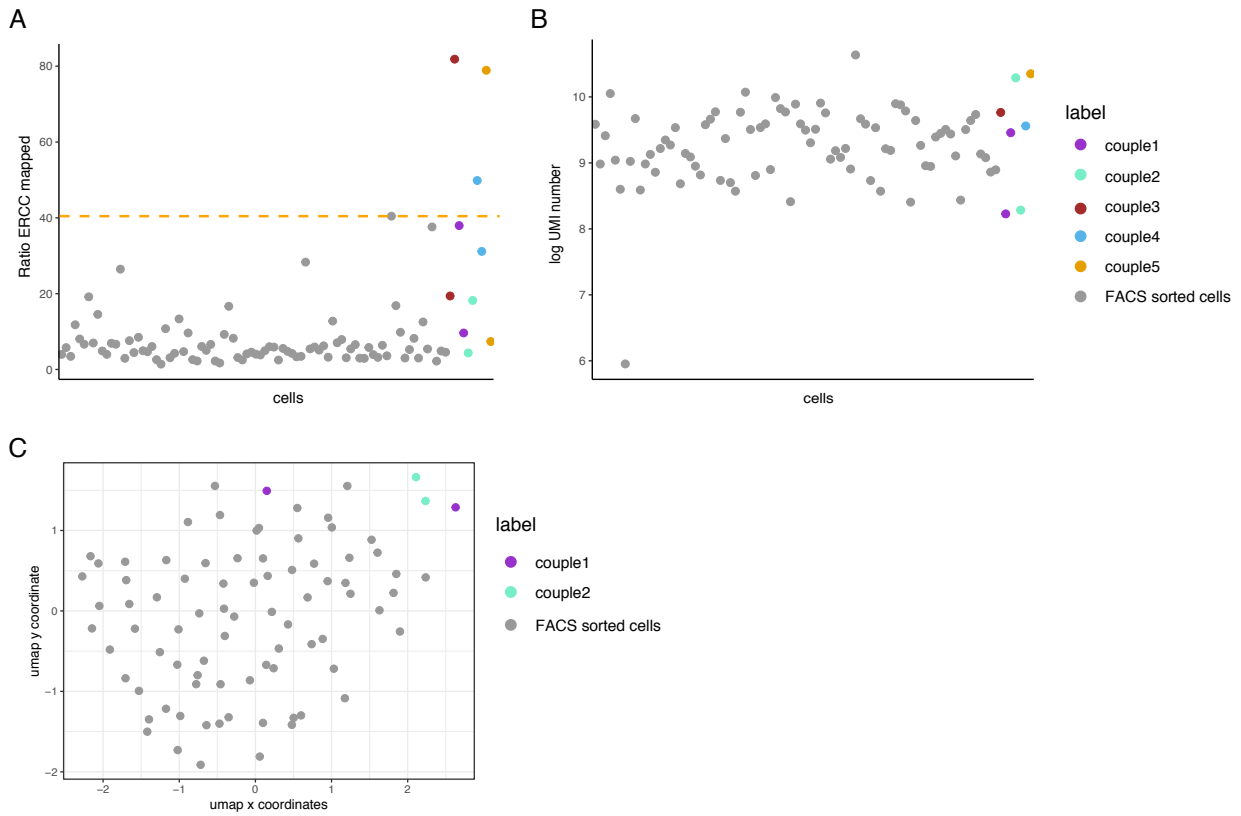


255
256 **Figure 3: Sister cell extraction.** (A) A single sister 6C2 cell was monitored using brightfield
257 imaging until it reached the extraction well. Fluorescence images of the extraction well before
258 (B) and after the extraction of a single sister cell (C). The scale bars are 100 μ m.
259

260 6C2 Experiment and Downstream Analysis

261 During this pilot experiment, performed on 6C2 cells, 5 couples of sister cells were followed and
262 extracted from the chip after one cell division (10 cells). The 10 isolated sister cells were directly
263 lysed and barcoded with unique cell barcode by using RT primers with a known sequence.
264 Before processing the library and as control, 86 6C2 FACS sorted single cells were added from a
265 population where relationship between the cells were unknown. FACS sorting cells were used as
266 controls as FACS sorting is the reference method to isolate single-cell for subsequent scRNA-seq
267 experiment. The library was then constructed as describe here³³ and sequenced. The raw data
268 were processed on our bio-informatics pipeline, filtered and normalized (more details in
269 material and methods). After data cleaning, 7 cells remained out of ten. Indeed, 3 cells were
270 probably damaged and have not been recovered as shown by their high content of ERCC spikes
271 RNA (Figure 4.A). Among the 7 cells, 2 complete couples of sister cells were recovered. The
272 sister cells isolated using our chip displayed the same amount of mean detected genes and

273 mean UMIs per cells as regular FACS sorted cells (Figure 4.B). Furthermore, applying UMAP
274 dimensionality reduction and projection method showed that chip-cultured and isolated cells
275 did not strongly segregate from the control FACS sorted cells (Figure 4.C).



276

Figure 4: scRNA-seq data visualization. (A) Plot of percent ERCC mapped in each cell. (B) Plot of log(UMIs) number per cells. (C) UMAP dimensions reduction and projection of the cells. Chip-cultured cells are colored and grouped by lineage and FACS sorted cells are white.

277

278 **Conclusions and Discussion**

279 In this study, we have described the development of a multilayer microfluidic device for tracking
280 non-adherent cells divisions at the single-cell level. The microfluidic device can trap single cells
281 in eight independently controlled proliferation chambers, isolate sister cells after division and
282 extract them for downstream analysis. We showed that the system is capable of tracking cells
283 over two cell generations using two cell models (a cell line and primary cells). The setup
284 incorporates an image processing-based device and automation of the cell loading, long-term
285 cell monitoring and cell extraction processes. In the presented experiments, both 6C2 (chicken
286 erythroleukemia cell line) and T2EC (primary chicken erythrocytic progenitors) cells proliferated
287 inside the chip, with a viability rate higher than 90%. The divided cells were separated using a
288 valve-based (T-shaped) structure and placed inside the 500 nl-volume extraction chambers,
289 which were compatible with downstream scRNA-seq analysis. Our general method allows to
290 recover selected single-cell and to record the genealogy information of the cell while providing
291 the same data quality required for subsequent scRNA-seq analysis as provided by regular FACS
292 sorting. The developed system provides a robust and automated platform for single-cell lineage
293 tracking studies at the resolution of cell division, and can easily be used for tracking non-
294 adherent cell types, including cell lines and primary cells as demonstrated in our work. In a
295 future perspective, the device could be used to perform perturbation experiments, including
296 differentiation induction and drugs, by changing culture reagent during the culture of the cells.
297 Moreover, throughput of the device can be improved by the increasing the number of parallel
298 proliferation chambers.

299

300 **Acknowledgements**

301 We thank the computational center of IN2P3 (Villeurbanne/France) and Pôle Scientifique de
302 Modélisation Numérique (PSMN, Ecole Normale Supérieure de Lyon) where computations were
303 performed. We acknowledge the contribution of the AniRA-Cytométrie core facility of SFR
304 BioSciences (UAR3444/US8). We thank the BioSyL Federation and the LabEx Ecofect (ANR-11-
305 LABX-0048) of the University of Lyon for inspiring scientific events.

306

307 **Funding**

308 This work was supported by funding from the French agency ANR (SinCity; ANR-17-CE12-0031).

309

310 **Availability of Data and Material**

311 The datasets supporting the conclusions of this article are available in the NIH repository,
312 accession number PRJNA882740, under embargo until publication.

313 R scripts are available on the Git repository <https://gitbio.ens-lyon.fr/cfournea/sincity>

314

315 **Authors' Contributions**

316 MKA contributed to the conceptualization of the study, designed and manufactured the chips,
317 performed the proliferation and extraction experiments, writing of the manuscript. CF
318 contributed to the conceptualization of the study, performed the proliferation and extraction
319 experiments, generated the scRNAseq data and analyzed it, writing of the manuscript. RP
320 participated in the proliferation experiments. SG, AP, AJDM and OG participated to the
321 conceptualization of the study, the project administration, the project supervision, and writing
322 of the manuscript.

323 **References**

- 324 (1) Elowitz, M. B. Stochastic Gene Expression in a Single Cell. *Science* **2002**, *297* (5584),
325 1183–1186. <https://doi.org/10.1126/science.1070919>.
- 326 (2) Symmons, O.; Raj, A. What's Luck Got to Do with It: Single Cells, Multiple Fates, and
327 Biological Nondeterminism. *Mol. Cell* **2016**, *62* (5), 788–802.
328 <https://doi.org/10.1016/j.molcel.2016.05.023>.
- 329 (3) Guillemain, A.; Stumpf, M. P. H. Noise and the Molecular Processes Underlying Cell Fate
330 Decision-Making. *Phys. Biol.* **2021**, *18* (1), 011002. <https://doi.org/10.1088/1478-3975/abc9d1>.
- 331 (4) Karamitros, D.; Stoilova, B.; Aboukhalil, Z.; Hamey, F.; Reinisch, A.; Samitsch, M.; Quek,
332 L.; Otto, G.; Repapi, E.; Doondeea, J.; Usukhbayar, B.; Calvo, J.; Taylor, S.; Goardon, N.; Six, E.;
333 Pflumio, F.; Porcher, C.; Majeti, R.; Gottgens, B.; Vyas, P. Heterogeneity of Human Lympho-
334 Myeloid Progenitors at the Single Cell Level. *Nat. Immunol.* **2018**, *19* (1), 85–97.
335 <https://doi.org/10.1038/s41590-017-0001-2>.
- 336 (5) Baslan, T.; Hicks, J. Unravelling Biology and Shifting Paradigms in Cancer with Single-Cell
337 Sequencing. *Nat. Rev. Cancer* **2017**, *17* (9), 557–569. <https://doi.org/10.1038/nrc.2017.58>.
- 338 (6) Guillemain, A.; Duchesne, R.; Crauste, F.; Gonin-Giraud, S.; Gandrillon, O. Drugs
339 Modulating Stochastic Gene Expression Affect the Erythroid Differentiation Process. *PLOS ONE*
340 **2019**, *14* (11), e0225166. <https://doi.org/10.1371/journal.pone.0225166>.
- 341 (7) Moris, N.; Pina, C.; Arias, A. M. Transition States and Cell Fate Decisions in Epigenetic
342 Landscapes. *Nat. Rev. Genet.* **2016**, *17* (11), 693–703. <https://doi.org/10.1038/nrg.2016.98>.
- 343 (8) Richard, A.; Boullu, L.; Herbach, U.; Bonnafoux, A.; Morin, V.; Vallin, E.; Guillemain, A.;
344 Papili Gao, N.; Gunawan, R.; Cosette, J.; Arnaud, O.; Kupiec, J.-J.; Espinasse, T.; Gonin-Giraud, S.;
345 Gandrillon, O. Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability
346 Preceding Irreversible Commitment in a Differentiation Process. *PLOS Biol.* **2016**, *14* (12),
347 e1002585. <https://doi.org/10.1371/journal.pbio.1002585>.
- 348 (9) Moussy, A.; Cosette, J.; Parmentier, R.; da Silva, C.; Corre, G.; Richard, A.; Gandrillon, O.;
349 Stockholm, D.; Páldi, A. Integrated Time-Lapse and Single-Cell Transcription Studies Highlight the
350 Variable and Dynamic Nature of Human Hematopoietic Cell Fate Commitment. *PLOS Biol.* **2017**,

- 351 15 (7), e2001867. <https://doi.org/10.1371/journal.pbio.2001867>.
- 352 (10) Mojtahedi, M.; Skupin, A.; Zhou, J.; Castaño, I. G.; Leong-Quong, R. Y. Y.; Chang, H.;
353 Trachana, K.; Giuliani, A.; Huang, S. Cell Fate Decision as High-Dimensional Critical State
354 Transition. *PLOS Biol.* **2016**, *14* (12), e2000640. <https://doi.org/10.1371/journal.pbio.2000640>.
- 355 (11) Dussiau, C.; Boussaroque, A.; Gaillard, M.; Bravetti, C.; Zaroili, L.; Knosp, C.; Friedrich, C.;
356 Asquier, P.; Willems, L.; Quint, L.; Bouscary, D.; Fontenay, M.; Espinasse, T.; Plesa, A.; Sujobert,
357 P.; Gandrillon, O.; Kosmider, O. Hematopoietic Differentiation Is Characterized by a Transient
358 Peak of Entropy at a Single-Cell Level. *BMC Biol.* **2022**, *20* (1), 60.
359 <https://doi.org/10.1186/s12915-022-01264-9>.
- 360 (12) Hu, M.; Krause, D.; Greaves, M.; Sharkis, S.; Dexter, M.; Heyworth, C.; Enver, T.
361 Multilineage Gene Expression Precedes Commitment in the Hemopoietic System. *Genes Dev.*
362 **1997**, *11* (6), 774–785. <https://doi.org/10.1101/gad.11.6.774>.
- 363 (13) Phillips, N. E.; Mandic, A.; Omid, S.; Naef, F.; Suter, D. M. Memory and Relatedness of
364 Transcriptional Activity in Mammalian Cell Lineages. *Nat. Commun.* **2019**, *10* (1).
365 <https://doi.org/10.1038/s41467-019-09189-8>.
- 366 (14) Kimmerling, R. J.; Lee Szeto, G.; Li, J. W.; Genshaft, A. S.; Kazer, S. W.; Payer, K. R.; de
367 Riba Borrajo, J.; Blainey, P. C.; Irvine, D. J.; Shalek, A. K.; Manalis, S. R. A Microfluidic Platform
368 Enabling Single-Cell RNA-Seq of Multigenerational Lineages. *Nat. Commun.* **2016**, *7* (1), 10220.
369 <https://doi.org/10.1038/ncomms10220>.
- 370 (15) Shaffer, S. M.; Emert, B. L.; Reyes Hueros, R. A.; Cote, C.; Harmange, G.; Schaff, D. L.;
371 Sizemore, A. E.; Gupte, R.; Torre, E.; Singh, A.; Bassett, D. S.; Raj, A. Memory Sequencing Reveals
372 Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors. *Cell*
373 **2020**, *182* (4), 947-959.e17. <https://doi.org/10.1016/j.cell.2020.07.003>.
- 374 (16) Muramoto, T.; Müller, I.; Thomas, G.; Melvin, A.; Chubb, J. R. Methylation of H3K4 Is
375 Required for Inheritance of Active Transcriptional States. *Curr. Biol.* **2010**, *20* (5), 397–406.
376 <https://doi.org/10.1016/j.cub.2010.01.017>.
- 377 (17) Bellec, M.; Dufourt, J.; Hunt, G.; Lenden-Hasse, H.; Trullo, A.; Zine El Aabidine, A.;
378 Lamarque, M.; Gaskill, M. M.; Faure-Gautron, H.; Mannervik, M.; Harrison, M. M.; Andrau, J.-C.;

- 379 Favard, C.; Radulescu, O.; Lagha, M. The Control of Transcriptional Memory by Stable Mitotic
380 Bookmarking. *Nat. Commun.* **2022**, *13*, 1176. <https://doi.org/10.1038/s41467-022-28855-y>.
- 381 (18) Weinreb, C.; Rodriguez-Fraticelli, A. E.; Camargo, F. D.; Klein, A. M. Lineage Tracing on
382 Transcriptional Landscapes Links State to Fate during Differentiation. *bioRxiv* **2018**.
383 <https://doi.org/10.1101/467886>.
- 384 (19) Bidy, B. A.; Waye, S. E.; Sun, T.; Morris, S. A. Single-Cell Analysis of Clonal Dynamics in
385 Direct Lineage Reprogramming: A Combinatorial Indexing Method for Lineage Tracing. *bioRxiv*
386 **2017**. <https://doi.org/10.1101/127860>.
- 387 (20) Brody, Y.; Kimmerling, R. J.; Maruvka, Y. E.; Benjamin, D.; Elacqua, J. J.; Haradhvala, N. J.;
388 Kim, J.; Mouw, K. W.; Frangaj, K.; Koren, A.; Getz, G.; Manalis, S. R.; Blainey, P. C. Quantification
389 of Somatic Mutation Flow across Individual Cell Division Events by Lineage Sequencing. *Genome*
390 *Res.* **2018**, *28* (12), 1901–1918. <https://doi.org/10.1101/gr.238543.118>.
- 391 (21) Gao, D.; Jin, F.; Zhou, M.; Jiang, Y. Recent Advances in Single Cell Manipulation and
392 Biochemical Analysis on Microfluidics. *Analyst* **2019**, *144* (3), 766–781.
393 <https://doi.org/10.1039/C8AN01186A>.
- 394 (22) Taniguchi, K.; Kajiyama, T.; Kambara, H. Quantitative Analysis of Gene Expression in a
395 Single Cell by QPCR. *Nat. Methods* **2009**, *6* (7), 503–506. <https://doi.org/10.1038/nmeth.1338>.
- 396 (23) Ziegenhain, C.; Vieth, B.; Parekh, S.; Reinius, B.; Guillaumet-Adkins, A.; Smets, M.;
397 Leonhardt, H.; Heyn, H.; Hellmann, I.; Enard, W. Comparative Analysis of Single-Cell RNA
398 Sequencing Methods. *Mol. Cell* **2017**, *65* (4), 631–643.e4.
399 <https://doi.org/10.1016/j.molcel.2017.01.023>.
- 400 (24) Kaiser, M.; Jug, F.; Julou, T.; Deshpande, S.; Pfohl, T.; Silander, O. K.; Myers, G.; van
401 Nimwegen, E. Monitoring Single-Cell Gene Regulation under Dynamically Controllable
402 Conditions with Integrated Microfluidics and Software. *Nat. Commun.* **2018**, *9* (1), 212.
403 <https://doi.org/10.1038/s41467-017-02505-0>.
- 404 (25) Mehling, M.; Tay, S. Microfluidic Cell Culture. *Curr. Opin. Biotechnol.* **2014**, *25*, 95–102.
405 <https://doi.org/10.1016/j.copbio.2013.10.005>.

- 406 (26) Lin, J.; Jordi, C.; Son, M.; Van Phan, H.; Drayman, N.; Abasiyanik, M. F.; Vistain, L.; Tu, H.-
407 L.; Tay, S. Ultra-Sensitive Digital Quantification of Proteins and mRNA in Single Cells. *Nat.*
408 *Commun.* **2019**, *10* (1), 3544. <https://doi.org/10.1038/s41467-019-11531-z>.
- 409 (27) Waldherr, S. Estimation Methods for Heterogeneous Cell Population Models in Systems
410 Biology. *J. R. Soc. Interface* **2018**, *15* (147), 20180530. <https://doi.org/10.1098/rsif.2018.0530>.
- 411 (28) Unger, M. A.; Chou, H.-P.; Thorsen, T.; Scherer, A.; Quake, S. R. Monolithic
412 Microfabricated Valves and Pumps by Multilayer Soft Lithography. *Science* **2000**, *288* (5463),
413 113–116. <https://doi.org/10.1126/science.288.5463.113>.
- 414 (29) Melin, J.; Roxhed, N.; Gimenez, G.; Griss, P.; van der Wijngaart, W.; Stemme, G. A Liquid-
415 Triggered Liquid Microvalve for on-Chip Flow Control. *Sens. Actuators B Chem.* **2004**, *100* (3),
416 463–468. <https://doi.org/10.1016/j.snb.2004.03.010>.
- 417 (30) Gandrillon, O.; Samarut, J. Role of the Different RAR Isoforms in Controlling the
418 Erythrocytic Differentiation Sequence. Interference with the v-ErbA and P135gag-Myb-Ets
419 Nuclear Oncogenes. *Oncogene* **1998**, *16* (5), 563–574. <https://doi.org/10.1038/sj.onc.1201550>.
- 420 (31) Gandrillon, O.; Schmidt, U.; Beug, H.; Samarut, J. TGF- β Cooperates with TGF- α to Induce
421 the Self-Renewal of Normal Erythrocytic Progenitors: Evidence for an Autocrine Mechanism.
422 *EMBO J.* **1999**, *18* (10), 2764–2781. <https://doi.org/10.1093/emboj/18.10.2764>.
- 423 (32) Jaitin, D. A.; Kenigsberg, E.; Keren-Shaul, H.; Elefant, N.; Paul, F.; Zaretsky, I.; Mildner, A.;
424 Cohen, N.; Jung, S.; Tanay, A.; Amit, I. Massively Parallel Single-Cell RNA-Seq for Marker-Free
425 Decomposition of Tissues into Cell Types. *Science* **2014**, *343* (6172), 776–779.
426 <https://doi.org/10.1126/science.1247651>.
- 427 (33) Zreika, S.; Fourneaux, C.; Vallin, E.; Modolo, L.; Seraphin, R.; Moussy, A.; Ventre, E.;
428 Bouvier, M.; Ozier-Lafontaine, A.; Bonnaffoux, A.; Picard, F.; Gandrillon, O.; Gonin-Giraud, S.
429 *Evidence for Close Molecular Proximity between Reverting and Undifferentiated Cells*; preprint;
430 *Cell Biology*, 2022. <https://doi.org/10.1101/2022.02.01.478637>.
- 431 (34) Di Tommaso, P.; Chatzou, M.; Floden, E. W.; Barja, P. P.; Palumbo, E.; Notredame, C.
432 Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* **2017**, *35* (4), 316–
433 319. <https://doi.org/10.1038/nbt.3820>.

434 (35) R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation
435 for Statistical Computing: Vienna, Austria, 2021.

436 (36) Cole, M. B.; Risso, D.; Wagner, A.; DeTomaso, D.; Ngai, J.; Purdom, E.; Dudoit, S.; Yosef,
437 N. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq.
438 *Cell Syst.* **2019**, *8* (4), 315-328.e8. <https://doi.org/10.1016/j.cels.2019.03.010>.

439 (37) Breda, J.; Zavolan, M.; van Nimwegen, E. Bayesian Inference of Gene Expression States
440 from Single-Cell RNA-Seq Data. *Nat. Biotechnol.* **2021**, *39* (8), 1008–1016.
441 <https://doi.org/10.1038/s41587-021-00875-x>.

442 (38) Hafemeister, C.; Satija, R. *Normalization and Variance Stabilization of Single-Cell RNA-*
443 *Seq Data Using Regularized Negative Binomial Regression*; preprint; Genomics, 2019.
444 <https://doi.org/10.1101/576827>.

445 (39) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W. H.; Ng, L. G.; Ginhoux, F.;
446 Newell, E. W. Dimensionality Reduction for Visualizing Single-Cell Data Using UMAP. *Nat.*
447 *Biotechnol.* **2018**, *37* (1), 38–44. <https://doi.org/10.1038/nbt.4314>.

448

Chapitre 4

Caractérisation moléculaire de la réversibilité phénotypique de progéniteurs érythrocytaires induits à se différencier

4.1 Introduction

Précédemment dans l'équipe, une analyse de l'expression génique à l'échelle de la cellule unique avait mis en évidence un point d'engagement critique au cours de la différenciation des progéniteurs érythrocytaires. En effet, il a été observé que dans les 8h à 24h suivant l'induction de la différenciation, les progéniteurs érythrocytaires présentaient une augmentation de la variabilité de l'expression génique quantifiée par l'entropie de Shannon. De plus, jusqu'à 24h de différenciation si ces cellules étaient replacées dans du milieu d'auto-renouvellement, elles étaient capables de proliférer à nouveau, ce qui n'était plus le cas à des temps plus tardifs (48h) [160].

En collaboration avec Souad Zreika, également doctorante en thèse dans l'équipe, nous avons réalisé une caractérisation moléculaire approfondie de ces cellules, qui n'avait pas été faite dans les travaux de Richard *et al.* [160]. Nous nous sommes demandées si les cellules induites 24h en différenciation puis replacées en milieu d'auto-renouvellement subissaient une réversion complète et étaient donc identiques, au niveau moléculaire, à des cellules indifférenciées ou si elles

conservaient des traces moléculaires de leur engagement en différenciation.

Ces travaux ont fait l'objet d'une publication dans BMC Biology [201] : Zreika S*, Fourneaux C*, Vallin E, Modolo L, Seraphin R, Moussy A, Ventre E, Bouvier M, Ozier-Lafontaine A, Bonnafoux A, Picard F, Gandrillon O[&] and Gonin-Giraud S[&]. Evidence for close molecular proximity between reverting and undifferentiated cells. BMC Biol. 2022. * Co-premiers auteurs ;
[&] co-derniers auteurs

4.2 Résumé des résultats

Dans cette étude, nous avons cultivé des progéniteurs érythrocytaires aviaires (T2EC) en état d'auto-renouvellement (point 0h) puis nous les avons induits à se différencier pendant 24h (point 24h), avant de diviser la population en deux. La première moitié des cellules a été placée dans du milieu de différenciation pendant 24h supplémentaires (point 48h de différenciation) et l'autre moitié a été placée dans du milieu d'auto-renouvellement pendant 24h également (point 48h réversion). Pour les 4 points de la cinétique, nous avons récupéré environ 200 cellules et avons analysé leurs transcriptomes séparément à l'aide de deux technologies de transcriptomique en cellule unique : le sc-RNA-seq et la scRT-qPCR.

Nous avons d'abord comparé les transcriptomes des quatre groupes (4 points de temps de la cinétique) de manière globale en utilisant la méthode de réduction de dimension et de projection UMAP. Nous avons ensuite effectué une analyse statistique pour comparer la proximité des transcriptomes des cellules entre chacun des groupes. Pour caractériser davantage les cellules, nous avons comparé aussi les profils d'expression génique entre les groupes et utilisé différents outils statistiques comme l'analyse d'expression différentielle, la sparse PLS et la comparaison des distributions de gènes en utilisant la distance de Wasserstein. Enfin, nous avons vérifié que les cellules en réversion ne provenaient pas d'une sous-population de cellules indifférenciées qui ne seraient jamais entrées dans le processus de différenciation.

Nos résultats principaux sont les suivants :

- 1) La réduction de dimension UMAP a montré que les cellules en réversion sont presque indiscernables des cellules indifférenciées ;

- 2) La comparaison des distributions des cellules n'a indiqué aucune différence significative entre les cellules indifférenciées et les cellules en réversion et l'analyse de l'expression différentielle n'a montré aucune différence majeure dans l'expression des gènes entre les cellules en réversion et les cellules indifférenciées ;
- 3) En examinant de plus près la distribution des gènes, les cellules en réversion ont conservé des traces de leur engagement dans le processus de différenciation. En effet, leurs profils d'expression génique, bien que très proches de ceux de cellules indifférenciées, ne sont pas totalement identiques. Confirmant ces résultats, l'analyse sPLS a révélé que l'expression de 3 gènes seulement permet de prédire l'appartenance des cellules au groupe « réversion » ou au groupe « indifférencié ». Bien que pour deux des trois gènes, la durée de la demi-vie des ARNm ne semble pas être la cause du délai observé, il n'est pas exclu qu'il s'agisse d'un simple délai, et que l'étude des cellules en réversion à un temps plus tardif montrerait un retour en arrière complet ;
- 4) Enfin, nous avons exclu l'hypothèse selon laquelle les cellules en réversion seraient des cellules indifférenciées qui ne se sont jamais engagées dans le processus de différenciation.

4.3 Principales conclusions

L'ensemble de nos résultats ont montré que les cellules induites à se différencier et capables de revenir à un état phénotypique caractéristique de cellules en auto-renouvellement ré-acquièrent également un état transcriptomique très proche de celui des cellules indifférenciées.

Notre étude met donc en lumière une plasticité physiologique et moléculaire au début de la différenciation érythrocytaire ; elle met aussi en évidence la nature probabiliste de la différenciation, par opposition à une vision déterministe et stéréotypée, grâce à l'utilisation combinée de différents outils statistiques et bio-informatiques.

Enfin, la prise de décision cellulaire est régie par la dynamique du réseau de régulation génique (GRN) sous-jacent, qui est lui même influencé par des signaux environnementaux. Nos travaux ouvrent donc des perspectives d'optimisation des GRN ; en effet cette possibilité de réversion imposerait une contrainte très forte qui devrait être prise en compte dans les algorithmes d'inférence des GRN.


4.4 Publication - article 3

RESEARCH ARTICLE

Open Access



Evidence for close molecular proximity between reverting and undifferentiated cells

Souad Zreika^{1,2†}, Camille Fourneau^{1†}, Elodie Vallin¹, Laurent Modolo¹, Rémi Seraphin¹, Alice Moussy³, Elias Ventre^{1,4,5}, Matteo Bouvier^{1,6}, Anthony Ozier-Lafontaine⁷, Arnaud Bonnaffoux^{1,6}, Franck Picard¹, Olivier Gandrillon^{1,4†} and Sandrine Gonin-Giraud^{1*†} 

Abstract

Background: According to Waddington's epigenetic landscape concept, the differentiation process can be illustrated by a cell akin to a ball rolling down from the top of a hill (proliferation state) and crossing furrows before stopping in basins or "attractor states" to reach its stable differentiated state. However, it is now clear that some committed cells can retain a certain degree of plasticity and reacquire phenotypical characteristics of a more pluripotent cell state. In line with this dynamic model, we have previously shown that differentiating cells (chicken erythrocytic progenitors (T2EC)) retain for 24 h the ability to self-renew when transferred back in self-renewal conditions. Despite those intriguing and promising results, the underlying molecular state of those "reverting" cells remains unexplored. The aim of the present study was therefore to molecularly characterize the T2EC reversion process by combining advanced statistical tools to make the most of single-cell transcriptomic data. For this purpose, T2EC, initially maintained in a self-renewal medium (0H), were induced to differentiate for 24H (24H differentiating cells); then, a part of these cells was transferred back to the self-renewal medium (48H reverting cells) and the other part was maintained in the differentiation medium for another 24H (48H differentiating cells). For each time point, cell transcriptomes were generated using scRT-qPCR and scRNAseq.

Results: Our results showed a strong overlap between 0H and 48H reverting cells when applying dimensional reduction. Moreover, the statistical comparison of cell distributions and differential expression analysis indicated no significant differences between these two cell groups. Interestingly, gene pattern distributions highlighted that, while 48H reverting cells have gene expression pattern more similar to 0H cells, they are not completely identical, which suggest that for some genes a longer delay may be required for the cells to fully recover. Finally, sparse PLS (sparse partial least square) analysis showed that only the expression of 3 genes discriminates 48H reverting and 0H cells.

Conclusions: Altogether, we show that reverting cells return to an earlier molecular state almost identical to undifferentiated cells and demonstrate a previously undocumented physiological and molecular plasticity during the differentiation process, which most likely results from the dynamic behavior of the underlying molecular network.

Keywords: Cell-fate reversion, Cell differentiation, Erythroid progenitors, Single-cell RNA-seq, Single-cell RT-qPCR

[†]Souad Zreika, Camille Fourneau, Olivier Gandrillon and Sandrine Gonin-Giraud contributed equally to this work.

*Correspondence: sandrine.giraud@ens-lyon.fr

¹ Laboratory of Biology and Modelling of the Cell, Université de Lyon, Ecole Normale Supérieure de Lyon, CNRS, UMR5239, Université Claude Bernard Lyon 1, Lyon, France
Full list of author information is available at the end of the article

Background

The integration and processing of endogenous and exogenous information constitute a fundamental requirement for cells to ensure functions and survival of unicellular or multicellular organisms. Cellular decision-making is then at the core of the physiological or pathological functioning of living organisms. Early



views of the mechanisms governing cell-fate decision-making, and in particular cell differentiation, were based on bulk population data, leading to an over-simplifying deterministic framework. In these first views, cell commitment to a predefined cell type was thought to be triggered through a stereotyped sequence of intermediate states under the influence of specific signals [1].

Single-cell approaches have allowed to change the scale of observation of many molecular processes and revealed that an important heterogeneity in gene expression lies at the heart of isogenic cell populations [2, 3]. Stochasticity in gene expression arises from different causes, such as the probabilistic nature of molecular interactions or transcriptional bursts [4]. Cell-to-cell variability is visible at all omics levels of gene expression, but is being widely studied at the transcriptomic level since various molecular biology tools are available for this scale of investigation [5]. Overall, this heterogeneity in gene expression has been shown to be critical for the process of differentiation, as it provides diversity without the cost of hard-wire-encoded fate programs [6, 7].

Furthermore, single-cell studies have also enabled the development of stochastic models to describe differentiation from single-cell transcriptomic data. One of the best-known models is Conrad Waddington's landscape, which also includes the non-genetic part of cell-to-cell heterogeneities [8]. According to Waddington's model, the shape of the landscape is determined by Gene Regulatory Networks (GRN) and state transitions are modelled as channeling events: a cell, presented as a ball, starts from a mountain top and crosses valleys before reaching a stable state by occupying basins or attractor states, shaped by an underlying GRN [9]. Once this stable state is reached, the state potential decreases and the associated cell fate is restricted or even irreversible [10].

However, it is now clearly accepted that some cells retain fate plasticity [11, 12]. Under the forced modification of transcription factors stoichiometry, a cell that has reached a differentiated state can return to a more pluripotent stage challenging the classical hierarchical view of differentiation [13, 14]. Quite interestingly, spontaneous fate reversion can be observed under a physiological or damaging condition where progenitors or even more committed cells return to an earlier stage, potentially more pluripotent, and reacquire progenitor or stem-cell-like phenotypes and characteristics [15–18]. In this view, our recent study has shown that chicken primary erythroid progenitor cells (T2EC) have retained the capacity to go back to a self-renewal state for up to 24H after the induction of differentiation before they irreversibly engaged in the differentiation process [19]. Despite intriguing and promising results, the molecular

determinants of this so-called fate reversion and the molecular characterization of the reverting cells remain unexplored.

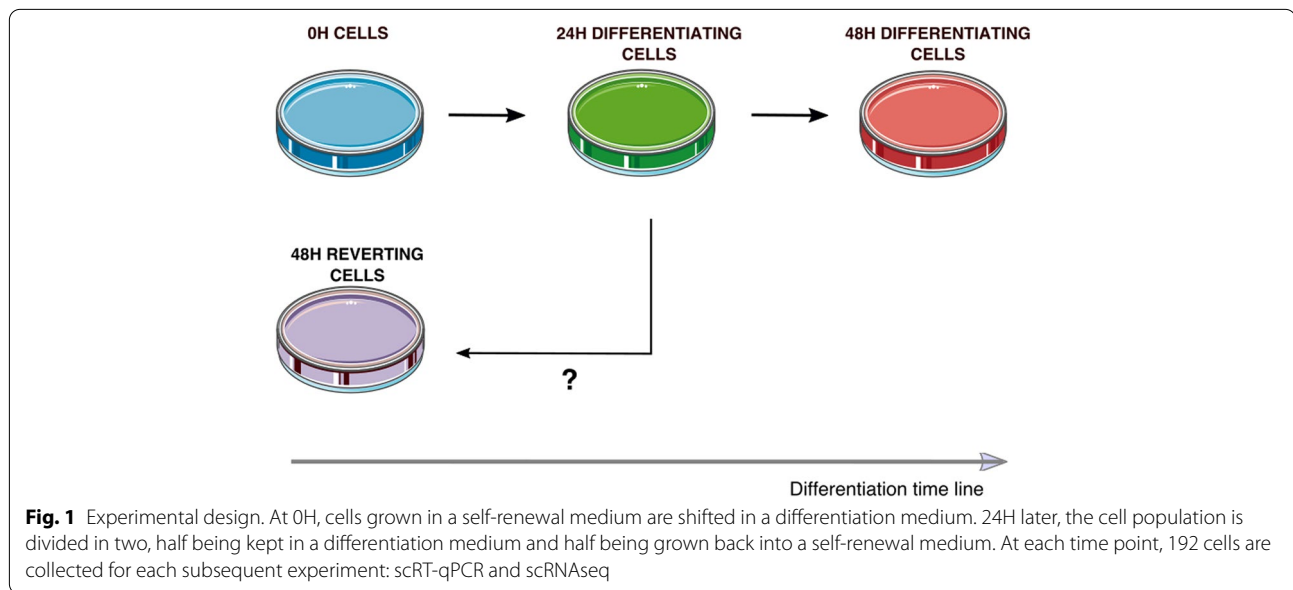
In this work, we go beyond the cellular and phenotypic characterization of the cell reversion process. We characterize the gene expression of primary erythroid progenitors and question if reverting cells undergo an actual fate reversion, i.e., in addition to regaining a comparable cellular state, reacquire a molecular state similar to undifferentiated cells.

For this, differentiation of self-renewing cells was induced by medium change during 24H. Then, we split the differentiating population so that half could pursue differentiation, and the second half was shifted back in a self-renewal medium (Fig. 1). To provide robust quantitative measurements of gene expression variability, we combined a highly sensitive targeted quantification method (scRT-qPCR) with genome-wide scRNAseq data to characterize the transcriptome of each population at the single-cell level: undifferentiated (0H), differentiating (24H and 48H), and reverting (48H reverting) cells. Our statistical analyses show that 48H reverting cells and undifferentiated cells were much more similar, whereas a separation was clearly visible between cells maintained in differentiation (48H differentiating cells) and cells in reversion (48H reverting cells). Furthermore, a statistical comparison of cell distributions indicated no significant differences between 0H cells and 48H reverting cells. Moreover, gene expression pattern distribution of 48H reverting cells showed a shift towards expression pattern distribution of 0H cells. Finally, we identified genes that discriminate 48H reverting cells and 0H cells. Using sparse PLS [20], we were able to show that the expression of 3 genes, *HBBA*, *TBC1D7*, and *HSP90AA1*, was discriminant between 48H reverting cells and 0H cells showing that reverting cells kept transcriptional traces of their induction to differentiation. In conclusion, our results show that reverting cells display gene expression patterns that are very similar to undifferentiated cells while retaining traces of their response to differentiation induction, which suggests an almost complete molecular reversion after 24H of differentiation induction.

Results

Robustness of single-cell transcriptomics analysis

We sought to characterize at the molecular level the cells that were induced to differentiate for 24 h and that retained the ability to proliferate when placed back into a self-renewal medium. We used two different complementary single-cell transcriptomics technologies, scRT-qPCR and scRNAseq. scRT-qPCR allows for highly sensitive quantification but is knowledge-driven and offers information of a limited number of genes while



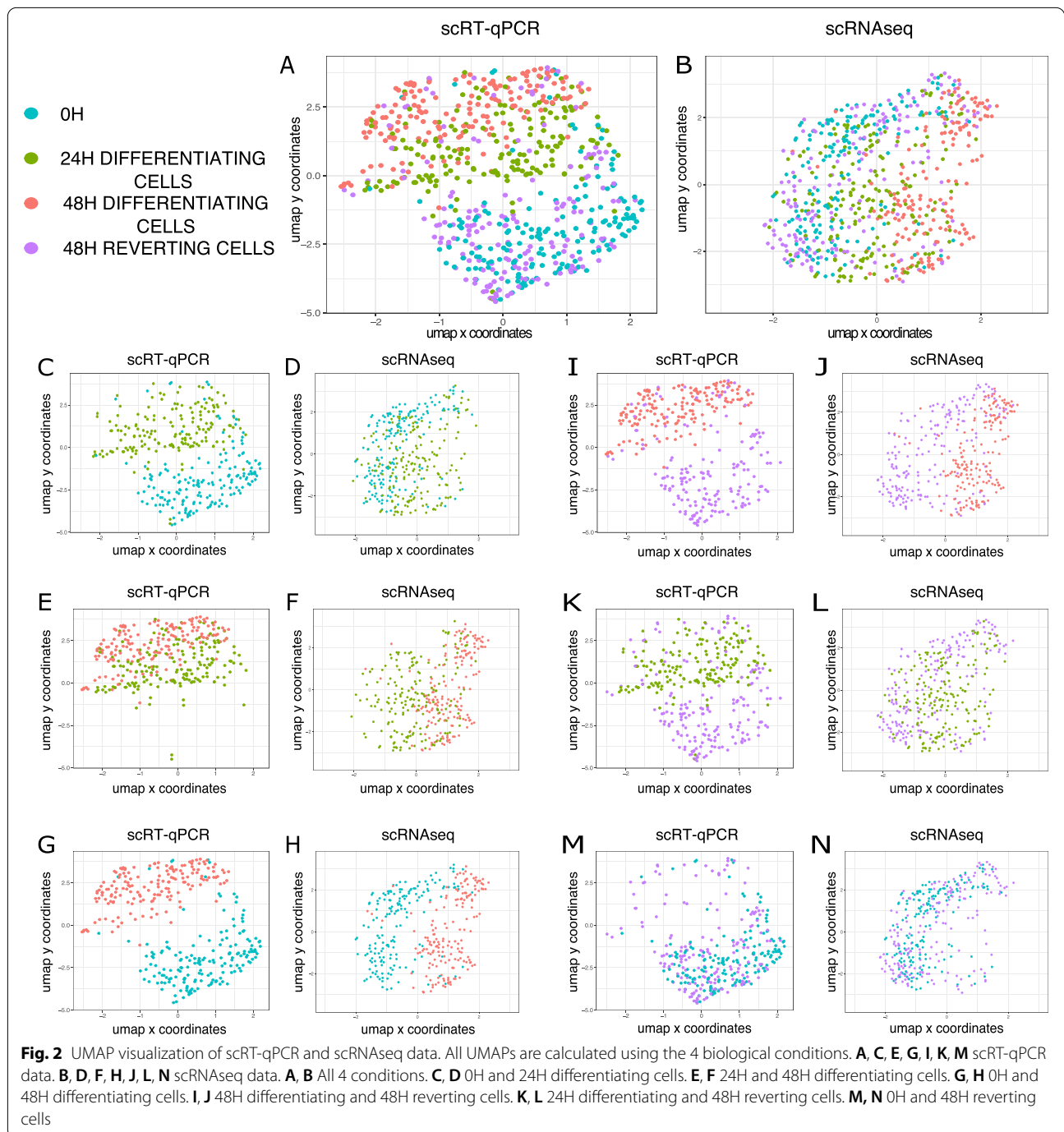
scRNAseq, although less precise for low expression level [21], enables genome-wide quantification without prior knowledge. Furthermore, using two different single-cell technologies allowed us to cross-validate our observations and point toward robust conclusions.

We first obtained by scRT-qPCR the expression level of 83 genes involved in T2EC differentiation in 173, 173, 168, and 171 cells for 0H, 24H, and 48H of differentiation and 48H reverting cells, respectively. Those genes are known to distinguish cells along the differentiation process and include sterol biosynthesis, metabolism, globin subunits, and transcription factors expressed by erythroid progenitors as published in [19]. The robustness of our measurements was confirmed by a Pearson's correlation of 0.85 (p -value = $2.2e-16$) between our experiments and the published data [19]. To investigate fate-reversion genome-wide by scRNAseq, we adapted the MARSseq (massively parallel single-cell RNA-Seq [22] — see the "Methods" section). Then, we obtained gene expression levels in 174, 181, 169, and 186 single cells for 0H, 24H, and 48H of differentiation and 48H reverting cells, respectively. The concordance between scRT-qPCR and scRNAseq data was confirmed by a Pearson's correlation of 0.73 (p -value = $1.34e-13$) between the 74 genes common to both datasets.

Similarity between reverting and undifferentiated cells revealed by dimension reduction

We used UMAP to uncover potential similarities between 48H reverting cells and subgroups of differentiating cells by projecting the 4 conditions (Fig. 2A,

scRT-qPCR data, and Fig. 2B, scRNAseq data). Then, we focused on the normal differentiation process using the 3 time points of differentiation (0H, 24H, and 48H differentiating cells) (Fig. 2C–H). For both experiments, pairwise representations show that 24H differentiating cells tend to overlap with both 0H cells (Fig. 2C, D) and 48H differentiating cells (Fig. 2E, F). On the contrary, the undifferentiated cells and 48H differentiating cells clearly differ (Fig. 2G, H). Interestingly, pairwise representations also reveal that 48H reverting cells separate well from the 48H differentiating cells (Fig. 2I, J) and from 24H cells (Fig. 2K, L), but are visually not distinguishable from the 0H cells (Fig. 2M, N). Almost identical results were observed when, instead of plotting cells on the UMAPs calculated from the mix of the 4 conditions, we recalculated the UMAPs for each pair of conditions (Additional file 1: Fig. S1). Principal Component Analysis (PCA) also captured this general separation of the data (Additional file 1: Fig. S2). Those analyses suggest that the transcriptomes of 48H reverting cells are more similar to the undifferentiated cells than to any other condition at both scales of observation. This was further confirmed by the pairwise statistical comparison of average scRNAseq distributions ([23] — see the "Methods" section). As shown in Table 1, the average transcriptomes of 48H reverting and 48H differentiating cells are significantly different, as well as of undifferentiated and 48H differentiating cells. In contrast, no significant difference in average transcriptomes was detected between 0H and 48H reverting conditions (p -value $\gg 0.05$), indicating a very close proximity of 48H reverting cells to undifferentiated cells.



48H reverting cells and undifferentiated cells have similar gene expression patterns

We then questioned if 48H reverting cells had gene expression patterns identical to 0H cells or retained, for some genes, an expression pattern more similar to 24H or 48H differentiating cells.

Pairwise scRNAseq DE (differential expression) analysis revealed that the “normal” erythrocyte differentiation

Table 1 *p*-value output of multivariate two tests between pair of conditions compared

	0H vs 48H reverting	0H vs 48H differentiating	48H reverting vs 48H differentiating
<i>p</i> -value	1.00	0	0.00000000369

process showed an increase in the expression of hemoglobin-related genes during the kinetics (hemoglobin subunit epsilon 1 (*HBBA*), hemoglobin alpha-locus 1 (*HBA1*), and hemoglobin alpha, subunit D (*HBAD*)) (Fig. 3A–C). On the other hand, 0H cells expressed a high level of *LDHA* (lactate dehydrogenase A), a marker for

glycolysis metabolism used by self-renewing cells [24], and *ID2* (inhibitor of DNA binding 2) coding for a transcription factor involved in differentiation inhibition [25].

Interestingly, when comparing 0H with 48H reverting cells, we saw only one gene that was significantly differentially expressed just above the threshold (Fig. 3D), the

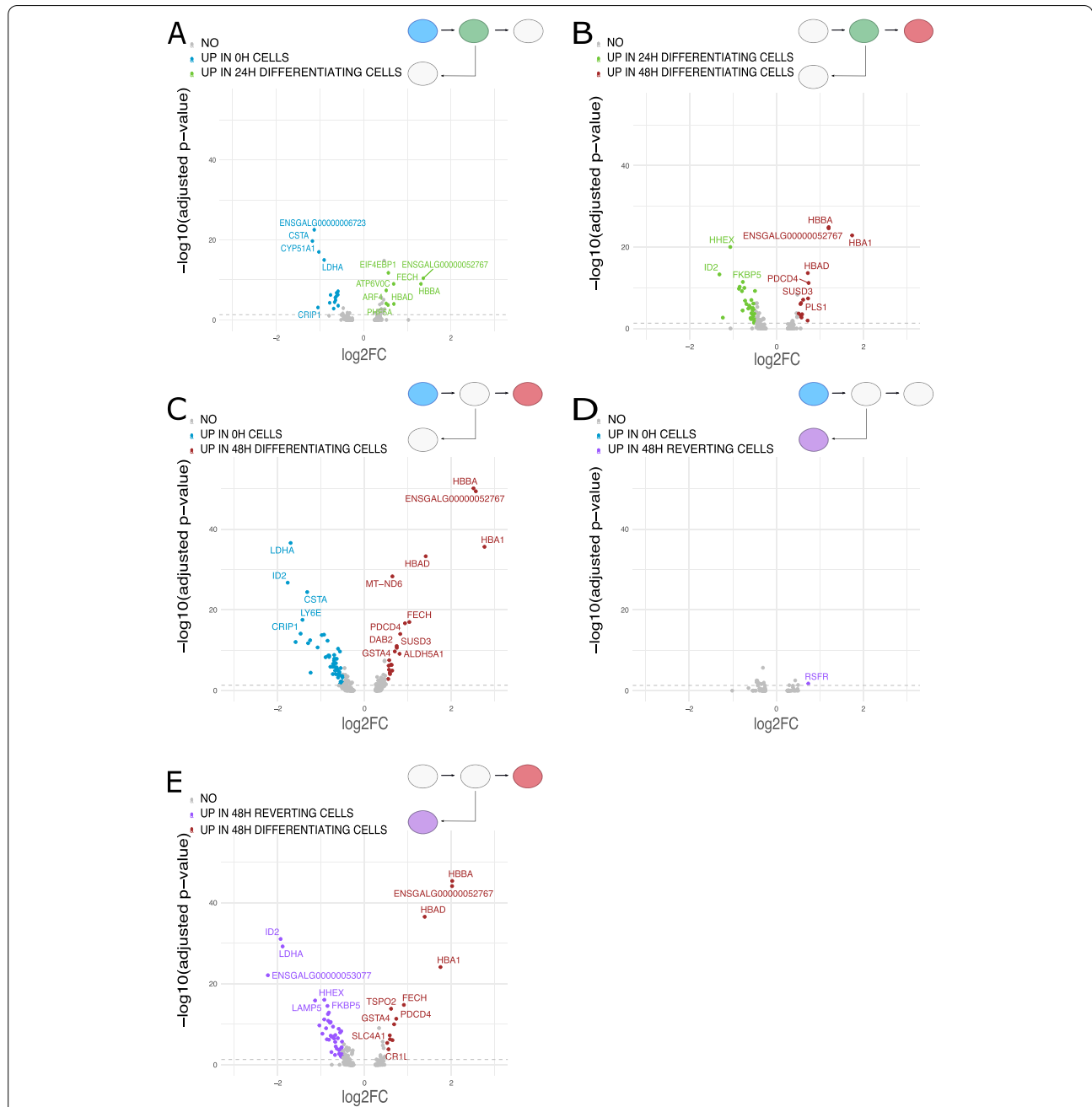


Fig. 3 Volcano plot of differentially expressed genes from scRNAseq data between conditions analyzed two by two. **A** 0H and 24H differentiating cells. **B** 24H differentiating and 48H differentiating cells. **C** 0H and 48H differentiating cells. **D** 0H and 48H reverting cells. **E** 48H reverting and 48H differentiating cells. Genes are considered significantly differentially expressed when the fold change is equal to or above 0.5 and the adjusted *p*-value is below 0.05

RSFR (RNase super family related) gene, that is highly expressed in precursor cells from chicken bone marrow [26]. Furthermore, when comparing 48H reverting with 48H differentiating cells, we found hemoglobin-related genes up in the differentiating cells and *LDHA* and *ID2* up in reverting cells (Fig. 3E).

We more closely investigated gene expression distributions within the different conditions to see how gene expression patterns would evolve during the reversion process (Fig. 4). We selected 8 genes differentially expressed and which expression increases or decreases during the differentiation process. *HBA1*, *HBBA*, *HBAD* (different hemoglobin subunits), and *FECH* (Ferrochelatase) are involved in hemoglobin and heme pathways and are more expressed by differentiating cells while *LDHA*, *ID2*, *CSTA* (cystatin A1), and *CRIP1* (Cysteine-rich intestinal protein1) are more expressed by self-renewing undifferentiated cells. We plotted and compared their distribution between the 4 conditions. For the genes involved in differentiation, we see a gradual shift in the distributions towards a higher level of expression as cells get more differentiated (Fig. 4A–D) and we see the opposite shift for genes involved in proliferation (Fig. 4E–H). In all cases, the 48H reverting cell expression patterns for those genes shifted back to patterns closer to the 0H cells. At the time of observation and especially for genes up in differentiation, the 48H reverting cell expression patterns are not completely similar to those of 0H cells. This was further confirmed by using a dedicated statistical tool, sparse PLS (see below).

To go further on gene distribution comparisons, we computed Wasserstein distances, a geometric distance well suited for comparing multimodal distributions, for each 2000 genes of the scRNAseq dataset between each condition two by two. We then obtain 6 distributions of Wasserstein distance values. Finally, we computed the Gini index as a measure of statistical dispersion in each distribution (the higher the Gini index is, the higher inequality among the values). We performed 100 bootstraps and compared the Gini values obtained (Fig. 5A). Distribution of Wasserstein distances between 0H cells and 48H reverting cells had the smallest average Gini index among all distributions (Fig. 5B). This result points towards a closer global transcriptional state between 48H reverting cells and 0H cells.

48H reverting cells retain molecular traces of a commitment into differentiation

To further characterize the molecular changes that persisted after reversion, we sought to identify predictive genes that discriminate the most the 48H reverting cells and the undifferentiated cells. We performed logistic regression combined with dimension reduction (partial

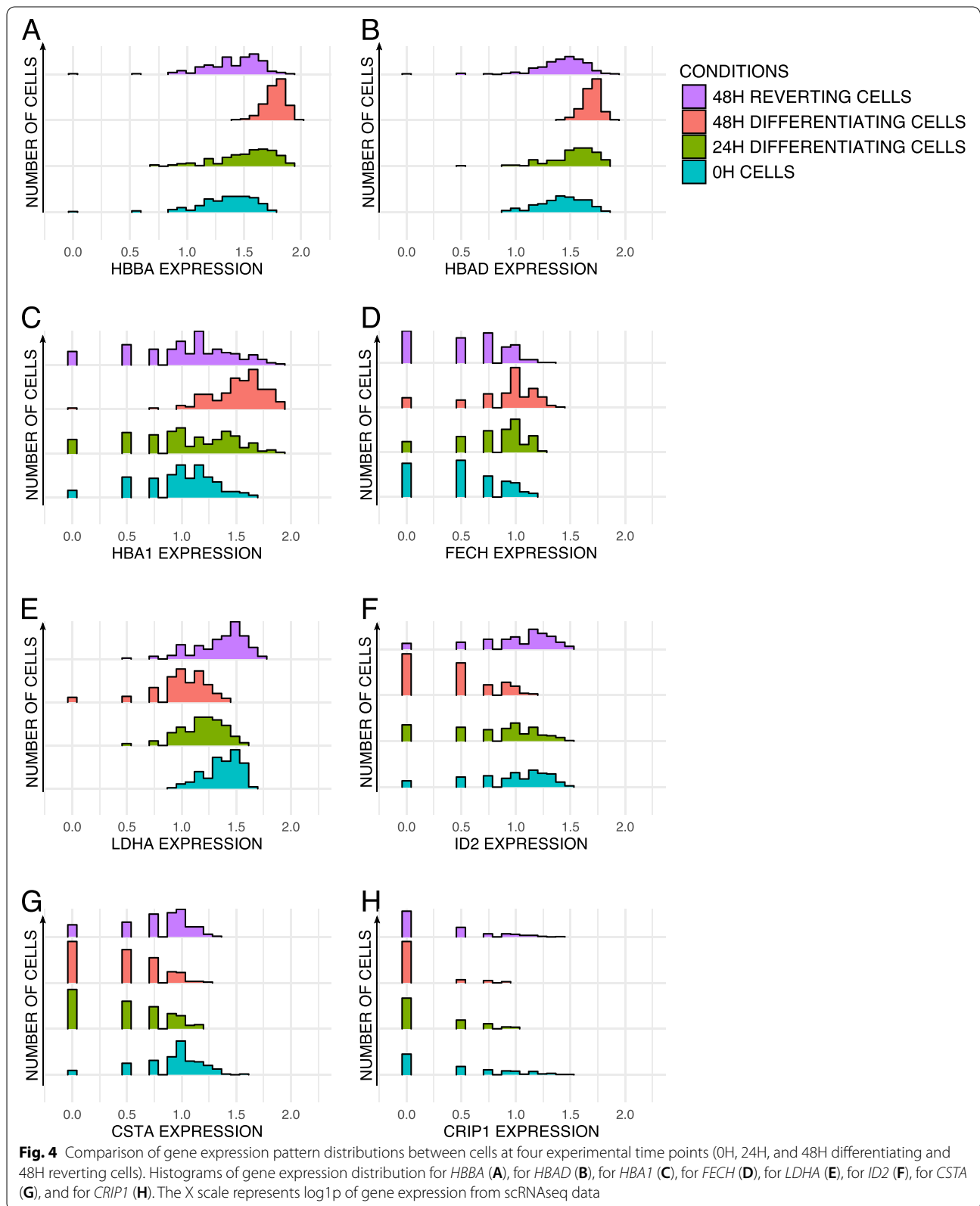
least square [20]) between 48H reverting cells and 0H cells and retained common most discriminating genes between scRT-qPCR and scRNAseq datasets. Interestingly, our results showed that only 3 common genes discriminate between the two cell groups: *HBBA*, *TBC1D7*, and *HSP90AA1*, the expression of which is shown in Fig. 6. *HBBA* is a subunit of the hemoglobin complex which carries oxygen, *TBC1D7* is presumed to have a role in regulating cell growth and differentiation [27], and *HSP90AA1* codes for an isoform of the HSP90 protein chaperone, which its specific transcription is known to be induced in response to insulin [28]. Looking closely, the 48H reverting cells have an intermediate expression level between differentiating cells and undifferentiated cells for the three predicted genes. The offset observed could be due to a longer duration of mRNA half-life at 24H of differentiation. We had previously performed a quantification of mRNA half-life during avian erythrocyte differentiation ([29] Additional file 1: Fig. S3). We focused on mRNA half-life at 24H for those three genes. *TBC1D7* and *HSP90AA1* have a relatively short half-life as opposed to *HBBA*. Other genes analyzed whose expression increases during differentiation, such as *DPP7*, *TPP1*, or *RPL22L1*, have also a long half-life duration mRNA, but only *HBBA* was identified in our statistical analysis as discriminating between undifferentiated and 48H reverting cells.

These results confirmed that the 48H reverting cells display a gene expression pattern very close to those of 0H cells while still retaining traces of their engagement into the differentiation process independently of the mRNA half-life. The molecular process explaining such “lagging genes” will have to be explored.

Cells are distributed as a continuum along the differentiation path

At that stage, two hypotheses could be made: (1) Either all cells have engaged into a differentiation process and do molecularly revert to a self-renewal transcriptional state or (2) at 24H of differentiation two subpopulations coexist: one that is still undifferentiated and would give rise to the 48H reverting cells and a second more differentiated which would lead to the 48H differentiating population and die in the reversion experiment.

We hypothesized that the existence of two subpopulations at 24 h should lead to a higher number of modes in the distribution of some genes at that time point. To test this hypothesis, we therefore estimated for each condition the most-likely number of modes for the probability distribution of each gene, as assessed through a Gamma mixture on scRNAseq (see the “Methods” section). We found no significant difference in the number of modes observed between the 4 populations (Fig. 7), which



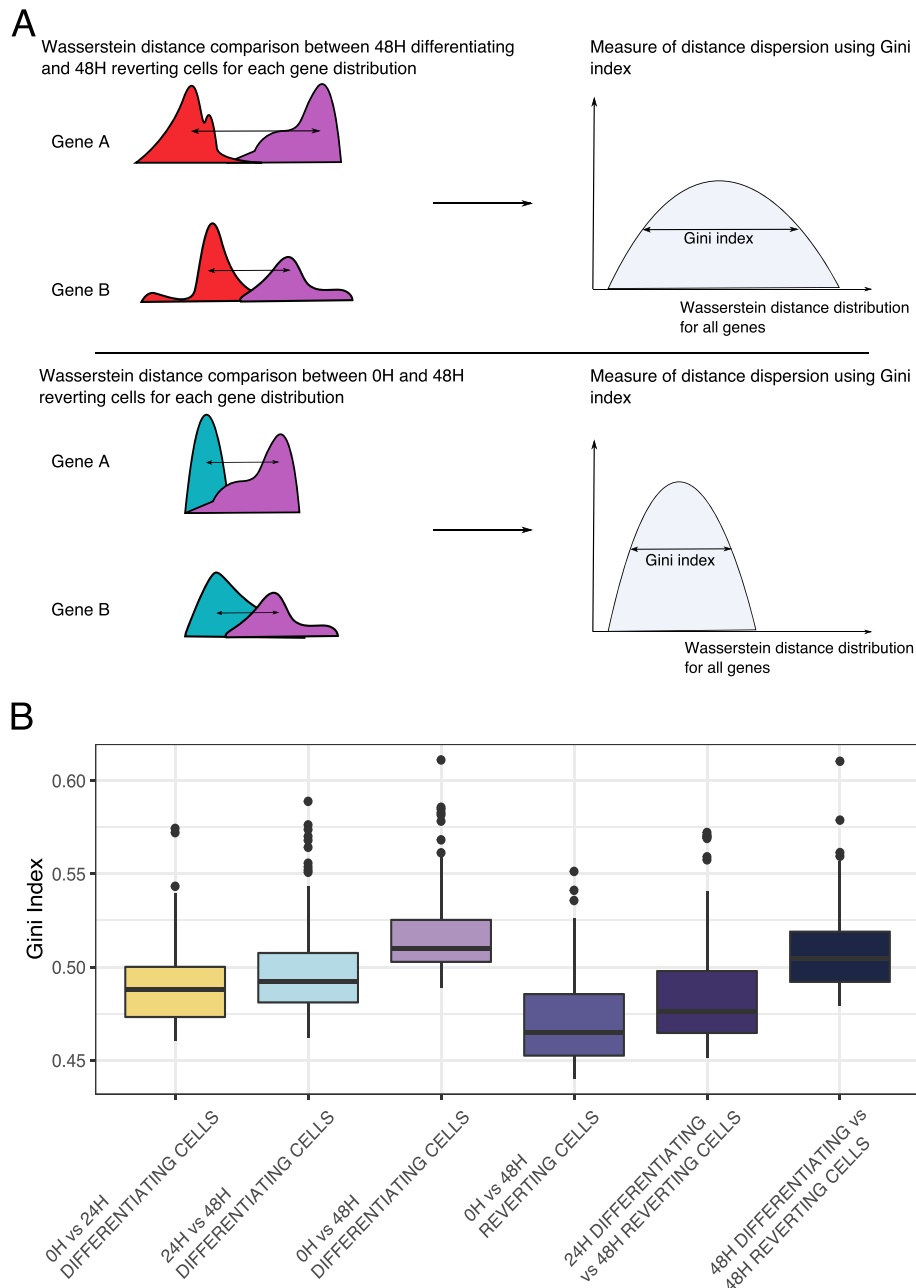
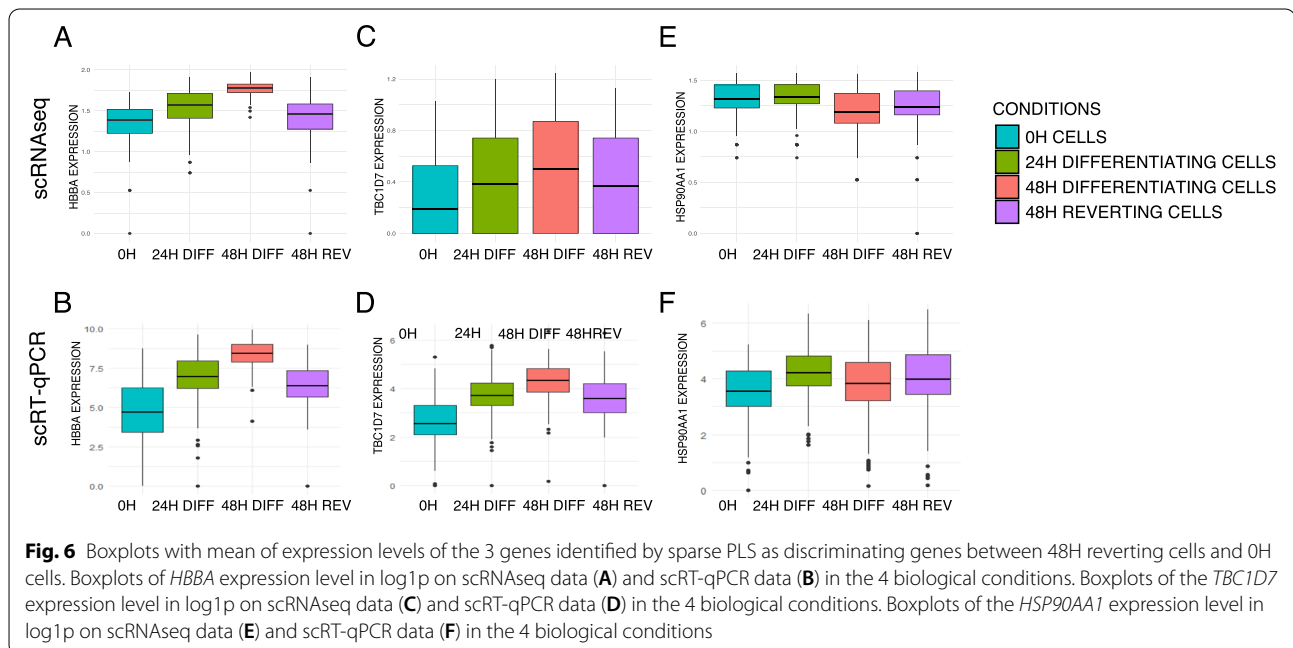


Fig. 5 Comparison of dispersion of gene distribution between cell populations. **A** Experiment design to compare gene distributions between the 4 biological conditions. Wasserstein distance is computed for each gene between pair of conditions, then dispersion of all gene distributions is calculated using Gini index. **B** Plot of Gini index values of Wasserstein distance distributions between conditions in pairs computed for each of the 2000 genes from scRNAseq data bootstrapped 100 times

confirms that the cells collected at 24H do not show more multi-stability than the other groups and are thus unlikely to be a mix of two populations.

The second hypothesis would also imply that in the 24H population, the cells engaged too far in the differentiation process would die a short time after media were

changed, while only the undifferentiated ones would survive. We then measured the viability rate during the kinetics and found no difference in viability between the conditions and especially between the 24H differentiation and the 48H reversion conditions (Additional file 1: Fig. S4).



Finally, the second hypothesis would also imply that the reverting cells are simply cells that have not yet entered the differentiation process. It would therefore be at odds with the evidence that the 48H reverting cells do retain traces of their engagement into the differentiation process (see upper).

Those results strongly suggest that the 24H cell population is not composed of two coexisting subpopulations of cells and that 48H reverting cells enter differentiation before going back to a transcriptomic state close to 0H cells.

Discussion

In the present study, we coupled two different single-cell transcriptomics techniques and state-of-the-art statistical approaches to demonstrate the fate reversibility of avian erythrocyte progenitors induced to differentiate for 24 h.

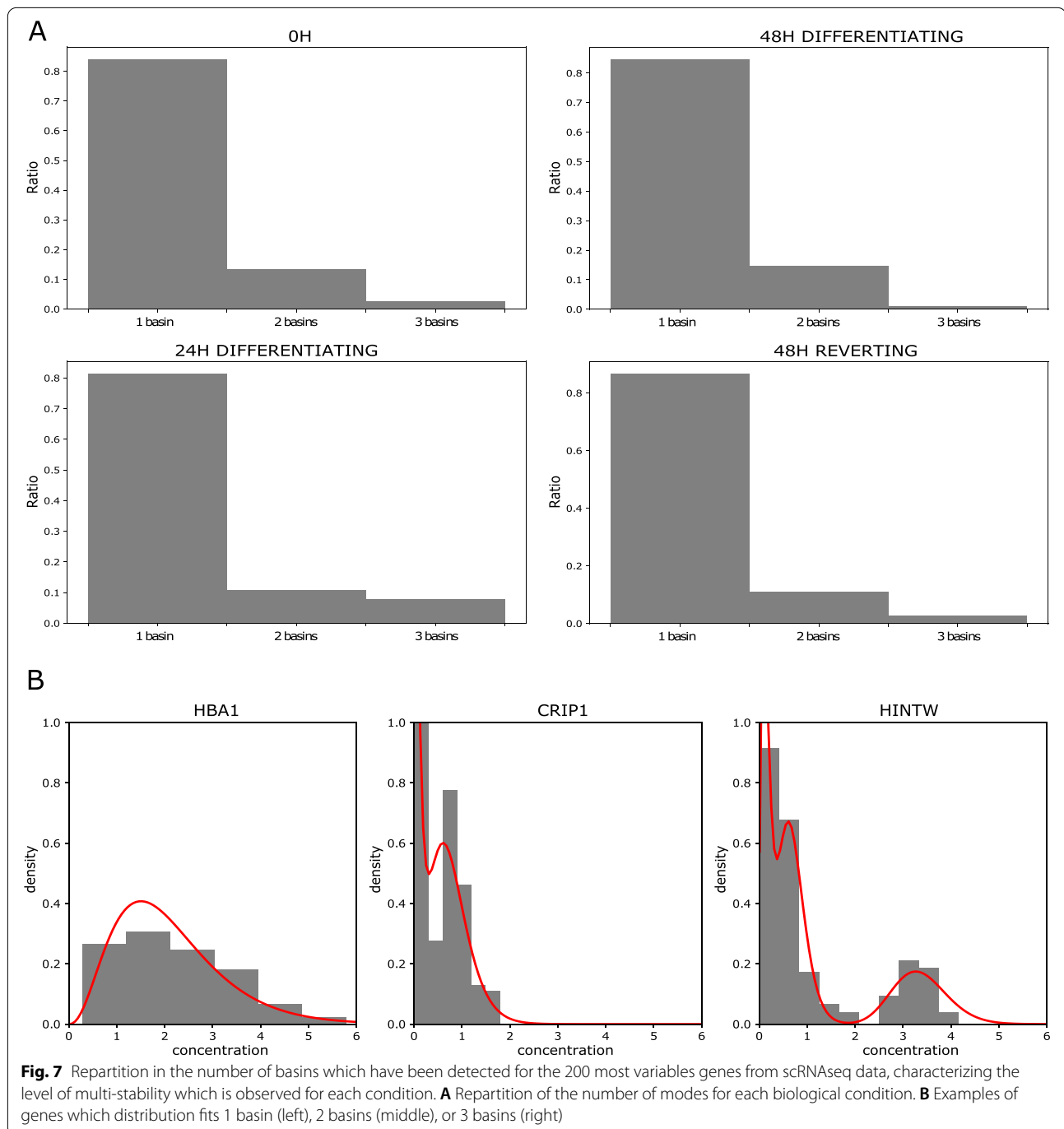
Our results revealed a very close proximity of reverting and undifferentiated cell transcriptomes. Indeed, statistical comparison of cell distributions showed no significant difference between 0H and 48H reverting cells while, as expected, significant changes in gene expression accompanied the differentiation sequence. The analysis of gene expression distribution patterns of the 48H reverting cells confirmed a switch toward the 0H cell gene expression profiles. First, DE analysis of scRNAseq data showed only one gene significantly differentially expressed between the two conditions. Second, Wasserstein distance analysis revealed closer distances between 48H reverting and 0H cells than to any other group of cells. Third, sparse

PLS analysis indicated that the expression level of only three genes, *HBBA*, *TBC1D7*, and *HSP90AA1*, was predictive of the 48H reverting and undifferentiated cells. Interestingly, the persistence of those three genes in 48H reverting cells could not be attributed solely to mRNA half-life duration. However, we cannot exclude that it could be a mere delay and thus a characterization of the reverting cells at a later time point may show a complete molecular reversion.

All of our results therefore favor the hypothesis that a vast majority of the 48H reverting cells responded to differentiation induction by modifying their gene expression profiles but then returned to the self-renewal transcriptional state.

One must note that this would not be the sole example of large-scale transcriptomic changes on (relatively) short time scales [18, 30]. The question as to whether such large-scale transcriptome changes are accompanied, or not, by (reversible) large-scale epigenetic changes remains an open question for future studies.

It has been described in the literature that during cellular decision-making, the cell state is maintained by dynamic interactions between positive and negative regulatory molecules [31] within the frame of a Gene Regulatory Network (GRN). These interactions can be repurposed by changing the stoichiometry of ubiquitous and specific regulatory molecules and factors [11, 13]. In our study, the analysis of gene expression patterns during the reversion process confirmed that the determination of the fate of erythrocyte progenitors is directed by the constraints of the dynamics of the GRN, influenced



by signals emitted by changing conditions of the environment surrounding the cells. In the absence of differentiation signals (or in the presence of self-renewal inducing signals), there is no ratchet in place that would prevent (at least at early stages in our case) the system to return back to its original quasi-steady state. This is in excellent agreement with the previous demonstration that there is a duration threshold for some GRN under which the

system can return back to its original state [32]. This was proposed to allow cells to discriminate between *bona fide* signals and random noise in their environment and could represent a physiological system for finely tuning the in vivo production of red blood cells while preserving the pool of progenitors. We recently proposed a methodology for inferring the GRN underlying T2EC differentiation [29]. For that, we kept in silico cells under constant

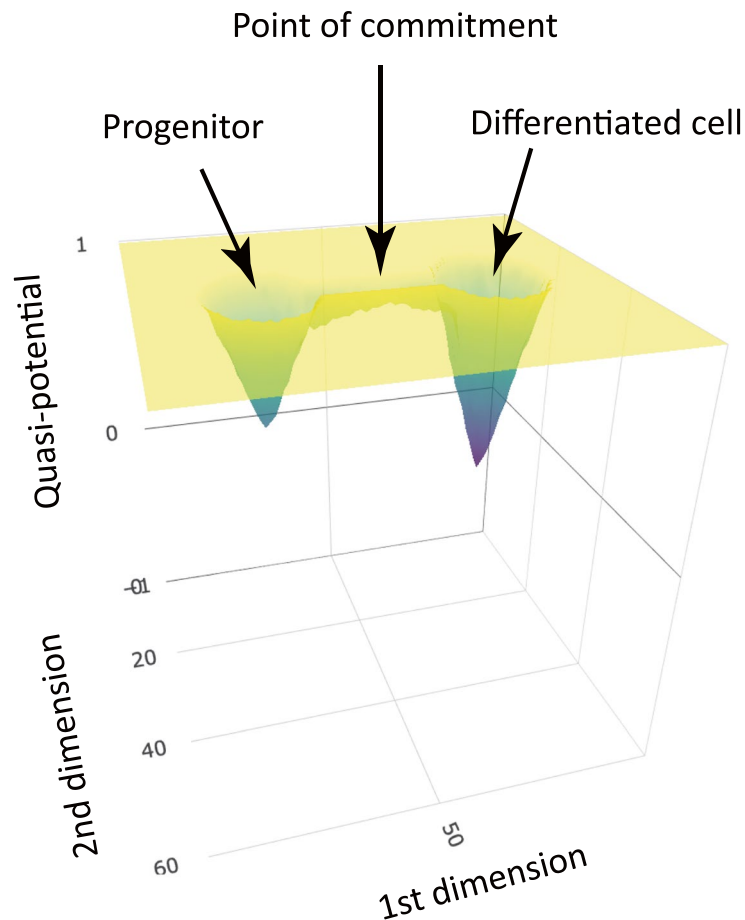


Fig. 8 A quasi-potential well depiction of the erythroid differentiation process. While the cells have not escaped the zone of influence of the progenitor attractor (i.e., when they have not passed the point of commitment, aka the point of no return [19]), the removal of the environmental influences results in their relaxing back to their original attractor state

differentiation stimulus. It would be of interest to see if the inferred GRN would be able to revert, up to a certain point where no “spontaneous” return is possible [19], to its original state. This would be a very strong constraint to impose and should severely limit the number of putative GRN able to reproduce experimental data and thus approaching the most accurate network.

Taken together, our results point towards a physiological plasticity and reversibility with respect to erythrocyte decision-making. It is also reminiscent of the plasticity observed in cancer stem cells that might not be specific to tumor cells [33]. In terms of the epigenetic landscape, our work implies that instead of a continuous gradient that the cells will roll down as in the classical Waddington’s depiction [8], they may go through an unstable state and may, sometimes, roll upwards over a bump in the landscape [34]. Thus, differentiation should be more appropriately described as cells moving from well to well, that is, from one metastable state [35–37] to another one

(Fig. 8). This view abides by the multi-stability framework where a complex quasi-potential landscape aims at describing both normal and pathological differentiation processes [37, 38], and exemplifies the fact that “commitment (is) a dynamical property of the landscape” [39]. It is important at this stage to remember that Waddington himself was aware that his drawing was but a simplification. Adapting and refining this landscape should not be considered as departing from his views. Such a non-monotonous landscape has been proposed to account for the depiction of regeneration in adult tissues [12] and is consistent with previously proposed dynamical principles of cell fate restriction [10]. It is in excellent accordance with the recent depiction that cells can “climb uphill on Waddington’s epigenetic landscape” during cranial neural crest cell development [15] and would also be more relevant to account for the “hesitant” behavior of human CD34+ stem cells in vitro [40] than a straight slope. It is beyond the scope of this discussion to go into more

details, but a cell “climbing uphill” should be seen as equivalent to “the landscape bending into a new valley.”

Conclusions

Our work has provided a detailed molecular characterization of the probabilistic nature of erythrocyte cell fate determination, influenced by the constraints of the underlying Gene Regulatory Network dynamics, and driven by environmental influences.

In conclusion, our results clearly depart from a deterministic view of the differentiation process and fully support the importance of gene expression stochasticity in all systems examined to date [4, 41–44], both in vitro [19, 21, 40, 45–48] and in vivo [49–51].

These new insights into the process of cell reversion could also lead to significant improvements of the executable GRN inference scheme [29].

Methods

Cellular biology

T2EC were extracted from the bone marrow of 19-day-old SPAFAS white leghorn chicken’s embryos (INRA, Tours, France). Cells were grown in self-renewal in a LM1 medium (α -MEM, 10% fetal bovine serum (FBS), 1 mM HEPES, 100 nM β -mercaptoethanol, 100 U/mL penicillin and streptomycin, 5 ng/mL TGF- α , 1 ng/mL TGF- β , and 1 mM dexamethasone) as previously described [52].

Differentiation was induced by removing the LM1 medium and placing the cells into a DM17 medium (α -MEM, 10% fetal bovine serum (FBS), 1 mM Hepes, 100 nM β -mercaptoethanol, 100 U/mL penicillin and streptomycin, 10 ng/mL insulin, and 5% anemic chicken serum (ACS [53])).

Differentiation kinetics were achieved by collecting a sub-fraction of the cells at different times after induction of differentiation (0H and 24H). After 24H, the DM17 medium was removed and half of the cells were placed back into the LM1 medium while the other half was kept in the DM17 medium to achieve 48H reversion and 48H differentiation time points respectively (Fig. 1).

Cell population mortality was assessed by counting dead and living cells from the different time points and conditions after Trypan blue staining and using a Malassez cell.

Single-cell sorting

For both single-cell transcriptomics methods, cells were sorted in 96-well plates using FACS Aria II μ , BD: 8 plates were produced for scRNAseq (2 plates per time point) and 8 plates were produced for scRT-qPCR (2 plates per time point). Since the first steps of library construction are performed per plate, we refer as “batch” the different plates.

Single-cell RT-qPCR analysis

All the manipulations related to the high-throughput scRT-qPCR experiments in microfluidics were performed according to the protocol recommended by the Fluidigm company (PN 68000088 K1, p.157-172). All steps from single-cell isolation to scRT-qPCR, gene selection, data generation, and cleaning are described in detail in [19]. The expression matrix was log_{1p} transformed before subsequent analysis.

Single-cell RNAseq

scRNAseq was performed using an adapted version of the MARSseq protocol [22]. Unless specified, all indicated concentrations correspond to final concentrations.

Individual cells were sorted into 96-well plates containing 4 μ L of lysis buffer and index RT primers (0.2% Triton (Sigma Aldrich), 0.4 U/ μ L RNaseOUT (ThermoFisher Scientific), 400nM RT_primers (Sigma Aldrich)). Index RT_primers (Table 2) contain oligo-dT chain to capture mRNA, a T7 RNA polymerase promoter for further in vitro transcription (IVT), unique cell barcodes for subsequent de-multiplexing, and unique molecular identifiers (UMIs) for PCR bias deduplication. After cell sorting, plates were immediately centrifuged and frozen on dry ice before storage at -80°C until reverse transcription (RT) was performed. The plates were put at 72°C for 3 min for denaturation. A total of 4 μ L of RT mix was added in each well (2mM dNTP (Thermo scientific), 20mM DTT, 2X First stranded buffer, 5 U/ μ L Superscript III RT enzyme (Superscript III RT enzyme kit Thermo scientific), 10% (W/V) PEG 8000 (Sigma Aldrich)). ERCC RNA spike-in (Thermo Scientific) was diluted into the RT mix (dilution 5×10^{-7}). The plates were then transferred into a thermocycler (program: 42°C -2min, 50°C -50min, 85°C -5min, 4°C hold).

Table 2 List and sequences of primers used for scRNAseq library construction

Primer name	5' to 3'
Index_RT_primers (cell BC and UMI)	5'-CGATTGAGGCCGTAATACGACTCACTATAGGG GCGACG TGTGCTCTCCGATCTXXXXXXXXNNNNNNNTTT TTTTTTTTTTTTTTTTTV-3'
P5N6_XXXX (Plate BC)	5'-CTACACGACGCTCTTCCG ATCTXXXXXXXXNNNN-3'
P5.rd1	5'-AATGATACGGCGACCACCGAGATCTCACTCTT TCC CTACACGACGCTCTTCCGATCT-3'
P7.rd2	5'-CAAGCAGAAGACGGCATACGAGAT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'

After reverse transcription, samples were pooled by plate and Exonuclease I (NEB) digestion was performed, followed by 1.2X AMPure beads purification (Beckman Coulter). Samples were eluted in 10mM Tris-HCl, pH=7.5. Second strand cDNA synthesis was performed with 1X SSS buffer and SSS enzyme (NebNext mRNA second strand synthesis kit NEB; thermocycler program: 16°C-150min, 65°C-20min, 4°C hold). Resulting double-strand cDNA were linearly amplified by IVT overnight (10mM ATP, 10mM GTP, 10mM UTP, 10mM GTP, 1X reaction buffer, 1/10 T7 RNA polymerase mix (HighScribe T7 High Yield RNA synthesis NEB)) at 37°C. IVT products were purified with 1.3X Ampure beads and eluted with 10mM Tris-HCl, 0.1mM EDTA. Amplified RNAs were fragmented (1X RNA fragmentation buffer (RNA fragmentation reagents Invitrogen)) at 70°C for 3 min. The fragmentation reaction was stopped with 34µL of STOP mix (0.3X Stop solution (RNA fragmentation reagents Invitrogen), TE buffer 1X (10mM Tris, 1mM EDTA, pH 8 - Invitrogen), and 0.7X AMPure beads to proceed with sample purification). Differing from the original MARSseq protocol, instead of ligation, a second RT was done to incorporate P5N6 primers (Table 2) containing random hexamers and specific barcodes to distinguish the different plates (5mM DTT, 500µM dNTP, 10µM P5N6_XXXX, 1X First stranded buffer, 10U/µL Superscript III RT enzyme, 2U/µL RNaseOUT; thermocycler program: 25°C 5min, 55°C 20min, 70°C 15min, 4°C hold). The cDNAs were then purified with 1.2x AMPure beads. Illumina primers (Table 2) were added by PCR (0.5 µM Mix primer P5.rd1/P7.Rd2, 1X KAPA Hifi HotStart PCR Mix (Kapa Biosystem); thermocycler program: 95°C 3min, 12 times [98°C 20s, 57°C 30s, 72°C 40s], 72°C 5min, 4°C hold), and PCR products were purified with 0.7x AMPure beads and eluted in 15µL.

Libraries were sequenced on a Next500 sequencer (Illumina) with a custom paired-end protocol to avoid a decrease of sequencing quality on read1 due to the high number of T added during polyA reading (130pb on read1 and 20pb on read2). We aimed for a depth of 200,000 raw reads per cell.

Bio-informatic pipeline

Fastq files were pre-processed through a bio-informatic pipeline developed in the team on the Nextflow platform [54]. Briefly, the first step removed Illumina adaptors. The second step de-multiplexed the sequences according to their plate barcodes. Then, all sequences containing at least 4T following cell barcode and UMI were kept. Using UMIttools whitelist, the cell barcodes and UMI were extracted from the reads. The sequences were then mapped on the reference transcriptome

(Gallus GallusGRCG6A.95 from Ensembl) and UMI were counted. Finally, a count matrix was generated for each plate.

Data filtering, normalization, and analysis

All analyses were carried out using R software (version 4.0.5; [55]). Matrixes from the eight plates were pooled together. Cells were filtered based on several criteria: reads number, gene number, count number, and ERCC content. For each criterion, the cutoff values were determined based on SCONE [56] pipeline and were calculated as follows:

$$\text{mean} - 3 * \text{SD}$$

We selected genes present in at least two cells. The filtered matrix was then normalized using SCTransform from the Seurat package [57] and we corrected for batch effect, time effect, and sequencing depth effect. The expression matrix was finally log₁p transformed.

Variable genes were identified using FindVariableFeatures from Seurat, vst method [58]. Based on visualization of gene variance, we retained the 2000 most variable features. Differentially expressed genes were identified using the FindMarkerGenes function from Seurat [58]. Analysis was done by pairwise comparisons between conditions; genes with log fold change ≥ 0.5 and adjusted *p*-value < 0.05 were kept as significant. More information on QC filtering is given in Additional file 1: Fig. S5.

Statistical analysis

All statistical analyses were performed using the R software (version 4.0.5; [55]). Dimensionality reduction and visualization were performed using UMAP [59]. UMAP was performed directly on the 2000 most variable genes (from the scRNAseq dataset) or 83 genes (from the scRT-qPCR dataset) using default parameters. PCA was performed using prcomp function from the stats R package (version 3.6.2). Adaptive sparse PLS for logistic regression was performed using the plsgenomics package [20]. For this analysis, scRT-qPCR data were scaled. Sparse PLS is a supervised statistical analysis that allows to predict the most discriminant variables between two groups.

Wasserstein distance computation was done using the Transport R package [60] and was accomplished for each gene of the scRNAseq dataset.

Gini indexes were calculated using the Ineq R package on Wasserstein distance distributions [61].

Bootstrapping was done using the sample_frac function from the Dplyr R package [62].

Estimation of multi-stability levels

For estimating the level of multi-stability in the data, we considered that the probability distribution of each

gene can be approximated by a Gamma distribution, or a mixture of Gamma distributions, since they are known to describe continuous single-cell data accurately [63]. More precisely, we parameterized the distribution of a gene i by:

$$p_i^m(x) = \sum_{j=1}^m \mu(j) \frac{x^{a_i^j-1} b_i^{d_i^j} e^{-b_i x}}{\Gamma(a_i^j)},$$

where Γ denotes the Gamma function: $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$. Note that only the parameters $(a_i^j)_{j=1, \dots, m}$ depend on the mixture component j : this is related to the distribution arising from the well-established two-state model of gene expression [64], when only the frequency of mRNA bursts is regulated, as described in [65].

For every condition, we constructed 10 training sets consisting of 80% of the cells in the population (randomly chosen), and we estimated the parameters $[(a_i^j)_{j=1, \dots, m}, b_i]$ with a MCMC algorithm for the numbers of mixture components $m = 1, 2, 3$ successively. We then considered that the optimal number of components for gene i was the one which minimized the average BIC score estimated on the 10 corresponding test sets.

Multivariate two-sample test

Samples were compared using a multivariate two-sample test based on the 2000 most variable genes. We suppose that the normalized gene expression X_1 and X_2 of two conditions (0H vs 48H reversion, 0H vs 48H differentiation, 48H reversion vs 48H differentiation) follow a multivariate Gaussian distribution $\mathcal{N}(\mu_1, \Sigma)$ and $\mathcal{N}(\mu_2, \Sigma)$ respectively, and we denote by $n = n_1 + n_2$ the total number of cells. Then, we test the null hypothesis $H_0: \mu_1 = \mu_2$ using the generalized Hotelling's T^2 test [23]. The data being high dimensional ($p > n$), the between-gene pooled covariance matrix is not invertible and is replaced by its Moore-Penrose inverse. In this setting, the asymptotic distribution of the generalized Hotelling statistics is $\chi^2(n-2)$. The p -values were adjusted according to the Benjamini-Hochberg correction [66]. Analysis was performed using the `fdahotelling` R package [67].

Abbreviations

ACS: Anemic chicken serum; DE: Differential expression; GRN: Gene Regulatory Networks; MARSseq: Massively parallel single-cell RNA-Seq; PCA: Principal Component Analysis; scRNAseq: Single-cell RNA sequencing; scRT-qPCR: Single-cell reverse transcription-quantitative polymerase chain reaction; SD: Standard deviation; sparse PLS: Sparse partial least square.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-022-01363-7>.

Additional file 1: Figure S1. [UMAP visualization of scRT-qPCR and scRNAseq data]. **Figure S2.** [PCA of scRT-qPCR and scRNAseq data]. **Figure S3.** [Half-life of mRNA table]. **Figure S4.** [Histograms of viability rate during reversion and differentiation processes]. **Figure S5.** [Summary table of scRNAseq data filtering steps and threshold values for each step].

Acknowledgements

We gratefully thank all members of the SBDM team and particularly Gerard Benoit for very fruitful discussions, suggestions, and commentaries on our project. We also thank Ghislain Durif for his great technical support during PLS computation and G. Yvert for helpful comments about the manuscript. We thank the computational center of IN2P3 (Villeurbanne/France) and Pôle Scientifique de Modélisation Numérique (PSMN, Ecole Normale Supérieure de Lyon) where computations were performed. We acknowledge the contribution of the AniRA-Cytométrie core facility of SFR BioSciences (UAR3444/US8). We thank the BioSyl Federation and the LabEx Ecofect (ANR-11-LABX-0048) of the University of Lyon for inspiring scientific events.

Authors' contributions

SZ and CF contributed equally to the conceptualization of the study, data generation, data analysis, statistical analysis, and interpretation, as well as writing the manuscript. CF participated in the pipeline development. EV1 provided technical support for scRT-qPCR and scRNAseq experiments. LM provided support for data analysis and participated to the pipeline development. RS participated to the pipeline development. AM provided support for the scRNAseq protocol. EV2, MB, AO, and FP performed the statistical analysis. AB participated to the conceptualization of the study. OG performed data analysis and obtained the funding. OG and SG participated to the conceptualization of the study, the project administration, the project supervision, and writing of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by funding from the French agency ANR (SinCity; ANR-17-CE12-0031).

Availability of data and materials

All data generated or analyzed during this study are included in this published article and its supplementary information file and publicly available repositories. Pipelines and analysis scripts are available at https://gitbio.ens-lyon.fr/LBMC/sbdm/mars_seq. Previous scRT-qPCR data are available in the SRA repository <http://www.ncbi.nlm.nih.gov/sra/SRP076011> [19]. Previous data mRNA half-life are available in the OSF repository <https://osf.io/gkedt/> [29]. The datasets supporting the conclusions of this article are available in the NIH repository, accession number PRJNA802343 [68], and in the OSF repository <https://osf.io/upw8d/>.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Author details

¹Laboratory of Biology and Modelling of the Cell, Université de Lyon, Ecole Normale Supérieure de Lyon, CNRS, UMR5239, Université Claude Bernard Lyon 1, Lyon, France. ²Azm Center for Research in Biotechnology and its Applications, LBA3B, EDST, Lebanese University, Tripoli 1300, Lebanon. ³Ecole Pratique des Hautes Etudes, PSL Research University, UMR5951, INSERM, Univ-Evry, Paris, France. ⁴Inria Team Dracula, Inria Center Grenoble Rhone-Alpes,

Grenoble, France. ⁵Institut Camille Jordan, CNRS UMR 5208, Université Claude Bernard Lyon 1, Villeurbanne, France. ⁶Vidium solutions, Lyon, France. ⁷Nantes Université, Centrale Nantes, Laboratoire de mathématiques Jean Leray, LMJL, F-44000 Nantes, France.

Received: 7 March 2022 Accepted: 27 June 2022

Published online: 06 July 2022

References

- Wolpert L. Do we understand development? *Science*. 1994;266:571–2.
- Elowitz MB. Stochastic gene expression in a single cell. *Science*. 2002;297:1183–6.
- Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A. Regulation of noise in the expression of a single gene. *Nat Genet*. 2002;31:69–73.
- Symmons O, Raj A. What's luck got to do with it: single cells, multiple fates, and biological nondeterminism. *Mol Cell*. 2016;62:788–802.
- Kolodziejczyk AA, Lönnberg T. Global and targeted approaches to single-cell transcriptome characterization. *Brief Funct Genomics*. 2018;17:209–19.
- Kalmar T, Lim C, Hayward P, Muñoz-Descalzo S, Nichols J, Garcia-Ojalvo J, et al. Regulated fluctuations in Nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biol*. 2009;7:e1000149.
- Blake WJ, Balázsi G, Kohanski MA, Isaacs FJ, Murphy KF, Kuang Y, et al. Phenotypic consequences of promoter-mediated transcriptional noise. *Mol Cell*. 2006;24:853–65.
- Waddington CH. *The strategy of the genes*. 1st ed: Routledge; 1957.
- Shi J, Teschendorff AE, Chen W, Chen L, Li T. Quantifying Waddington's epigenetic landscape: a comparison of single-cell potency measures. *Brief Bioinform*. 2018;21:248–61.
- Moris N, Pina C, Arias AM. Transition states and cell fate decisions in epigenetic landscapes. *Nat Rev Genet*. 2016;17:693–703.
- Baron MH. Reversibility of the differentiated state in somatic cells. *Curr Opin Cell Biol*. 1993;5:1050–6.
- Rajagopal J, Stanger BZ. Plasticity in the adult: how should the Waddington diagram be applied to regenerating tissues? *Dev Cell*. 2016;36:133–7.
- Johnson NC, Dillard ME, Baluk P, McDonald DM, Harvey NL, Frase SL, et al. Lymphatic endothelial cell identity is reversible and its maintenance requires Prox1 activity. *Genes Dev*. 2008;22:3282–91.
- Ladewig J, Koch P, Brüstle O. Leveling Waddington: the emergence of direct programming and the loss of cell fate hierarchies. *Nat Rev Mol Cell Biol*. 2013;14:225–36.
- Zalc A, Sinha R, Gulati GS, Wesche DJ, Daszczuk P, Swigut T, et al. Reactivation of the pluripotency program precedes formation of the cranial neural crest. *Science*. 2021;371:eabb4776.
- Buczacki SJA, Zecchini HI, Nicholson AM, Russell R, Vermeulen L, Kemp R, et al. Intestinal label-retaining cells are secretory precursors expressing Lgr5. *Nature*. 2013;495:65–9.
- Tata PR, Mou H, Pardo-Saganta A, Zhao R, Prabhu M, Law BM, et al. Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature*. 2013;503:218–23.
- Chang HH, Hemberg M, Barahona M, Ingber DE, Huang S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*. 2008;453:544–7.
- Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, et al. Single-cell-based analysis highlights a surge in cell-to-cell molecular variability preceding irreversible commitment in a differentiation process. *PLoS Biol*. 2016;14:e1002585 <http://www.ncbi.nlm.nih.gov/sra/SRP076011>.
- Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. *Stat Appl Genet Mol Biol*. 2008;7(1).
- Mojtahedi M, Skupin A, Zhou J, Castaño IG, Leong-Quong RYY, Chang H, et al. Cell fate decision as high-dimensional critical state transition. *PLoS Biol*. 2016;14:e2000640.
- Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretzky I, et al. Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science*. 2014;343:776–9.
- Secchi P, Stamm A, Vantini S. Inference for the mean of large p small n data: a finite-sample high-dimensional generalization of Hotelling's theorem. *Electron J Stat*. 2013;7:2005–31.
- Richard A, Vallin E, Romestaing C, Roussel D, Gandrillon O, Gonin-Giraud S. Erythroid differentiation displays a peak of energy consumption concomitant with glycolytic metabolism rearrangements. *PLoS One*. 2019;14:e0221472.
- Yokota Y, Mori S, Narumi O, Kitajima K. In vivo function of a differentiation inhibitor, Id2. *IUBMB Life*. 2001;51:207–14.
- Klenova EM, Botezato I, Laudet V, Goodwin GH, Wallace JC, Lobanenkov VV. Isolation of a cDNA clone encoding the RNASE-superfamily-related gene highly expressed in chicken bone marrow cells. *Biochem Biophys Res Commun*. 1992;185:231–9.
- Dibble CC, Elis W, Menon S, Qin W, Klekota J, Asara JM, et al. TBC1D7 is a third subunit of the TSC1-TSC2 complex upstream of mTORC1. *Mol Cell*. 2012;47:535–46.
- Zuehlke AD, Beebe K, Neckers L, Prince T. Regulation and function of the human HSP90AA1 gene. *Gene*. 2015;570:8–16.
- Bonnafox A, Herbach U, Richard A, Guillemain A, Gonin-Giraud S, Gros P-A, et al. WASABI: a dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics*. 2019;20:220 <https://osf.io/gkedt/>.
- Nichols JM, Antolović V, Reich JD, Brameyer S, Paschke P, Chubb JR. Cell and molecular transitions during efficient dedifferentiation. *eLife*. 2020;9:e55435.
- Blau HM. Differentiation requires continuous active control. *Annu Rev Biochem*. 1992;61:1213–30.
- Sokolik C, Liu Y, Bauer D, McPherson J, Broecker M, Heimberg G, et al. Transcription factor competition allows embryonic stem cells to distinguish authentic signals from noise. *Cell Syst*. 2015;1:117–29.
- Thankamony AP, Saxena K, Murali R, Jolly MK, Nair R. Cancer stem cell plasticity – a deadly deal. *Front Mol Biosci*. 2020;7:79.
- Kimmel JC, Yi N, Roy M, Hendrickson DG, Kelley DR. Differentiation reveals latent features of aging and an energy barrier in murine myogenesis. *Cell Rep*. 2021;35:109046.
- Guillemain A, Stumpf MPH. Noise and the molecular processes underlying cell fate decision-making. *Phys Biol*. 2021;18:011002.
- Pisco AO, Fouquier d'Hérouël A, Huang S. Conceptual confusion: the case of epigenetics. preprint. bioRxiv. 2016:053009. <https://doi.org/10.1101/053009>.
- Ventre E, Espinasse T, Bréhier C-E, Calvez V, Lepoutre T, Gandrillon O. Reduction of a stochastic model of gene expression: Lagrangian dynamics gives access to basins of attraction as cell types and metastability. *J Math Biol*. 2021;83:59.
- Huang S. Reprogramming cell fates: reconciling rarity with robustness. *BioEssays*. 2009;31:546–60.
- Sáez M, Blassberg R, Camacho-Aguilar E, Siggia ED, Rand DA, Briscoe J. Statistically derived geometrical landscapes capture principles of decision-making dynamics during cell fate transitions. *Cell Syst*. 2022;13:12–28.e3.
- Moussy A, Cossette J, Parmentier R, da Silva C, Corre G, Richard A, et al. Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment. *PLoS Biol*. 2017;15:e2001867.
- Smith S, Grima R. Single-cell variability in multicellular organisms. *Nat Commun*. 2018;9:345.
- Eling N, Morgan MD, Marioni JC. Challenges in measuring and understanding biological noise. *Nat Rev Genet*. 2019;20:536–48.
- Losick R, Desplan C. Stochasticity and cell fate. *Science*. 2008;320:65–8.
- Paldi A. Stochastic or deterministic? That is the Question. *Org J Biol Sci*. 2020;4:77–9.
- Guillemain A, Duchesne R, Crauste F, Gonin-Giraud S, Gandrillon O. Drugs modulating stochastic gene expression affect the erythroid differentiation process. *PLoS One*. 2019;14:e0225166.
- Moris N, Edri S, Seyres D, Kulkarni R, Domingues AF, Balayo T, et al. Histone acetyltransferase KAT2A stabilizes pluripotency with control of transcriptional heterogeneity. *Stem Cells Dayt Ohio*. 2018;36:1828–38.
- Stumpf PS, Smith RCG, Lenz M, Schuppert A, Müller F-J, Babbie A, et al. Stem cell differentiation as a non-Markov stochastic process. *Cell Syst*. 2017;5:268–282.e7.

48. Stockholm D, Edom-Vovard F, Coutant S, Sanatine P, Yamagata Y, Corre G, et al. Bistable cell fate specification as a result of stochastic fluctuations and collective spatial cell behaviour. *PLoS One*. 2010;5:e14441.
49. Wernet MF, Mazzoni EO, Çelik A, Duncan DM, Duncan I, Desplan C. Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature*. 2006;440:174–80.
50. Dussiau C, Boussaroque A, Gaillard M, Bravetti C, Zaroili L, Knosp C, et al. Hematopoietic differentiation is characterized by a transient peak of entropy at a single-cell level. *BMC Biol*. 2022;20:60.
51. Toh K, Saunders D, Verd B, Stevenon B. Zebrafish neuromesodermal progenitors undergo a critical state transition *in vivo*. *bioRxiv*. 2022.02.25.481986. <https://doi.org/10.1101/2022.02.25.481986>.
52. Gandrillon O, Schmidt U, Beug H, Samarut J. TGF- β cooperates with TGF- α to induce the self-renewal of normal erythrocytic progenitors: evidence for an autocrine mechanism. *EMBO J*. 1999;18:2764–81.
53. Gandrillon O, Samarut J. Role of the different RAR isoforms in controlling the erythrocytic differentiation sequence. Interference with the v-erbA and p135gag-myb-ets nuclear oncogenes. *Oncogene*. 1998;16:563–74.
54. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. *Nat Biotechnol*. 2017;35:316–9.
55. R Core Team. R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing; 2021.
56. Cole MB, Risso D, Wagner A, DeTomaso D, Ngai J, Purdom E, et al. Performance assessment and selection of normalization procedures for single-cell RNA-Seq. *Cell Syst*. 2019;8:315–328.e8.
57. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*. 2019;20:296.
58. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single cell data. *Cell*. 2019;177(7):1888–1902.e21.
59. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. 2018;37:38–44.
60. Schuhmacher D, Bähre B, Gottschlich C, Hartmann V, Heinemann F, Schmitzer B. transport: computation of optimal transport plans and Wasserstein distances. R package version 0.12-2. 2020. <https://cran.r-project.org/package=transport>.
61. Zeileis A. ineq: measuring inequality, concentration, and poverty; 2014.
62. Hadley Wickham, Romain François, Lionel Henry, Kirill Müller. dplyr: a grammar of data manipulation. 2021.
63. Albayrak C, Jordi CA, Zechner C, Lin J, Bichsel CA, Khammash M, et al. Digital quantification of proteins and mRNA in single mammalian cells. *Mol Cell*. 2016;61:914–24.
64. Peccoud J, Ycart B. Markovian modeling of gene-product synthesis. *Theor Popul Biol*. 1995;48(2):222–34.
65. Ventre E. Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics. *In Silico Biol*. 2021;14(3–4):89–113.
66. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Methodol*. 1995;57:289–300.
67. Stamm A, Pini A, Vantini S. fdahotelling: inference for functional data analysis in R. R (>= 3.1.3). <https://github.com/astamm/fdahotelling>.
68. Molecular characterization of pre-commitment cell reversion during erythroid differentiation. NCBI BioProject accession: PRJNA802343. <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA802343>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



Chapitre 5

Discussion

L'intégration des signaux endogènes et exogènes et la réponse adaptée qui en résulte permettent aux cellules de prendre les décisions nécessaires et indispensables aux fonctions et à la survie des organismes. Parmi ces processus de décision, la différenciation est l'acquisition par une cellule d'une fonction spécifique associée à un type cellulaire. Les études réalisées en cellules uniques ont permis de mettre en évidence qu'à cette échelle, les processus de différenciation sont des processus stochastiques régis principalement par les fluctuations aléatoires des différents produits régulant l'expression génique [25, 34]. Dans de nombreux systèmes biologiques *in vitro* [162, 165] et *in vivo* [137, 202], il a aussi été montré que la différenciation s'accompagne d'une augmentation transitoire de la variabilité de l'expression génique. Cette augmentation transitoire a notamment été démontrée lors de la différenciation des progéniteurs érythrocytaires aviaires T2EC [160]. Cependant, la façon dont cette variabilité d'expression génique est propagée au cours des générations cellulaires dans une population de cellules isogéniques est encore peu investiguée. D'autant plus que plusieurs études récentes montrent qu'il existe une mémoire transcriptionnelle caractérisée par la transmission clonale de niveaux d'expression de certains gènes, dont l'expression est autrement variable dans la population. Ces études ont investigué la mémoire transcriptionnelle de différentes manières : à l'échelle d'un petit nombre de gènes rapporteurs en cellules uniques [175] ; à l'échelle du transcriptome en cellules uniques à l'aide d'une puce microfluidique [198] ; et à l'échelle du transcriptome sur de petits clones de cellules [172]. Néanmoins ces études n'adressent pas directement la question du devenir de la mémoire transcriptionnelle lors de la différenciation cellulaire. Or, il semble que la transmission d'états transcriptionnels actifs soit essentielle au maintien des programmes de différenciation.

L'enjeu de ma thèse a donc été d'étudier comment les cellules concilient les contraintes de la mémoire transcriptionnelle et l'augmentation de la variabilité d'expression génique au cours du processus de différenciation.

Dans un premier temps, je me suis intéressée au lien entre la variabilité de l'expression génique et les relations de parenté de progéniteurs érythrocytaires aviaires (T2EC) qui, soit s'auto-renouvellent, soit se différencient, et après une ou deux divisions cellulaires.

Dans un deuxième temps et suite à une étude réalisée précédemment dans l'équipe, j'ai caractérisé le transcriptome de cellules en différenciation replacées dans du milieu d'auto-renouvellement, dans l'objectif de confronter la plasticité phénotypique des cellules et leur plasticité moléculaire.

5.1 Mémoire transcriptionnelle : transmission des niveaux d'ARNm lors de la différenciation érythrocytaire

Dans la première partie de mon travail nous nous sommes questionnés sur comment les contraintes de la mémoire transcriptionnelle s'intègrent avec l'augmentation de la variabilité d'expression génique nécessaire au processus de différenciation.

Pour répondre à cette question, nous avons besoin de comparer les transcriptomes de cellules après une et deux divisions cellulaires par sc-RNA-seq, dans un contexte d'auto-renouvellement et dans un contexte de différenciation.

Nous avons donc dû développer plusieurs techniques originales permettant de récupérer le transcriptome de cellules généalogiquement apparentées en conservant toujours l'information de leurs liens de parenté.

Les résultats de la comparaison des transcriptomes à l'aide des distances de Manhattan entre des cellules soeurs de deux types cellulaires (HSC humaines et progéniteurs érythrocytaires aviaires) et dans deux conditions biologiques (auto-renouvellement et différenciation), ont montré qu'il existe une mémoire transcriptionnelle entre des cellules soeurs et que cette mémoire est présente aussi bien pendant l'auto-renouvellement que pendant la différenciation. Enfin, nous avons identifié que la mémoire est portée par un sous-groupe de gènes, les gènes « mémoire » que nous avons mis en évidence à l'aide d'un modèle statistique à effets mixtes. Ces gènes ont une abondance relative plutôt élevée et sont majoritairement impliqués dans la fonction érythrocytaire.

De manière intéressante, nos résultats sur les cellules soeurs suggèrent que la mémoire transcriptionnelle est moins maintenue au cours de la différenciation que lors de l'auto-renouvellement. L'analyse des transcriptomes de cellules cousines, soit après deux divisions cellulaires (48h en auto-renouvellement ou en différenciation) nous a permis de montrer que la mémoire transcriptionnelle est effacée graduellement au cours des divisions cellulaires et plus rapidement pendant la différenciation que pendant l'auto-renouvellement (Figure 5.1).

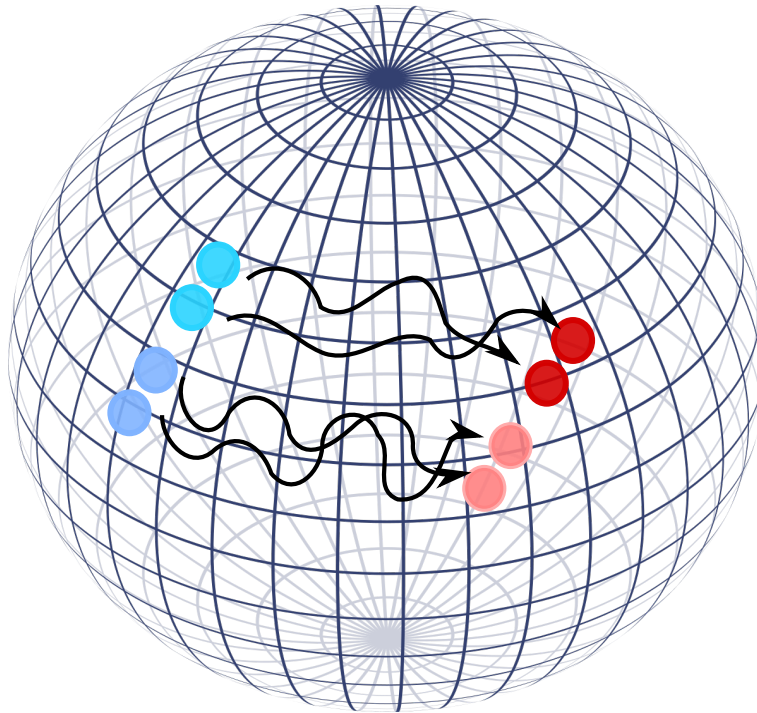


FIGURE 5.1 – Schéma représentant l'effacement progressif de la mémoire transcriptionnelle au cours de la différenciation érythrocytaire.

La sphère représente l'espace d'expression des gènes où les cellules bleues en état d'auto-renouvellement vont recevoir le signal de différenciation. Durant ce processus, les contraintes de la mémoire transcriptionnelle pousse les cellules soeurs à commencer par emprunter des chemins similaires puis l'augmentation de la SEG fait que les cellules apparentés s'éloignent et empruntent des routes différentes jusqu'à atteindre l'état différencié en rouge.

Pour aller plus loin, nos travaux ont permis de montrer qu'il serait aussi possible d'utiliser la puce microfluidique développée dans le chapitre 3 dans un contexte multigénérationnel, grâce à la relocalisation des cellules filles à chaque division dans de nouvelles chambres. Il sera également possible de changer, au cours de la culture des cellules, le milieu de culture pour induire les cellules à se différencier.

On peut alors se demander quels sont les mécanismes qui sous-tendent la mémoire transcrip-

tionnelle. Il a d'abord été proposé qu'il s'agit majoritairement de mécanismes passifs tels que la partition égale pendant la mitose des facteurs régulant la dynamique de la transcription. Dans ce cas la mémoire serait majoritairement permise par la durée de demi-vie longue des molécules régulant la transcription (principalement les protéines mais aussi les ARNm).

Cependant, ces mécanismes passifs ne semblent pas pouvoir expliquer à eux seuls la mémoire. En effet, dans nos données l'intensité de corrélation des gènes mémoire ne semblent pas proportionnelle à la durée de la demi-vie des ARNm.

La mémoire pourrait en fait être transmise par une combinaison de mécanismes passifs comme ceux cités ci-dessus mais aussi de mécanismes actifs probablement d'origine épigénétique. Très peu de travaux démontrent des relations causales entre profils épigénétiques et mémoire transcriptionnelle en cellule unique ; néanmoins, certaines études montrent des relations de corrélation impliquant des modifications épigénétiques qui contribueraient à maintenir l'identité moléculaire des cellules au travers des divisions cellulaires.

Dans une étude réalisée chez *Dictyostelium*, les auteurs ont montré que la fréquence de transcription du gène *Act5* est héritée dans les cellules filles et que cette hérédité passe par la méthylation du résidu lysine 4 de l'histone 3 (H3K4), cette marque étant elle-même transmise au cours des divisions cellulaires. Cette marque est généralement associée à une transcription active [173]. Dans des cellules de mélanomes, Shaffer *et al.* ont observé un appauvrissement des marques d'acétylation de H3K27 et un gain des marques de triméthylation de H3K27 sur les gènes mémoires identifiés par Memory-seq par rapport à des gènes contrôles choisis aléatoirement [172]. De plus, il a été montré que les régions promotrices de certains gènes conservent des marques épigénétiques, notamment di et triméthylation de H3K4 ainsi qu'une tri-méthylation de H3K27 pendant la mitose [203].

Sur la base de ces observations, nous pourrions regarder dans nos cellules soeurs et cousines si l'environnement chromatinien des gènes mémoires est également marqué de cette manière par rapport à des gènes non mémoires. Puis, en fonction des marques impliquées, nous pourrions utiliser des drogues connues pour impacter des enzymes participant au dépôt ou au maintien de ces marques. Par exemple dans le cas de l'acétylation/dé-acétylation des résidus d'histones, la drogue MB3, déjà utilisée dans l'équipe et décrite pour inhiber l'histone acétyl-transférase KAT2A. Le traitement des T2EC en différenciation avec cette drogue avait induit une augmen-

tation de la variabilité de l'expression génique et avait modifié la dynamique de différenciation des cellules [161]. Une autre drogue que nous pourrions utiliser, la Trichostatine A, inhibe elle les HDAC de classe I et II. L'analyse de l'effet de ces traitements sur la mémoire transcriptionnelle dans les deux conditions biologiques permettrait d'explorer plus amplement la piste du rôle des marques épigénétiques dans l'environnement chromatinien des gènes mémoires.

Les mécanismes de mémoire pourraient aussi impliquer des facteurs de transcription particuliers comme cela semble être le cas dans l'étude de Shaffer *et al.*. Les auteurs ont identifié par ATAC-seq un enrichissement pour les motifs correspondants à 6 facteurs de transcription spécifiques dans les régions accessibles des gènes mémoires comparé aux régions accessibles entourant des gènes contrôles choisis aléatoirement. De plus, le knockout de 4 de ces facteurs de transcription a montré de forts effets sur les niveaux de transcription des gènes mémoires, indiquant que la régulation de l'expression de ces gènes mémoires implique très probablement ces facteurs de transcription [172].

Ces facteurs de transcription pourraient être impliqués dans le « bookmarking mitotique ». Le « bookmarking mitotique » ou marquage mitotique fait référence à la rétention de facteurs impliqués dans la régulation des gènes, y compris les facteurs de transcription, sur la chromatine mitotique. Ce système de rétention de régulateurs permet ensuite la réactivation rapide et précise de l'expression des gènes marqués dans les cellules filles et favorise des profils d'expression similaires entre elles [204].

Les facteurs de transcription potentiellement impliqués dans ce marquage mitotique pourraient tout à fait être types cellulaires spécifiques. Dans nos cellules nous pourrions par exemple cibler le facteur de transcription GATA1. Premièrement parce que GATA1 est un facteur de transcription essentiel à la différenciation érythrocytaire normale [205], deuxièmement car il a été montré comme pouvant être impliqué dans du marquage mitotique [206], et troisièmement car 66% de nos gènes mémoires sont des cibles potentielles de GATA1 (en se référant à la base de données ENCODE).

Cependant, la manière dont les facteurs de transcription restent associés à la chromatine pendant la mitose n'est pas encore bien comprise. Une étude très récente suggère que le marquage mitotique pourrait impliquer des interactions entre les facteurs de transcription et des motifs particuliers de résidus d'histones, ce qui permettrait de maintenir certains états chromatiniens

à travers la mitose [207]. En particulier, il a été montré que dans des lignées de cellules érythroïdes, la déplétion de GATA-1 réduit l'acétylation de H3K27 au niveau des sites de fixation du facteur de transcription CTCF dans les régions des enhanceurs des gènes spécifiques des cellules érythroïdes [208].

De plus, il a été montré, grâce à la méthode EU-RNA-seq (5-ethynyluridine RNA-seq) qui permet de quantifier spécifiquement la transcription pendant la mitose, que la transcription à bas bruit de certains gènes était maintenue pendant la mitose [209]. Ces études remettent en question le modèle de « silence transcriptomique » durant la mitose. Un modèle explicatif simple a été proposé : pendant la mitose, le maintien de la chromatine ouverte au niveau des promoteurs ainsi que la transcription active de certains gènes, même à un faible niveau, permettent une réexpression rapide et robuste des gènes à la sortie de la mitose, fonctionnant ainsi comme un mécanisme épigénétique qui garantit la propagation de la mémoire transcriptionnelle à travers la mitose [210]. Il est également possible que la transcription mitotique passe par de la rétention de facteurs de transcription comme on vient de le voir. Ainsi nous pourrions définir par EU-RNA-seq si les gènes mémoires identifiés dans les T2EC sont transcriptionnellement actifs pendant la mitose et mèneraient à des niveaux d'expression corrélés entre cellules soeurs comme observé.

La mémoire transcriptionnelle pourrait donc avoir un rôle dans la formation des tissus au cours des processus du développement et de l'homéostasie adulte. On sait que la SEG est à l'origine de changements dans les profils d'expression des gènes entre les cellules, changements qui seraient alors renforcés et stabilisés clonalement par la mémoire transcriptionnelle. La mémoire transcriptionnelle pourrait ainsi être le mécanisme par lequel les cellules conserveraient l'information, après la mitose, de l'état transcriptionnel dans lequel elles se trouvaient avant leur division. Cette mémoire serait donc l'étape intermédiaire qui réconcilie la robustesse des processus développementaux et la variabilité de l'expression génique observée au sein des populations cellulaires.

5.2 Mémoire transcriptionnelle : plasticité moléculaire lors de la différenciation érythrocytaire

Une étude précédemment réalisée dans l'équipe avait mis en évidence un pic de variabilité, quantifié par l'entropie de Shannon, entre 8h et 24h après induction de la différenciation des progéniteurs érythrocytaires. Dans cette étude, l'équipe avait aussi observé que ces cellules induites à se différencier pendant 24h puis, replacées dans le milieu d'auto-renouvellement ré-acquéraient la capacité de proliférer de la même manière que des cellules indifférenciées, ce qui n'était pas le cas après 48h de différenciation [160]. Cette observation avait ouvert la question de l'état transcriptomique de ces cellules induites transitoirement à se différencier et capables de revenir en arrière en tout cas d'un point de vue prolifératif.

La deuxième partie de mon travail de thèse a donc consisté à déterminer si la réversion phénotypique des progéniteurs était également accompagnée d'une réversion moléculaire.

Dans l'ensemble, nos résultats ont montré que la très grande majorité des cellules retournaient à un état transcriptomique très proche de celui de cellules indifférenciées. Nos observations favorisent un modèle plastique de la différenciation plutôt qu'un modèle linéaire et stéréotypé ; en effet, la différenciation est presque complètement réversible à l'échelle moléculaire tant que les cellules n'ont pas passé le point d'engagement au-delà duquel l'engagement est irréversible.

Cette capacité de réversion moléculaire pourrait permettre aux cellules de discriminer entre des fluctuations aléatoires de signaux liées à un environnement dynamique et un vrai signal de différenciation. Comme tous processus cellulaires, la différenciation érythrocytaire est contrainte par la dynamique du GRN sous-jacents. Ainsi, il a été montré qu'il existe un seuil pour certains GRN. Tant que le seuil n'est pas franchit, et malgré les fluctuations des niveaux d'expression des gènes influencés par l'environnement, le système biologique peut revenir à son état d'origine [211]. Notre équipe a précédemment développé une méthode d'inférence de réseaux qui a été appliquée à la différenciation érythrocytaire aviaire normale générant plusieurs centaines de réseaux candidats [118]. Il serait donc très intéressant d'imposer cette contrainte de la réversion afin d'affiner le nombre de réseaux candidats à l'aide cet outil.

Cependant, dans les cellules en réversion, nous avons observé que les niveaux d'expression de

certaines gènes n'étaient pas identiques aux niveaux d'expression de cellules indifférenciées. En effet, les cellules en réversion présentent pour ces gènes des niveaux d'expression intermédiaires entre des cellules indifférenciées et des cellules en cours de différenciation. Cette observation peut suggérer deux choses : 1) Bien que la durée des demi-vie des ARNm ne semble pas être la cause d'un possible délai, ces gènes pourraient avoir des temps de turn-over plus long (impliquant une stabilité plus élevée ou un taux de dégradation plus faible en moyenne que pour d'autres gènes) ; les cellules, observées à des temps plus tardifs post-réversion, pourraient alors présenter des niveaux d'expression pour ces gènes identiques à des cellules indifférenciées. 2) Si à un temps plus long, les cellules en réversion présentent toujours des niveaux d'expression différents pour ces gènes « retardés » comparés aux cellules indifférenciées, cela pourrait suggérer une forme de mémoire de l'engagement des cellules dans la différenciation.

Il serait donc pertinent de questionner également si le phénomène de réversion moléculaire et phénotypique s'accompagne de changements épigénétiques plus ou moins pérennes. Ces changements épigénétiques pourraient être médiés : 1) par un maintien de l'accessibilité de la chromatine au niveau des régions régulatrices de certains gènes, et notamment ceux pour lesquels une différence dans le niveau d'expression par rapport aux cellules indifférenciées à été observée. L'accessibilité des régions régulatrices de ces gènes pourrait être évaluée par ATAC-seq en cellule unique ; 2) par marquage des régions régulatrices des gènes, particulièrement des enhanceurs, comme vu dans l'introduction bibliographique. Ce marquage des régions régulatrices pourrait passer par des modifications de résidus d'histones spécifiques dans l'environnement chromatinien de ces gènes. En faisant l'hypothèse préalable des marques qui seraient impliquées dans le maintien, il serait possible d'évaluer leur présence dans les régions régulatrices des gènes identifiés plus haut par « CUT and TAG » en cellule unique [212], méthode dérivant du ChIP-seq et est adaptée à la cellule unique.

De plus, comme il a été montré qu'à 24h, la différenciation érythrocytaire s'accompagne d'un pic de variabilité, il serait particulièrement intéressant de déterminer si la population de cellules en réversion conserve de hauts niveaux de variabilité et donc d'entropie ou bien si les cellules en réversion retrouvent un niveau d'entropie comparable à celui de cellules indifférenciées.

Des niveaux d'entropie différents de ceux de cellules indifférenciées pourrait suggérer que la réponse à un nouveau signal de différenciation des cellules en réversion pourrait suivre une

cinétique différente.

L'ensemble de ces questions pourraient permettre de déterminer si un « priming » peut avoir lieu lors de la différenciation érythrocytaire comme c'est le cas dans d'autres systèmes biologiques [182, 213, 214]. Ce « priming » pourrait favoriser une différenciation plus rapide des cellules en réversion, en permettant une réponse rapide en présence d'un nouveau signal de différenciation. Mais il est aussi envisageable que ce « priming » empêche les cellules, peut être partiellement, de se re-différencier. Cela pourrait permettre de favoriser le compartiment cellulaire « progéniteurs » plutôt que celui des érythrocytes matures, et maintenir leur présence dans la moelle osseuse. Finalement, la plasticité moléculaire des progéniteurs érythrocytaires pourrait représenter un phénomène physiologique permettant de réguler avec précision la production *in vivo* de globules rouges tout en préservant le pool de progéniteurs.

Bibliographie

1. HAMBURGER, V. & LEVI-MONTALCINI, R. Proliferation, differentiation and degeneration in the spinal ganglia of the chick embryo under normal and experimental conditions. en. *Journal of Experimental Zoology* **111**. _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/jez.1401457-501>. ISSN : 1097-010X (1949).
2. AHLGREN, S. en. in *Methods in Cell Biology* 153-165 (Elsevier, 2008). ISBN : 978-0-12-564174-6.
3. O'FARRELL, P. H., EDGAR, B. A., LAKICH, D. & LEHNER, C. F. Directing Cell Division During Development. *Science* **246**. Publisher : American Association for the Advancement of Science, 635-640 (nov. 1989).
4. CLEVERS, H. The intestinal crypt, a prototype stem cell compartment. eng. *Cell* **154**, 274-284. ISSN : 1097-4172 (juill. 2013).
5. TAKANO, H., EMA, H., SUDO, K. & NAKAUCHI, H. Asymmetric division and lineage commitment at the level of hematopoietic stem cells : inference from differentiation in daughter cell and granddaughter cell pairs. eng. *The Journal of Experimental Medicine* **199**, 295-302. ISSN : 0022-1007 (fév. 2004).
6. WILLIAMS, S. E. & FUCHS, E. Oriented divisions, fate decisions. eng. *Current Opinion in Cell Biology* **25**, 749-758. ISSN : 1879-0410 (déc. 2013).
7. LI, Y., WANG, R., QIAO, N., PENG, G., ZHANG, K., TANG, K., HAN, J.-D. J. & JING, N. Transcriptome analysis reveals determinant stages controlling human embryonic stem cell commitment to neuronal cells. en. *Journal of Biological Chemistry* **292**. Number : 48 Reporter : Journal of Biological Chemistry, 19590-19604. ISSN : 0021-9258, 1083-351X (déc. 2017).

8. GRINENKO, T., EUGSTER, A., THIELECKE, L., RAMASZ, B., KRÜGER, A., DIETZ, S., GLAUCHE, I., GERBAULET, A., von BONIN, M., BASAK, O., CLEVERS, H., CHAVAKIS, T. & WIELOCKX, B. Hematopoietic stem cells can differentiate into restricted myeloid progenitors before cell division in mice. en. *Nature Communications* **9**, 1898. ISSN : 2041-1723 (déc. 2018).
9. BELOKHVOSTOVA, D., BERZANSKYTE, I., CUJBA, A.-M., JOWETT, G., MARSHALL, L., PRUELLER, J. & WATT, F. M. Homeostasis, regeneration and tumour formation in the mammalian epidermis. *The International journal of developmental biology* **62**, 571-582. ISSN : 0214-6282 (2018).
10. FUCHS, Y. & STELLER, H. Programmed Cell Death in Animal Development and Disease. *Cell* **147**, 742-758. ISSN : 0092-8674 (nov. 2011).
11. GALLUZZI, L. *et al.* Essential versus accessory aspects of cell death : recommendations of the NCCD 2015. en. *Cell Death & Differentiation* **22**, 58-73. ISSN : 1350-9047, 1476-5403 (jan. 2015).
12. GUDIPATY, S. A., CONNER, C. M., ROSENBLATT, J. & MONTELL, D. J. Unconventional Ways to Live and Die : Cell Death and Survival in Development, Homeostasis, and Disease. eng. *Annual Review of Cell and Developmental Biology* **34**, 311-332. ISSN : 1530-8995 (oct. 2018).
13. ELLIS, R. E., YUAN, J. Y. & HORVITZ, R. H. Mechanisms and functions of cell death. eng. *Annual Review of Cell Biology* **7**, 663-698. ISSN : 0743-4634 (1991).
14. DEKKERS, M. P., NIKOLETOPOULOU, V. & BARDE, Y.-A. Death of developing neurons : New insights and implications for connectivity. *The Journal of Cell Biology* **203**, 385-393. ISSN : 0021-9525 (nov. 2013).
15. OWEN, J. J. T. & JENKINSON, E. J. Apoptosis and T-Cell Repertoire Selection in the Thymus. en. *Annals of the New York Academy of Sciences* **663**, 305-310. ISSN : 0077-8923, 1749-6632 (nov. 1992).
16. SANDERS, E. J. & WRIDE, M. Programmed Cell Death in Development. en. *Int Rev Cytol*, 69 (1995).
17. TESTA, U. Apoptotic mechanisms in the control of erythropoiesis. eng. *Leukemia* **18**, 1176-1199. ISSN : 0887-6924 (juill. 2004).

18. WOLPERT, L. Do we understand development ? en. *Science* **266**, 571-572. ISSN : 0036-8075, 1095-9203 (oct. 1994).
19. WEISSMAN, I. L. Stem Cells : Units of Development, Units of Regeneration, and Units in Evolution. *Cell* **100**, 157-168. ISSN : 0092-8674 (2000).
20. WOLFF, L. & HUMENIUK, R. Concise Review : Erythroid Versus Myeloid Lineage Commitment : Regulating the Master Regulators. en. *Stem Cells* **31**, 1237-1244. ISSN : 1066-5099, 1549-4918 (juill. 2013).
21. NOVICK, A. & WEINER, M. ENZYME INDUCTION AS AN ALL-OR-NONE PHENOMENON. en. *Proceedings of the National Academy of Sciences* **43**, 553-566. ISSN : 0027-8424, 1091-6490 (juill. 1957).
22. LYON, M. F. Gene Action in the X-chromosome of the Mouse (*Mus musculus* L.) *Nature* **190**, 372-373. ISSN : 1476-4687 (avr. 1961).
23. LEVSKY, J. M. & SINGER, R. H. Gene expression and the myth of the average cell. en. *Trends in Cell Biology* **13**, 4-6. ISSN : 0962-8924 (jan. 2003).
24. BOX, G. E. P. & DRAPER, N. R. *Empirical model-building and response surfaces* Pages : xiv, 669. ISBN : 978-0-471-81033-9 (John Wiley & Sons, Oxford, England, 1987).
25. SYMMONS, O. & RAJ, A. What's Luck Got to Do with It : Single Cells, Multiple Fates, and Biological Nondeterminism. en. *Molecular Cell* **62**. Number : 5 Reporter : Molecular Cell, 788-802. ISSN : 10972765 (juin 2016).
26. ARKIN, A., ROSS, J. & MCADAMS, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633-1648. ISSN : 0016-6731 (août 1998).
27. WERNET, M. F., MAZZONI, E. O., ÇELIK, A., DUNCAN, D. M., DUNCAN, I. & DESPLAN, C. Stochastic spineless expression creates the retinal mosaic for colour vision. *Nature* **440**, 10.1038/nature04615. ISSN : 0028-0836 (mars 2006).
28. ZECHNER, C., NERLI, E. & NORDEN, C. Stochasticity and determinism in cell fate decisions. en. *Development* **147**, dev181495. ISSN : 1477-9129, 0950-1991 (juill. 2020).
29. WADDINGTON, C. H. *The Strategy of the Genes* 1st ed. (Routledge, 1957).

30. MCADAMS, H. H. & ARKIN, A. Stochastic mechanisms in gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 814-819. ISSN : 0027-8424 (fév. 1997).
31. KO, M. S. H. A stochastic model for gene induction. en. *Journal of Theoretical Biology* **153**, 181-194. ISSN : 0022-5193 (nov. 1991).
32. KO, M. S., NAKAUCHI, H. & TAKAHASHI, N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. en. *The EMBO Journal* **9**, 2835-2842. ISSN : 02614189 (sept. 1990).
33. ELOWITZ, M. B. & LEIBLER, S. A synthetic oscillatory network of transcriptional regulators. en. *Nature* **403**, 335-338. ISSN : 0028-0836, 1476-4687 (jan. 2000).
34. ELOWITZ, M. B. Stochastic Gene Expression in a Single Cell. en. *Science* **297**. Number : 5584 Reporter : Science, 1183-1186. ISSN : 00368075, 10959203 (août 2002).
35. GREGOR, T., WIESCHAUS, E. F., MCGREGOR, A. P., BIALEK, W. & TANK, D. W. Stability and nuclear dynamics of the Bicoid morphogen gradient. *Cell* **130**, 141-152. ISSN : 0092-8674 (juill. 2007).
36. DURRIEU, L., KIRRMAYER, D., SCHNEIDT, T., KATS, I., RAGHAVAN, S., HUFNAGEL, L., SAUNDERS, T. E. & KNOP, M. Bicoid gradient formation mechanism and dynamics revealed by protein lifetime analysis. *Molecular Systems Biology* **14**, e8355. ISSN : 1744-4292 (sept. 2018).
37. CHABOT, J. R., PEDRAZA, J. M., LUITEL, P. & van OUDENAARDEN, A. Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock. en. *Nature* **450**, 1249-1252. ISSN : 0028-0836, 1476-4687 (déc. 2007).
38. TALIA, S. D., SKOTHEIM, J. M., BEAN, J. M., SIGGIA, E. D. & CROSS, F. R. The effects of molecular noise and size control on variability in the budding yeast cell cycle. en. *Nature* **448**, 947-951. ISSN : 0028-0836, 1476-4687 (août 2007).
39. HUH, D. & PAULSSON, J. Non-genetic heterogeneity from stochastic partitioning at cell division. en. *Nature Genetics* **43**, 95-100. ISSN : 1061-4036, 1546-1718 (fév. 2011).
40. ROSENFELD, N., YOUNG, J. W., ALON, U., SWAIN, P. S. & ELOWITZ, M. B. Gene Regulation at the Single-Cell Level. *Science* **307**. Publisher : American Association for the Advancement of Science, 1962-1965 (mars 2005).

41. OZBUDAK, E. M., THATTAI, M., KURTSEY, I., GROSSMAN, A. D. & van OUDENAARDEN, A. Regulation of noise in the expression of a single gene. en. *Nature Genetics* **31**, 69-73. ISSN : 1061-4036, 1546-1718 (mai 2002).
42. BAHAR HALPERN, K., CASPI, I., LEMZE, D., LEVY, M., LANDEN, S., ELINAV, E., ULITSKY, I. & ITZKOVITZ, S. Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports* **13**, 2653-2662. ISSN : 2211-1247 (déc. 2015).
43. BATTICH, N., STOEGER, T. & PELKMANS, L. Control of Transcript Variability in Single Mammalian Cells. en. *Cell* **163**, 1596-1610. ISSN : 0092-8674 (déc. 2015).
44. CHUBB, J. R., TRCEK, T., SHENOY, S. M. & SINGER, R. H. Transcriptional pulsing of a developmental gene. eng. *Current biology : CB* **16**, 1018-1025. ISSN : 0960-9822 (mai 2006).
45. KÆRN, M., ELSTON, T. C., BLAKE, W. J. & COLLINS, J. J. Stochasticity in gene expression : from theories to phenotypes. en. *Nature Reviews Genetics* **6**. Number : 6 Publisher : Nature Publishing Group, 451-464. ISSN : 1471-0064 (juin 2005).
46. SUTER, D. M., MOLINA, N., GATFIELD, D., SCHNEIDER, K., SCHIBLER, U. & NAEF, F. Mammalian genes are transcribed with widely different bursting kinetics. eng. *Science (New York, N.Y.)* **332**, 472-474. ISSN : 1095-9203 (avr. 2011).
47. LEVESQUE, M. J. & RAJ, A. Single-chromosome transcriptional profiling reveals chromosomal gene expression regulation. en. *Nature Methods* **10**. Number : 3 Publisher : Nature Publishing Group, 246-248. ISSN : 1548-7105 (mars 2013).
48. VIÑUELAS, J., KANEKO, G., COULON, A., VALLIN, E., MORIN, V., MEJIA-POUS, C., KUPIEC, J.-J., BESLON, G. & GANDRILLON, O. Quantifying the contribution of chromatin dynamics to stochastic gene expression reveals long, locus-dependent periods between transcriptional bursts. eng. *BMC biology* **11**, 15. ISSN : 1741-7007 (fév. 2013).
49. RAJ, A., PESKIN, C. S., TRANCHINA, D., VARGAS, D. Y. & TYAGI, S. Stochastic mRNA Synthesis in Mammalian Cells. en. *PLoS Biology* **4** (éd. SCHIBLER, U.) e309. ISSN : 1545-7885 (sept. 2006).
50. WARREN, L., BRYDER, D., WEISSMAN, I. L. & QUAKE, S. R. Transcription factor profiling in individual hematopoietic progenitors by digital RT-PCR. en. *Proceedings of the National Academy of Sciences* **103**, 17807-17812. ISSN : 0027-8424, 1091-6490 (nov. 2006).

51. De KROM, M., van de CORPUT, M., von LINDERN, M., GROSVELD, F. & STROUBOULIS, J. Stochastic Patterns in Globin Gene Expression Are Established prior to Transcriptional Activation and Are Clonally Inherited. en. *Molecular Cell* **9**, 1319-1326. ISSN : 1097-2765 (juin 2002).
52. WEINBERGER, L., VOICHEK, Y., TIROSH, I., HORNUNG, G., AMIT, I. & BARKAI, N. Expression noise and acetylation profiles distinguishes HDACs functions. *Molecular cell* **47**, 193-202. ISSN : 1097-2765 (juill. 2012).
53. BLAKE, W. J., KAERN, M., CANTOR, C. R. & COLLINS, J. J. Noise in eukaryotic gene expression. en. *Nature* **422**, 633-637. ISSN : 0028-0836, 1476-4687 (avr. 2003).
54. BLAKE, W. J., BALAZSI, G., KOHANSKI, M. A., ISAACS, F. J., MURPHY, K. F., KUANG, Y., CANTOR, C. R., WALT, D. R. & COLLINS, J. J. Phenotypic Consequences of Promoter-Mediated Transcriptional Noise. *Molecular Cell* **24**, 853-865. ISSN : 1097-2765 (2006).
55. HABERLE, V. & STARK, A. Eukaryotic core promoters and the functional basis of transcription initiation. en. *Nature Reviews Molecular Cell Biology* **19**. Number : 10 Reporter : Nature Reviews Molecular Cell Biology, 621-637. ISSN : 1471-0072, 1471-0080 (oct. 2018).
56. JACKSON, D. A., HASSAN, A. B., ERRINGTON, R. J. & COOK, P. R. Visualization of focal sites of transcription within human nuclei. *The EMBO Journal* **12**, 1059-1065. ISSN : 0261-4189 (mars 1993).
57. WANSINK, D. G., SCHUL, W., van der KRAAN, I., van STEENSEL, B. & van DRIEL Roel and de Jong, L. Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus. *The Journal of Cell Biology* **122**, 283-293. ISSN : 0021-9525 (juill. 1993).
58. OSBORNE, C. S., CHAKALOVA, L., BROWN, K. E., CARTER, D., HORTON, A., DEBRAND, E., GOYENECHEA, B., MITCHELL, J. A., LOPES, S., REIK, W. & FRASER, P. Active genes dynamically colocalize to shared sites of ongoing transcription. en. *Nature Genetics* **36**, 1065-1071. ISSN : 1061-4036, 1546-1718 (oct. 2004).
59. CARTER, D. R., ESKIW, C. & COOK, P. R. Transcription factories. *Biochemical Society Transactions* **36**, 585-589. ISSN : 0300-5127, 1470-8752 (1^{er} août 2008).

60. SWAIN, P. S., ELOWITZ, M. B. & SIGGIA, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. en. *Proceedings of the National Academy of Sciences* **99**, 12795-12800. ISSN : 0027-8424, 1091-6490 (oct. 2002).
61. JOHNSTON, R. J. & DESPLAN, C. Interchromosomal Communication Coordinates Intrinsically Stochastic Expression Between Alleles. en. *Science* **343**. Number : 6171 Reporter : Science, 661-665. ISSN : 0036-8075, 1095-9203 (fév. 2014).
62. NEMAZEE, D. Receptor Selection in B and T Lymphocytes. *Annual review of immunology* **18**, 10.1146/annurev.immunol.18.1.19. ISSN : 0732-0582 (2000).
63. GOLDING, I., PAULSSON, J., ZAWILSKI, S. M. & COX, E. C. Real-Time Kinetics of Gene Activity in Individual Bacteria. en. *Cell* **123**, 1025-1036. ISSN : 00928674 (déc. 2005).
64. DAR, R. D., RAZOOKY, B. S., SINGH, A., TRIMELONI, T. V., MCCOLLUM, J. M., COX, C. D., SIMPSON, M. L. & WEINBERGER, L. S. Transcriptional burst frequency and burst size are equally modulated across the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 17454-17459. ISSN : 1091-6490 (23 oct. 2012).
65. LARSON, D. R., FRITZSCH, C., SUN, L., MENG, X., LAWRENCE, D. S. & SINGER, R. H. Direct observation of frequency modulated transcription in single cells using light activation. *eLife* **2**, e00750. ISSN : 2050-084X (24 sept. 2013).
66. HANSEN, M. M., DESAI, R. V., SIMPSON, M. L. & WEINBERGER, L. S. Cytoplasmic Amplification of Transcriptional Noise Generates Substantial Cell-to-Cell Variability. en. *Cell Systems* **7**. Number : 4 Reporter : Cell Systems, 384-397.e6. ISSN : 24054712 (oct. 2018).
67. CAI, L., FRIEDMAN, N. & XIE, X. S. Stochastic protein expression in individual cells at the single molecule level. en. *Nature* **440**, 358-362. ISSN : 0028-0836, 1476-4687 (mars 2006).
68. CORRE, G., STOCKHOLM, D., ARNAUD, O., KANEKO, G., VIÑUELAS, J., YAMAGATA, Y., NEILDEZ-NGUYEN, T. M. A., KUPIEC, J.-J., BESLON, G., GANDRILLON, O. & PALDI, A. Stochastic Fluctuations and Distributed Control of Gene Expression Impact Cellular Memory. en. *PLoS ONE* **9** (éd. MACARTHUR, B. D.) e115574. ISSN : 1932-6203 (déc. 2014).

69. NEWMAN, J. R. S., GHAEMMAGHAMI, S., IHMELS, J., BRESLOW, D. K., NOBLE, M., DERISI, J. L. & WEISSMAN, J. S. Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. en. *Nature* **441**, 840-846. ISSN : 0028-0836, 1476-4687 (juin 2006).
70. FRASER, H. B., HIRSH, A. E., GIAEVER, G., KUMM, J. & EISEN, M. B. Noise Minimization in Eukaryotic Gene Expression. en. *PLoS Biology* **2** (éd. KEN WOLFE) e137. ISSN : 1545-7885 (avr. 2004).
71. MUNDT, M., ANDERS, A., MURRAY, S. M. & SOURJIK, V. A System for Gene Expression Noise Control in Yeast. *ACS Synthetic Biology* **7**. Publisher : American Chemical Society, 2618-2626 (16 nov. 2018).
72. BECSKEI, A. & SERRANO, L. Engineering stability in gene networks by autoregulation. eng. *Nature* **405**, 590-593. ISSN : 0028-0836 (juin 2000).
73. ANTOLOVIC, V., MIERMONT, A., CORRIGAN, A. M. & CHUBB, J. R. Generation of Single-Cell Transcript Variability by Repression. *Current biology : CB* **27**, 1811-1817.e3. ISSN : 1879-0445 (19 juin 2017).
74. RASER, J. M. & O'SHEA, E. K. Noise in Gene Expression : Origins, Consequences, and Control. *Science* **309**, 2010-2013. ISSN : 0036-8075, 1095-9203 (23 sept. 2005).
75. SINGH, A. & BOKES, P. Consequences of mRNA Transport on Stochastic Variability in Protein Levels. en. *Biophysical Journal* **103**, 1087-1096. ISSN : 00063495 (sept. 2012).
76. HU, T., WEI, L., LI, S., CHENG, T., ZHANG, X. & WANG, X. Single-cell Transcriptomes Reveal Characteristics of MicroRNAs in Gene Expression Noise Reduction. en. *Genomics, Proteomics & Bioinformatics* **19**, 394-407. ISSN : 16720229 (juin 2021).
77. LESTAS, I., VINNICOMBE, G. & PAULSSON, J. Fundamental limits on the suppression of molecular fluctuations. en. *Nature* **467**, 174-178. ISSN : 0028-0836, 1476-4687 (sept. 2010).
78. KUSSELL, E. & LEIBLER, S. Phenotypic Diversity, Population Growth, and Information in Fluctuating Environments. en. *Science* **309**, 2075-2078. ISSN : 0036-8075, 1095-9203 (sept. 2005).
79. SUEL, G. M., GARCIA-OJALVO, J., LIBERMAN, L. M. & ELOWITZ, M. B. An excitable gene regulatory circuit induces transient cellular differentiation. en. *Nature* **440**, 545-550. ISSN : 0028-0836, 1476-4687 (mars 2006).

80. THATTAI, M. & van OUDENAARDEN, A. Stochastic gene expression in fluctuating environments. *Genetics* **167**, 523-530. ISSN : 0016-6731 (mai 2004).
81. VASSAR, R., NGAI, J. & AXEL, R. Spatial segregation of odorant receptor expression in the mammalian olfactory epithelium. en. *Cell* **74**, 309-318. ISSN : 00928674 (juill. 1993).
82. MOMBAERTS, P. Odorant receptor gene choice in olfactory sensory neurons : the one receptor one neuron hypothesis revisited. en. *Current Opinion in Neurobiology* **14**, 31-36. ISSN : 09594388 (fév. 2004).
83. NATHANS, J. The Evolution and Physiology of Human Color Vision : Insights from Molecular Genetic Studies of Visual Pigments. en. *Neuron* **24**, 299-312. ISSN : 0896-6273 (oct. 1999).
84. DOE, C. Q. & SKEATH, J. B. Neurogenesis in the insect central nervous system. en. *Current Opinion in Neurobiology* **6**, 18-24. ISSN : 09594388 (fév. 1996).
85. CHANG, H. H., HEMBERG, M., BARAHONA, M., INGBER, D. E. & HUANG, S. Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. en. *Nature* **453**, 544-547. ISSN : 0028-0836, 1476-4687 (mai 2008).
86. BEACH, D. L., SALMON, E. D. & BLOOM, K. Localization and anchoring of mRNA in budding yeast. *Current Biology* **9**, 569-S1. ISSN : 0960-9822 (1999).
87. BERTRAND, E., CHARTRAND, P., SCHAEFER, M., SHENOY, S. M., SINGER, R. H. & LONG, R. M. Localization of ASH1 mRNA Particles in Living Yeast. *Molecular Cell* **2**, 437-445. ISSN : 1097-2765 (1998).
88. FEMINO, A. M., FAY, F. S., FOGARTY, K. & SINGER, R. H. Visualization of Single RNA Transcripts in Situ. en. *Science* **280**, 585-590. ISSN : 0036-8075, 1095-9203 (avr. 1998).
89. RAJ, A., van den BOGAARD, P., RIFKIN, S. A., van OUDENAARDEN, A. & TYAGI, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nature methods* **5**, 877-879. ISSN : 1548-7091 (oct. 2008).
90. CHANG, C.-H., MAU-HSU, D., CHEN, K.-C., WEI, C.-W., CHIU, C.-Y. & YOUNG, T.-H. Evaluation of digital real-time PCR assay as a molecular diagnostic tool for single-cell analysis. en. *Scientific Reports* **8**, 3432. ISSN : 2045-2322 (déc. 2018).

91. PORTER, J. R., TELFORD, W. G. & BATCHELOR, E. Single-cell Gene Expression Profiling Using FACS and qPCR with Internal Standards. en. *Journal of Visualized Experiments*, 55219. ISSN : 1940-087X (fév. 2017).
92. KOLODZIEJCZYK, A. A. & LÖNNBERG, T. Global and targeted approaches to single-cell transcriptome characterization. en. *Briefings in Functional Genomics* **17**, 209-219. ISSN : 2041-2649, 2041-2657 (juill. 2018).
93. ZIEGENHAIN, C., VIETH, B., PAREKH, S., REINIUS, B., GUILLAUMET-ADKINS, A., SMETS, M., LEONHARDT, H., HEYN, H., HELLMANN, I. & ENARD, W. Comparative Analysis of Single-Cell RNA Sequencing Methods. en. *Molecular Cell* **65**. Number : 4 Reporter : Molecular Cell, 631-643.e4. ISSN : 10972765 (fév. 2017).
94. ZHANG, M. J., NTRANOS, V. & TSE, D. Determining sequencing depth in a single-cell RNA-seq experiment. en. *Nature Communications* **11**. Number : 1 Reporter : Nature Communications, 774. ISSN : 2041-1723 (déc. 2020).
95. HICKS, S. C., TOWNES, F. W., TENG, M. & IRIZARRY, R. A. *Missing Data and Technical Variability in Single-Cell RNA- Sequencing Experiments* en. preprint (Genomics, août 2015).
96. SVENSSON, V., NATARAJAN, K. N., LY, L.-H., MIRAGAIA, R. J., LABALETTE, C., MACAULAY, I. C., CVEJIC, A. & TEICHMANN, S. A. Power analysis of single-cell RNA-sequencing experiments. en. *Nature Methods* **14**. Number : 4 Reporter : Nature Methods, 381-387. ISSN : 1548-7091, 1548-7105 (avr. 2017).
97. DEEKE, J. M. & GAGNON-BARTSCH, J. A. Stably expressed genes in single-cell RNA sequencing. *Journal of Bioinformatics and Computational Biology* **18**. Publisher : World Scientific Publishing Co., 2040004. ISSN : 0219-7200 (fév. 2020).
98. KIM, H.-J., BOOTH, G., SAUNDERS, L., SRIVATSAN, S., MCFALINE-FIGUEROA, J. L. & TRAPNELL, C. Nuclear oligo hashing improves differential analysis of single-cell RNA-seq. *Nature Communications* **13**. Number : 1 Publisher : Nature Publishing Group, 2666. ISSN : 2041-1723 (13 mai 2022).
99. GRIFFITHS, J. A., SCIALDONE, A. & MARIONI, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. en. *Molecular Systems Biology* **14**. Number : 4 Reporter : Molecular Systems Biology, e8046. ISSN : 1744-4292, 1744-4292, 1744-4292 (avr. 2018).

100. GALOW, A.-M., KUSSAUER, S., WOLFIEN, M., BRUNNER, R. M., GOLDAMMER, T., DAVID, R. & HOEFLICH, A. Quality control in scRNA-Seq can discriminate pacemaker cells : the mtRNA bias. en. *Cellular and Molecular Life Sciences*. ISSN : 1420-682X, 1420-9071 (août 2021).
101. COLE, M. B., RISSO, D., WAGNER, A., DETOMASO, D., NGAI, J., PURDOM, E., DUDOIT, S. & YOSEF, N. Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. en. *Cell Systems* **8**. Number : 4 Reporter : Cell Systems, 315-328.e8. ISSN : 24054712 (avr. 2019).
102. HAFEMEISTER, C. & SATIJA, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. en. *bioRxiv*. Reporter : bioRxiv (mars 2019).
103. BREDA, J., ZAVOLAN, M. & van NIMWEGEN, E. Bayesian inference of gene expression states from single-cell RNA-seq data. en. *Nature Biotechnology* **39**, 1008-1016. ISSN : 1087-0156, 1546-1696 (août 2021).
104. GAO, M., LING, M., TANG, X., WANG, S., XIAO, X., QIAO, Y., YANG, W. & YU, R. *Comparison of High-Throughput Single-Cell RNA Sequencing Data Processing Pipelines* en. preprint (Bioinformatics, fév. 2020).
105. VIETH, B., PAREKH, S., ZIEGENHAIN, C., ENARD, W. & HELLMANN, I. A systematic evaluation of single cell RNA-seq analysis pipelines. en. *Nature Communications* **10**. Number : 1 Reporter : Nature Communications. ISSN : 2041-1723 (déc. 2019).
106. SUN, S., ZHU, J., MA, Y. & ZHOU, X. Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis. en. *Genome Biology* **20**. Number : 1 Reporter : Genome Biology, 269. ISSN : 1474-760X (déc. 2019).
107. JOLLIFFE, I. T. *Principal Component Analysis* en (2002).
108. LINDERMAN, G. C., RACHH, M., HOSKINS, J. G., STEINERBERGER, S. & KLUGER, Y. Fast Interpolation-based t-SNE for Improved Visualization of Single-Cell RNA-Seq Data. *Nature methods* **16**, 243-245. ISSN : 1548-7091 (mars 2019).
109. BECHT, E., MCINNES, L., HEALY, J., DUTERTRE, C.-A., KWOK, I. W. H., NG, L. G., GINHOUX, F. & NEWELL, E. W. Dimensionality reduction for visualizing single-cell data

- using UMAP. en. *Nature Biotechnology* **37**. Number : 1 Reporter : Nature Biotechnology, 38-44. ISSN : 1087-0156, 1546-1696 (jan. 2019).
110. KISELEV, V. Y., ANDREWS, T. S. & HEMBERG, M. Challenges in unsupervised clustering of single-cell RNA-seq data. en. *Nature Reviews Genetics* **20**. Number : 5 Reporter : Nature Reviews Genetics, 273-282. ISSN : 1471-0056, 1471-0064 (mai 2019).
111. MIRCEA, M., HOCHANE, M., FAN, X., CHUVA DE SOUSA LOPES, S. M., GARLASCHELLI, D. & SEMRAU, S. Phiclust : a clusterability measure for single-cell transcriptomics reveals phenotypic subpopulations. *Genome Biology* **23**, 18. ISSN : 1474-760X (10 jan. 2022).
112. MOU, T., DENG, W., GU, F., PAWITAN, Y. & VU, T. N. Reproducibility of Methods to Detect Differentially Expressed Genes from Single-Cell RNA Sequencing. en. *Frontiers in Genetics* **10**. Reporter : Frontiers in Genetics, 1331. ISSN : 1664-8021 (jan. 2020).
113. SINGH, A. & SOLTANI, M. Quantifying Intrinsic and Extrinsic Variability in Stochastic Gene Expression Models. *PLoS ONE* **8**, e84301. ISSN : 1932-6203 (déc. 2013).
114. VALLEJOS, C. A., RICHARDSON, S. & MARIONI, J. C. Beyond comparisons of means : understanding changes in gene expression at the single-cell level. *Genome Biology* **17**, 70. ISSN : 1474-760X (avr. 2016).
115. GANDRILLON, O., GAILLARD, M., ESPINASSE, T., GARNIER, N. B., DUSSIAU, C., KOSMIDER, O. & SUJOBERT, P. Entropy as a measure of variability and stemness in single-cell transcriptomics. en. *Current Opinion in Systems Biology* **27**, 100348. ISSN : 2452-3100 (sept. 2021).
116. SAELENS, W., CANNOODT, R., TODOROV, H. & SAEYS, Y. A comparison of single-cell trajectory inference methods. en. *Nature Biotechnology* **37**. Number : 5 Publisher : Nature Publishing Group, 547-554. ISSN : 1546-1696 (mai 2019).
117. VENTRE, E. Reverse engineering of a mechanistic model of gene expression using metastability and temporal dynamics. *In Silico Biology* **14**, 89-113. ISSN : 1434-3207 (2021).
118. BONNAFFOUX, A., HERBACH, U., RICHARD, A., GUILLEMIN, A., GONIN-GIRAUD, S., GROS, P.-A. & GANDRILLON, O. WASABI : a dynamic iterative framework for gene regulatory network inference. en. *BMC Bioinformatics* **20**, 220. ISSN : 1471-2105 (déc. 2019).

119. CACACE, E., COLLOMBET, S. & THIEFFRY, D. Logical modeling of cell fate specification- Application to T cell commitment. *Current Topics in Developmental Biology* **139**, 205-238. ISSN : 1557-8933 (2020).
120. MORIS, N., PINA, C. & ARIAS, A. M. Transition states and cell fate decisions in epigenetic landscapes. en. *Nature Reviews Genetics* **17**, 693-703. ISSN : 1471-0056, 1471-0064 (nov. 2016).
121. BROWN, G. Towards a New Understanding of Decision-Making by Hematopoietic Stem Cells. eng. *International Journal of Molecular Sciences* **21**, E2362. ISSN : 1422-0067 (mars 2020).
122. MEVEL, R., DRAPER, J. E., LIE-A-LING, M., KOUSKOFF, V. & LACAUD, G. RUNX transcription factors : orchestrators of development. *Development (Cambridge, England)* **146**, dev148296. ISSN : 1477-9129 (5 sept. 2019).
123. DZIERZAK, E. & PHILIPSEN, S. Erythropoiesis : Development and Differentiation. *Cold Spring Harbor Perspectives in Medicine* **3**, a011601. ISSN : 2157-1422 (avr. 2013).
124. BONNET, D. Haematopoietic stem cells. eng. *The Journal of Pathology* **197**, 430-440. ISSN : 0022-3417 (juill. 2002).
125. LE DOUARIN, N. M., DIETERLEN-LIÈVRE, F. & OLIVER, P. D. Ontogeny of primary lymphoid organs and lymphoid stem cells : ONTOGENY OF LYMPHOID ORGANS AND STEM CELLS. en. *American Journal of Anatomy* **170**, 261-299. ISSN : 00029106 (juill. 1984).
126. MOORE, M. A. S. & OWEN, J. J. T. EXPERIMENTAL STUDIES ON THE DEVELOPMENT OF THE THYMUS. *The Journal of Experimental Medicine* **126**, 715-726. ISSN : 0022-1007 (oct. 1967).
127. TAVIAN, M., BIASCH, K., SINKA, L., VALLET, J. & PEAULT, B. Embryonic origin of human hematopoiesis. en. *The International Journal of Developmental Biology* **54**, 1061-1065. ISSN : 0214-6282 (2010).
128. WATT, F. M. & HOGAN, B. L. Out of Eden : stem cells and their niches. eng. *Science (New York, N.Y.)* **287**, 1427-1430. ISSN : 0036-8075 (fév. 2000).

129. BROWN, G. & SANCHEZ-GARCIA, I. Is lineage decision-making restricted during tumoral reprogramming of haematopoietic stem cells? *Oncotarget* **6**, 43326-43341. ISSN : 1949-2553 (22 déc. 2015).
130. TSIFTSOGLU, A. S., VIZIRIANAKIS, I. S. & STROUBOULIS, J. Erythropoiesis : Model systems, molecular regulators, and developmental programs. en. *IUBMB Life* **61**. _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/iub.226>, 800-830. ISSN : 1521-6551 (2009).
131. RIEGER, M. A. & SCHROEDER, T. Hematopoiesis. *Cold Spring Harbor Perspectives in Biology* **4**, a008250. ISSN : 1943-0264 (déc. 2012).
132. BHOOPALAN, S. V., HUANG, L. J.-S. & WEISS, M. J. Erythropoietin regulation of red blood cell production : from bench to bedside and back. *F1000Research* **9**, F1000 Faculty Rev-1153. ISSN : 2046-1402 (2020).
133. DEMIN, I., CRAUSTE, F., GANDRILLON, O. & VOLPERT, V. A multi-scale model of erythropoiesis. *Journal of Biological Dynamics* **4**. Publisher : Taylor & Francis _eprint : <https://doi.org/10.1080/17513750902777642>, 59-70. ISSN : 1751-3758 (jan. 2010).
134. LAURENTI, E. & GÖTTGENS, B. From haematopoietic stem cells to complex differentiation landscapes. *Nature* **553**, 418-426. ISSN : 0028-0836 (jan. 2018).
135. WAGERS, A. J., CHRISTENSEN, J. L. & WEISSMAN, I. L. Cell fate determination from stem cells. en. *Gene Therapy* **9**, 606-612. ISSN : 0969-7128, 1476-5462 (mai 2002).
136. KONDO, M., WEISSMAN, I. L. & AKASHI, K. Identification of Clonogenic Common Lymphoid Progenitors in Mouse Bone Marrow. en. *Cell* **91**, 661-672. ISSN : 0092-8674 (nov. 1997).
137. DUSSIAU, C., BOUSSAROQUE, A., GAILLARD, M., BRAVETTI, C., ZAROILI, L., KNOSP, C., FRIEDRICH, C., ASQUIER, P., WILLEMS, L., QUINT, L., BOUSCARY, D., FONTENAY, M., ESPINASSE, T., PLESA, A., SUJOBERT, P., GANDRILLON, O. & KOSMIDER, O. Hematopoietic differentiation is characterized by a transient peak of entropy at a single-cell level. *BMC Biology* **20**, 60. ISSN : 1741-7007 (mars 2022).
138. ENVER, T., HEYWORTH, C. M. & DEXTER, M. T. Do Stem Cells Play Dice? en. *Blood* **92**, 348-351. ISSN : 1528-0020, 0006-4971 (juill. 1998).

139. ORKIN, S. H. & ZON, L. I. Hematopoiesis : An Evolving Paradigm for Stem Cell Biology. en. *Cell* **132**, 631-644. ISSN : 00928674 (fév. 2008).
140. HOPPE, P. S., SCHWARZFISCHER, M., LOEFFLER, D., KOKKALIARIS, K. D., HILSENBECK, O., MORITZ, N., ENDELE, M., FILIPCZYK, A., GAMBARDILLA, A., AHMED, N., ETZRODT, M., COUTU, D. L., RIEGER, M. A., MARR, C., STRASSER, M. K., SCHAUBERGER, B., BURTSCHER, I., ERMAKOVA, O., BÜRGER, A., LICKERT, H., NERLOV, C., THEIS, F. J. & SCHROEDER, T. Early myeloid lineage choice is not initiated by random PU.1 to GATA1 protein ratios. en. *Nature* **535**, 299-302. ISSN : 0028-0836, 1476-4687 (juill. 2016).
141. CLEVERS, H. What Is Your Conceptual Definition of Cell Type in the Context of a Mature Organism? English. *Cell Systems* **4**. Publisher : Elsevier, 255-259. ISSN : 2405-4712 (mars 2017).
142. MOIGNARD, V., MACAULAY, I. C., SWIERS, G., BUETTNER, F., SCHÜTTE, J., CALERO-NIETO, F. J., KINSTON, S., JOSHI, A., HANNAH, R., THEIS, F. J., JACOBSEN, S. E., de BRUIJN, M. & GÖTTGENS, B. Characterisation of transcriptional networks in blood stem and progenitor cells using high-throughput single cell gene expression analysis. *Nature cell biology* **15**, 363-372. ISSN : 1465-7392 (avr. 2013).
143. VELTEN, L., HAAS, S. F., RAFFEL, S., BLASZKIEWICZ, S., ISLAM, S., HENNIG, B. P., HIRCHE, C., LUTZ, C., BUSS, E. C., NOWAK, D., BOCH, T., HOFMANN, W.-K., HO, A. D., HUBER, W., TRUMPP, A., ESSERS, M. A. & STEINMETZ, L. M. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature cell biology* **19**, 271-281. ISSN : 1465-7392 (avr. 2017).
144. KARAMITROS, D., STOILOVA, B., ABOUKHALIL, Z., HAMEY, F., REINISCH, A., SAMITSCH, M., QUEK, L., OTTO, G., REPAPI, E., DOONDEEA, J., USUKHBAYAR, B., CALVO, J., TAYLOR, S., GOARDON, N., SIX, E., PFLUMIO, F., PORCHER, C., MAJETI, R., GOTTGENS, B. & VYAS, P. Heterogeneity of human lympho-myeloid progenitors at the single cell level. *Nature immunology* **19**, 85-97. ISSN : 1529-2908 (jan. 2018).
145. LLOYD, J. A. en. in *Erythropoiesis : Methods and Protocols* (éd. LLOYD, J. A.) 1-10 (Springer, New York, NY, 2018). ISBN : 978-1-4939-7428-3.
146. KOVILAKATH, A., MOHAMAD, S., HERMES, F., WANG, S. Z., GINDER, G. D. & LLOYD, J. A. in *Erythropoiesis* (éd. LLOYD, J. A.) Series Title : Methods in Molecular Biology, 259-274 (Springer New York, New York, NY, 2018).

147. ELLIOTT, S., PHAM, E. & MACDOUGALL, I. C. Erythropoietins : a common mechanism of action. eng. *Experimental Hematology* **36**, 1573-1584. ISSN : 0301-472X (déc. 2008).
148. HATTANGADI, S. M., WONG, P., ZHANG, L., FLYGARE, J. & LODISH, H. F. From stem cell to red cell : regulation of erythropoiesis at multiple levels by multiple proteins, RNAs, and chromatin modifications. *Blood* **118**, 6258-6268. ISSN : 0006-4971 (déc. 2011).
149. LEDUC, M., GAUTIER, E.-F., GUILLEMIN, A., BROUSSARD, C., SALNOT, V., LACOMBE, C., GANDRILLON, O., GUILLONNEAU, F. & MAYEUX, P. Deep proteomic analysis of chicken erythropoiesis. en. *bioRxiv*. Reporter : bioRxiv (mars 2018).
150. CANNON, M., PHILLIPS, H., SMITH, S., MITCHELL, S., LANDES, K., DESAI, P., BYRD, J. & LAPALOMBELLA, R. Red blood cells differentiated in vitro using sequential liquid and semi-solid culture as a pre-clinical model. *Experimental Hematology & Oncology* **10**, 50. ISSN : 2162-3619 (déc. 2021).
151. BEACON, T. H. & DAVIE, J. R. The chicken model organism for epigenomic research. eng. *Genome* **64**, 476-489. ISSN : 1480-3321 (avr. 2021).
152. SMITH, N. & ENGELBERT, V. E. Erythropoiesis in chicken peripheral blood. en. *Canadian Journal of Zoology* **47**, 1269-1273. ISSN : 0008-4301, 1480-3283 (nov. 1969).
153. GINDER, G. D., WOOD, W. I. & FELSENFELD, G. Isolation and characterization of recombinant clones containing the chicken adult beta-globin gene. en. *Journal of Biological Chemistry* **254**, 8099-8102. ISSN : 00219258 (oct. 1979).
154. SHENG, G. Primitive and definitive erythropoiesis in the yolk sac : a birds eye view. en. *The International Journal of Developmental Biology* **54**, 1033-1043. ISSN : 0214-6282 (2010).
155. DIETERLEN-LIEVRE, F. On the origin of haemopoietic stem cells in the avian embryo : an experimental approach. en, 13.
156. CAPRIOLI, A., JAFFREDO, T., GAUTIER, R., DUBOURG, C. & DIETERLEN-LIÈVRE, F. Blood-borne seeding by hematopoietic and endothelial precursors from the allantois. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 1641-1646. ISSN : 0027-8424 (fév. 1998).
157. WONG, E. A. & UNI, Z. Centennial Review : The chicken yolk sac is a multifunctional organ. en. *Poultry Science* **100**, 100821. ISSN : 0032-5791 (mars 2021).

158. GANDRILLON, O., SCHMIDT, U., BEUG, H. & SAMARUT, J. TGF- β cooperates with TGF- α to induce the self-renewal of normal erythrocytic progenitors : evidence for an autocrine mechanism. en. *The EMBO Journal* **18**. Number : 10 Reporter : The EMBO Journal, 2764-2781. ISSN : 0261-4189, 1460-2075 (mai 1999).
159. GANDRILLON, O. & SAMARUT, J. Role of the different RAR isoforms in controlling the erythrocytic differentiation sequence. Interference with the v-erbA and p135gag-myb-ets nuclear oncogenes. *Oncogene* **16**, 563-574. ISSN : 0950-9232 (5 fév. 1998).
160. RICHARD, A., BOULLU, L., HERBACH, U., BONNAFOUX, A., MORIN, V., VALLIN, E., GUILLEMIN, A., PAPILI GAO, N., GUNAWAN, R., COSETTE, J., ARNAUD, O., KUPIEC, J.-J., ESPINASSE, T., GONIN-GIRAUD, S. & GANDRILLON, O. Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. en. *PLOS Biology* **14** (éd. TEICHMANN, S. A.) Number : 12 Reporter : PLOS Biology, e1002585. ISSN : 1545-7885 (déc. 2016).
161. GUILLEMIN, A., DUCHESNE, R., CRAUSTE, F., GONIN-GIRAUD, S. & GANDRILLON, O. Drugs modulating stochastic gene expression affect the erythroid differentiation process. en. *PLOS ONE* **14** (éd. PINA, C.) Number : 11 Reporter : PLOS ONE, e0225166. ISSN : 1932-6203 (nov. 2019).
162. MOUSSY, A., COSETTE, J., PARMENTIER, R., da SILVA, C., CORRE, G., RICHARD, A., GANDRILLON, O., STOCKHOLM, D. & PÁLDI, A. Integrated time-lapse and single-cell transcription studies highlight the variable and dynamic nature of human hematopoietic cell fate commitment. en. *PLOS Biology* **15** (éd. HUANG, S.) e2001867. ISSN : 1545-7885 (juill. 2017).
163. SEMRAU, S., GOLDMANN, J. E., SOUMILLON, M., MIKKELSEN, T. S., JAENISCH, R. & van OUDENAARDEN, A. Dynamics of lineage commitment revealed by single-cell transcriptomics of differentiating embryonic stem cells. en. *Nature Communications* **8**. Number : 1 Reporter : Nature Communications. ISSN : 2041-1723 (déc. 2017).
164. STUMPF, P. S., SMITH, R. C., LENZ, M., SCHUPPERT, A., MÜLLER, F.-J., BAPTIE, A., CHAN, T. E., STUMPF, M. P., PLEASE, C. P., HOWISON, S. D., ARAI, F. & MACARTHUR, B. D. Stem Cell Differentiation as a Non-Markov Stochastic Process. en. *Cell Systems* **5**. Number : 3 Reporter : Cell Systems, 268-282.e7. ISSN : 24054712 (sept. 2017).

165. MOJTAHEDI, M., SKUPIN, A., ZHOU, J., CASTAÑO, I. G., LEONG-QUONG, R. Y. Y., CHANG, H., TRACHANA, K., GIULIANI, A. & HUANG, S. Cell Fate Decision as High-Dimensional Critical State Transition. en. *PLOS Biology* **14**. Number : 12 Reporter : PLOS Biology, e2000640. ISSN : 1545-7885 (déc. 2016).
166. REBHAHN, J. A., DENG, N., SHARMA, G., LIVINGSTONE, A. M., HUANG, S. & MOSMANN, T. R. An animated landscape representation of CD4⁺ T-cell differentiation, variability, and plasticity : Insights into the behavior of populations versus cells. *European Journal of Immunology* **44**, 2216-2229. ISSN : 0014-2980 (août 2014).
167. KAUFMANN, B. B., YANG, Q., METTETAL, J. T. & van OUDENAARDEN, A. Heritable Stochastic Switching Revealed by Single-Cell Genealogy. en. *PLoS Biology* **5** (éd. GELFAND, M. S.) e239. ISSN : 1545-7885 (sept. 2007).
168. DUFFY, K. R. & HODGKIN, P. D. Intracellular competition for fates in the immune system. en. *Trends in Cell Biology* **22**, 457-464. ISSN : 0962-8924 (sept. 2012).
169. FERRARO, T., ESPOSITO, E., MANCINI, L., NG, S., LUCAS, T., COPPEY, M., DOSTATNI, N., WALCZAK, A. M., LEVINE, M. & LAGHA, M. Transcriptional memory in the Drosophila embryo. *Current biology : CB* **26**, 212-218. ISSN : 0960-9822 (jan. 2016).
170. D'URSO, A. & BRICKNER, J. H. Mechanisms of epigenetic memory. en. *Trends in Genetics* **30**. Number : 6 Reporter : Trends in Genetics, 230-236. ISSN : 01689525 (juin 2014).
171. PTASHNE, M. On the use of the word epigenetic. English. *Current Biology* **17**. Publisher : Elsevier, R233-R236. ISSN : 0960-9822 (avr. 2007).
172. SHAFFER, S. M., EMERT, B. L., REYES HUEROS, R. A., COTE, C., HARMANGE, G., SCHAFF, D. L., SIZEMORE, A. E., GUPTE, R., TORRE, E., SINGH, A., BASSETT, D. S. & RAJ, A. Memory Sequencing Reveals Heritable Single-Cell Gene Expression Programs Associated with Distinct Cellular Behaviors. en. *Cell* **182**. Number : 4 Reporter : Cell, 947-959.e17. ISSN : 00928674 (août 2020).
173. MURAMOTO, T., MULLER, I., THOMAS, G., MELVIN, A. & CHUBB, J. R. Methylation of H3K4 Is required for inheritance of active transcriptional states. *Curr Biol* **20**, 397-406. ISSN : 1879-0445 (Electronic) 0960-9822 (Linking) (2010).

174. HENIKOFF, S. & GREALLY, J. M. Epigenetics, cellular memory and gene regulation. en. *Current Biology* **26**. Number : 14 Reporter : Current Biology, R644-R648. ISSN : 09609822 (juill. 2016).
175. PHILLIPS, N. E., MANDIC, A., OMIDI, S., NAEF, F. & SUTER, D. M. Memory and relatedness of transcriptional activity in mammalian cell lineages. en. *Nature Communications* **10**. Number : 1 Reporter : Nature Communications. ISSN : 2041-1723 (déc. 2019).
176. MOLD, J. E., WEISSMAN, M., RATZ, M., HAGEMANN-JENSEN, M., HARD, J., ERIKSSON, C.-J., TOOSI, H., BERGENSTRAHLE, J. A., von BERLIN, L., MARTIN, M., BLOM, K., LAGERGREN, J., LUNDEBERG, J., SANDBERG, R., MICHAELSSON, J. & FRISEN, J. *Clonally heritable gene expression imparts a layer of diversity within cell types* en. preprint (Systems Biology, fév. 2022).
177. SIGAL, A., MILO, R., COHEN, A., GEVA-ZATORSKY, N., KLEIN, Y., LIRON, Y., ROSENFELD, N., DANON, T., PERZOV, N. & ALON, U. Variability and memory of protein levels in human cells. en. *Nature* **444**, 643-646. ISSN : 0028-0836, 1476-4687 (nov. 2006).
178. SCHWANHÄUSSER, B., WOLF, J., SELBACH, M. & BUSSE, D. Synthesis and degradation jointly determine the responsiveness of the cellular proteome : Insights & Perspectives. en. *BioEssays* **35**, 597-601. ISSN : 02659247 (juill. 2013).
179. CLAVERÍA, C., GIOVINAZZO, G., SIERRA, R. & TORRES, M. Myc-driven endogenous cell competition in the early mammalian embryo. en. *Nature* **500**, 39-44. ISSN : 0028-0836, 1476-4687 (août 2013).
180. SHAFFER, S. M., DUNAGIN, M. C., TORBORG, S. R., TORRE, E. A., EMERT, B., KREPLER, C., BEQIRI, M., SPROESSER, K., BRAFFORD, P. A., XIAO, M., EGGAN, E., ANASTOPOULOS, I. N., VARGAS-GARCIA, C. A., SINGH, A., NATHANSON, K. L., HERLYN, M. & RAJ, A. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. en. *Nature* **546**, 431-435. ISSN : 0028-0836, 1476-4687 (juin 2017).
181. MEIR, Z., MUKAMEL, Z., CHOMSKY, E., LIFSHITZ, A. & TANAY, A. Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. *Nature genetics* **52**, 709-718. ISSN : 1061-4036 (1^{er} juill. 2020).
182. LI, B., ZEIS, P., ALEKSEENKO, A., LIN, G., TEKKEDIL, M. M., STEINMETZ, L. M. & PELECHANO, V. *Differential regulation of mRNA stability modulates transcriptional memory and facilitates environmental adaptation* en. preprint (Genomics, mai 2022).

183. ZALC, A., SINHA, R., GULATI, G. S., WESCHE, D. J., DASZCZUK, P., SWIGUT, T., WEISSMAN, I. L. & WYSOCKA, J. Reactivation of the pluripotency program precedes formation of the cranial neural crest. en. *Science* **371**, eabb4776. ISSN : 0036-8075, 1095-9203 (fév. 2021).
184. WANG, A., YUE, F., LI, Y., XIE, R., HARPER, T., PATEL, N. A., MUTH, K., PALMER, J., QIU, Y., WANG, J., LAM, D. K., RAUM, J. C., STOFFERS, D. A., REN, B. & SANDER, M. Epigenetic priming of enhancers predicts developmental competence of hESC-derived endodermal lineage intermediates. *Cell Stem Cell* **16**, 386-399. ISSN : 1875-9777 (2 avr. 2015).
185. DUNN, J., MCCUAIG, R., TU, W. J., HARDY, K. & RAO, S. Multi-layered epigenetic mechanisms contribute to transcriptional memory in T lymphocytes. *BMC Immunology* **16**, 27. ISSN : 1471-2172 (6 mai 2015).
186. NICHOLS, J. M., ANTOLOVIĆ, V., REICH, J. D., BRAMEYER, S., PASCHKE, P. & CHUBB, J. R. Cell and molecular transitions during efficient dedifferentiation. en. *eLife* **9**, e55435. ISSN : 2050-084X (avr. 2020).
187. BUCZACKI, S. J. A., ZECCHINI, H. I., NICHOLSON, A. M., RUSSELL, R., VERMEULEN, L., KEMP, R. & WINTON, D. J. Intestinal label-retaining cells are secretory precursors expressing Lgr5. en. *Nature* **495**, 65-69. ISSN : 0028-0836, 1476-4687 (mars 2013).
188. TATA, P. R., MOU, H., PARDO-SAGANTA, A., ZHAO, R., PRABHU, M., LAW, B. M., VINARSKY, V., CHO, J. L., BRETON, S., SAHAY, A., MEDOFF, B. D. & RAJAGOPAL, J. Dedifferentiation of committed epithelial cells into stem cells in vivo. en. *Nature* **503**, 218-223. ISSN : 0028-0836, 1476-4687 (nov. 2013).
189. MARTINEZ-FERNANDEZ, A., NELSON, T. J. & TERZIC, A. Nuclear reprogramming strategy modulates differentiation potential of induced pluripotent stem cells. *Journal of Cardiovascular Translational Research* **4**, 131-137. ISSN : 1937-5395 (avr. 2011).
190. LUU, P.-L., GEROVSKA, D., SCHÄFFLER, H. R. & ARAËZO-BRAVO, M. J. Rules governing the mechanism of epigenetic reprogramming memory. *Epigenomics* **10**, 149-174. ISSN : 1750-192X (fév. 2018).
191. RICHARDS, J. L., ZACHARIAS, A. L., WALTON, T., BURDICK, J. T. & MURRAY, J. I. A quantitative model of normal *Caenorhabditis elegans* embryogenesis and its disruption after stress. *Developmental Biology* **374**, 12-23. ISSN : 0012-1606 (1^{er} fév. 2013).

192. SULSTON, J. E. & HORVITZ, H. R. Post-embryonic cell lineages of the nematode, *Caenorhabditis elegans*. eng. *Developmental Biology* **56**, 110-156. ISSN : 0012-1606 (mars 1977).
193. LIVET, J., WEISSMAN, T. A., KANG, H., DRAFT, R. W., LU, J., BENNIS, R. A., SANES, J. R. & LICHTMAN, J. W. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature* **450**, 56-62. ISSN : 0028-0836, 1476-4687 (nov. 2007).
194. WEINREB, C., RODRIGUEZ-FRATICELLI, A., CAMARGO, F. D. & KLEIN, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. en. *Science* **367**, eaaw3381. ISSN : 0036-8075, 1095-9203 (fév. 2020).
195. ALEMANY, A., FLORESCU, M., BARON, C. S., PETERSON-MADURO, J. & van OUDENAARDEN, A. Whole-organism clone tracing using single-cell sequencing. *Nature* **556**, 108-112. ISSN : 0028-0836, 1476-4687 (avr. 2018).
196. SCHMITZ, J., TÄUBER, S., WESTERWALBESLOH, C., von LIERES, E., NOLL, T. & GRÜNBERGER, A. *Development and application of a cultivation platform for mammalian suspension cell lines with single-cell resolution (MaSC)* en. preprint (Bioengineering, juill. 2020).
197. MULAS, C., HODGSON, A. C., KOHLER, T. N., AGLEY, C. C., HUMPHREYS, P., KLEINE-BRÜGGENEY, H., HOLLFELDER, F., SMITH, A. & CHALUT, K. J. Microfluidic platform for 3D cell culture with live imaging and clone retrieval. en. *Lab on a Chip* **20**. Number : 14 Reporter : Lab on a Chip, 2580-2591. ISSN : 1473-0197, 1473-0189 (2020).
198. KIMMERLING, R. J., LEE SZETO, G., LI, J. W., GENSHAFT, A. S., KAZER, S. W., PAYER, K. R., de RIBA BORRAJO, J., BLAINEY, P. C., IRVINE, D. J., SHALEK, A. K. & MANALIS, S. R. A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. en. *Nature Communications* **7**. Number : 1 Reporter : Nature Communications. ISSN : 2041-1723 (avr. 2016).
199. AGGARWAL, C. C., HINNEBURG, A. & KEIM, D. A. *On the Surprising Behavior of Distance Metrics in High Dimensional Space* in *Database Theory — ICDT 2001* (éd. VAN DEN BUSSCHE, J. & VIANU, V.) (Springer Berlin Heidelberg, Berlin, Heidelberg, 2001), 420-434. ISBN : 978-3-540-44503-6.
200. HUIZING, G.-J., PEYRÁ©, G. & CANTINI, L. Optimal Transport improves cell-cell similarity inference in single-cell omics data. *Bioinformatics (Oxford, England)*, btac084. ISSN : 1367-4811 (14 fév. 2022).

201. ZREIKA, S., FOURNEAUX, C., VALLIN, E., MODOLO, L., SERAPHIN, R., MOUSSY, A., VENTRE, E., BOUVIER, M., OZIER-LAFONTAINE, A., BONNAFFOUX, A., PICARD, F., GANDRILLON, O. & GONIN-GIRAUD, S. Evidence for close molecular proximity between reverting and undifferentiated cells. *BMC Biology* **20**, 155. ISSN : 1741-7007 (déc. 2022).
202. TOH, K., SAUNDERS, D., VERD, B. & STEVENTON, B. Zebrafish Neuromesodermal Progenitors Undergo a Critical State Transition in vivo. *bioRxiv* (2022).
203. VALLS, E., SÁNCHEZ-MOLINA, S. & MARTÁNEZ-BALBÁS, M. A. Role of Histone Modifications in Marking and Activating Genes through Mitosis *. *Journal of Biological Chemistry* **280**. Publisher : Elsevier. ISSN : 0021-9258, 1083-351X (30 déc. 2005).
204. TEVES, S. S., AN, L., HANSEN, A. S., XIE, L., DARZACQ, X. & TJIAN, R. A dynamic mode of mitotic bookmarking by transcription factors. *eLife* **5**, e22280. ISSN : 2050-084X (19 nov. 2016).
205. FERREIRA, R., OHNEDA, K., YAMAMOTO, M. & PHILIPSEN, S. GATA1 Function, a Paradigm for Transcription Factors in Hematopoiesis. *Molecular and Cellular Biology* **25**, 1215-1227. ISSN : 0270-7306 (fév. 2005).
206. KADAUKE, S., UDUGAMA, M. I., PAWLICKI, J. M., ACHTMAN, J. C., JAIN, D. P., CHENG, Y., HARDISON, R. C. & BLOBEL, G. A. Tissue-specific Mitotic Bookmarking by Hematopoietic Transcription Factor GATA1. *Cell* **150**, 725-737. ISSN : 0092-8674 (17 août 2012).
207. BELLEC, M., DUFOURT, J., HUNT, G., LENDEN-HASSE, H., TRULLO, A., ZINE EL AABIDINE, A., LAMARQUE, M., GASKILL, M. M., FAURE-GAUTRON, H., MANNERVIK, M., HARRISON, M. M., ANDRAU, J.-C., FAVARD, C., RADULESCU, O. & LAGHA, M. The control of transcriptional memory by stable mitotic bookmarking. *Nature Communications* **13**, 1176. ISSN : 2041-1723 (4 mars 2022).
208. KIM, Y. W., KANG, Y., KANG, J. & KIM, A. GATA-1-dependent histone H3K27 acetylation mediates erythroid cell-specific chromatin interaction between CTCF sites. *The FASEB Journal* **34**. _eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1096/fj.202001526R>, 14736-14749. ISSN : 1530-6860 (2020).
209. PALOZOLA, K. C., DONAHUE, G. & ZARET, K. S. EU-RNA-seq for in vivo labeling and high throughput sequencing of nascent transcripts. *STAR Protocols* **2**, 100651. ISSN : 2666-1667 (17 sept. 2021).

210. PALOZOLA, K. C., LERNER, J. & ZARET, K. S. A changing paradigm of transcriptional memory propagation through mitosis. *Nature reviews. Molecular cell biology* **20**, 55-64. ISSN : 1471-0072 (jan. 2019).
211. SOKOLIK, C., LIU, Y., BAUER, D., MCPHERSON, J., BROEKER, M., HEIMBERG, G., QI, L. S., SIVAK, D. A. & THOMSON, M. Transcription Factor Competition Allows Embryonic Stem Cells to Distinguish Authentic Signals from Noise. en. *Cell Systems* **1**, 117-129. ISSN : 24054712 (août 2015).
212. KAYA-OKUR, H. S., WU, S. J., CODOMO, C. A., PLEDGER, E. S., BRYSON, T. D., HENIKOFF, J. G., AHMAD, K. & HENIKOFF, S. CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nature Communications* **10**, 1930. ISSN : 2041-1723 (déc. 2019).
213. VIHervaara, A., MAHAT, D. B., HIMANEN, S. V., BLOM, M. A., LIS, J. T. & SISTONEN, L. Stress-Induced Transcriptional Memory Accelerates Promoter-Proximal Pause-Release and Decelerates Termination over Mitotic Divisions. *Molecular cell* **81**, 1715-1731.e6. ISSN : 1097-2765 (15 avr. 2021).
214. PASCUAL-GARCIA, P., LITTLE, S. C. & CAPELSON, M. Nup98-dependent transcriptional memory is established independently of transcription. *eLife* **11** (éd. SINGER, R. H. & STRUHL, K.) Publisher : eLife Sciences Publications, Ltd, e63404. ISSN : 2050-084X (15 mars 2022).