

THESE DE DOCTORAT DE

L'UNIVERSITE DE NANTES

ECOLE DOCTORALE N° 605

Biologie Santé

Spécialité : *Bioinformatique*

Par

Dimitri MEISTERMANN

Modélisation du développement préimplantatoire humain à partir de données de transcriptome de cellule unique

Thèse présentée et soutenue à Nantes, le 13 mars 2020

Unité de recherche : UMR 1064 Centre de Recherche en Transplantation et Immunologie

Rapporteurs avant soutenance :

Constance CIAUDO Assistant professor, Inst. f. Molecular Health Sciences, ETH Zurich

Antonio RAUSELL Chargé de recherche, IMAGINE INSTITUTE

Composition du Jury :

Président : **Richard REDON** Directeur de recherche, L'unité de recherche de l'institut du thorax (UMR 1087), Université de Nantes

Examineurs : **Constance CIAUDO** Assistant professor, Inst. f. Molecular Health Sciences, ETH Zurich

Antonio RAUSELL Chargé de recherche, IMAGINE INSTITUTE

Olivier GANDRILLON Directeur de recherche, LBMC (UMR 5239), ENS Lyon

Dir. de thèse : **Jérémie BOURDON** Professeur des universités LS2N (UMR 6004), Université de Nantes

Co-dir. de thèse : **Laurent DAVID** MCU-PH, CRTI (UMR 1064), Université de Nantes

« L'ennui dans ce monde, c'est que les idiots sont sûrs d'eux et les gens sensés pleins de doutes. »

Bertrand Russell

Remerciements

Mes premiers remerciements vont d'abord à mes encadrants de thèse, Laurent et Jérémie. D'aussi loin que je me souviens j'ai toujours eu la passion de la Science, et j'ai eu très tôt l'objectif de travailler dans la recherche. Grâce à vous deux ce rêve qui semblait jadis bien lointain est devenu réalité. Grâce à vous deux je peux regarder en face mon enfance et lui dire : « on l'a fait ! ». Une deuxième raison pour vous remercier concerne la réalisation de ma thèse, vos qualités humaines et scientifiques m'ont permis d'être heureux dans mon travail.

Je remercie ensuite ma famille : ma mère et mon père, ma fratrie, Amandine et Jérémy, et ma grand-mère, Louissette. Vous n'avez jamais cessé de croire en moi et de m'encourager. Il est absolument évident que sans l'environnement que vous m'avez fourni, je ne serais jamais arrivé aussi loin.

Je remercie toute ma belle-famille : Annick, Gaby, Anne-Claire, Cédric, Mélanie, Johannick, Marie et Rocco. L'énergie apportée par les repas de famille, que ce soit par les quantités ingérées ou par les instants de franche rigolade ont été un atout non négligeable dans la réalisation de ce travail.

Je remercie les biologistes qui ont travaillé avec moi en étroite collaboration sur les projets de l'équipe : Stéphanie, Julie, Sophie, Alexandre, Arnaud, Betty et Gaël. Sans vous, le biologiste qui est en moi serait mort à petit feu. Vous avez fait abstractions de ma formation de bioinformaticien et m'avez parlé comme si j'étais un des vôtres, ce qui avec le recul est primordiale pour mon parcours scientifique.

Je remercie l'équipe Combi du LS2N, particulièrement Yohann qui m'a aidé à créer des scripts durant ma première année de thèse. Vous m'avez toujours accueilli sans broncher, et vos conseils d'analyses se sont toujours avérés précieux. Ceci était tout particulièrement vrai pour le début de ma thèse, où vous étiez les seules personnes que je côtoyais qui pouvaient critiquer ma méthodologie en statistique.

Je remercie les membres de mon laboratoire, le CRTI. Tout d'abord les chercheurs de l'équipe : Carole, Ignacio, Mathieu, Jérôme. Vous avez très vite placé votre confiance en moi, que ce soit pour des collaborations, ou pour votre écoute lors de mes longues sessions de labmeeting où j'essayais de distiller, lentement mais sûrement, une culture de la bioinformatique. Je remercie aussi tous les collègues et anciens collègues qui sont devenues des amis proches : Raphaël, Lucas, Mathieu, Eros, Estelle, Magalie, Sophie et Nicolas. Notre relation n'est pas restée que professionnelle pour rien : vous m'êtes tout à fait sympathique.

Je remercie la plateforme IPS : Anne, Caroline, Quentin et les autres. Pour certains vous m'avez accueilli pour la première fois en stage de master 1, c'est-à-dire il y a fort longtemps. Votre gentillesse a toujours été sans faille à mon égard

Je remercie le centre d'assistance médicale à la procréation de Nantes. Et spécialement Thomas, qui malgré son emploi du temps résolument chargé, est toujours prêt à répondre à mes questions. Je suis admiratif de la synergie que vous formez avec Laurent, et de la

qualité de la science qui en découle. Accessoirement, sans Thomas, je n'aurais pas eu de financement de thèse. Pour cet acte de confiance, encore merci.

Je remercie l'UMR Phan, et particulièrement Valérie, la directrice de thèse de ma chère et tendre, ainsi que Julie, Anne-Lise, Diane, Thomas, Valentine, Marine, vous êtes devenu très vite une belle brochette d'amis.

Je remercie la communauté de la bioinformatique, et particulièrement française. Ce sont des gens désintéressés et passionnés. Il existe une vraie entraide, et un vrai esprit de communauté qui se fait rare aujourd'hui dans la recherche. Pour les nombreuses heures gagnées après discussion ou lectures de tutoriaux, chapeau !

Je remercie mes anciens stagiaires officiels ou officieux : Hayat, Johanna, Valentin, Simon, Thierno et bien d'autre ; qui ont essuyé les plâtres de mon apprentissage de la pédagogie, mais qui sont tous restés des amis.

Je souhaite aussi remercier toute ma promo de master, particulièrement Ulysse, Victor, Jennifer, Anne-Sophie, Xavier, Kenzo, Clara. On a vécu des moments... inoubliables, et dont on rigolera pendant longtemps encore. En dehors de l'aspect poilant de vos personnalités. Vous avez des qualités scientifiques remarquables, et j'ai dû me battre pour rester à votre niveau. Nous formons un réseau solidaire, dont je suis certain que chacun d'entre nous y gagnera.

Julien, Anthony, vous faites partie de mes plus vieux et plus proches amis. D'aucun dirait que vous êtes mes « bestah ». Je ne pourrais donc jamais vous remercier assez pour m'avoir supporté depuis si longtemps. Vous êtes mes mètres étalons de la gentillesse !

Je souhaite aussi remercier tous ceux que j'oublie mais qui méritent quand même leur présence ici. Vous êtes sorti de ma mémoire, mais pas de mon cœur !

À celle sans qui rien de tout ça ne serait possible. Morgane, tu m'as tout donné. Je ne sais pas très bien ce que tu me trouves mais saches que grâce à ton attention et ton soutien indéfectible, je suis le plus heureux des Homo sapiens. Que de choses ont changées depuis nos premières amourettes, et même, depuis le début de nos thèses : nous sommes maintenant mariés et tu m'as donné un fils, Arthur. Un petit bébé si mignon qu'il est difficile de ne pas le pouponner. C'est aussi une nouvelle raison de vivre : un petit être à laquelle tout apprendre. Morgane, Arthur, je vous aime.

Je dédie ma thèse à la mémoire de mon grand-père, simple agriculteur, et pourtant l'une des personnes les plus intelligentes que je n'ai jamais rencontré. Tu m'as appris à toujours voir plus loin que le bout de mon nez.

Table des matières

Préambule : introduction vulgarisée	1
1 Introduction.....	4
1.1 Du zygote à la nidation : le développement préimplantatoire	4
1.1.1 La fécondation <i>in vitro</i> : méthode principale pour permettre aux couples infertiles d'avoir un enfant.....	6
1.1.2 Les mécanismes à l'origine des premières spécifications de lignées cellulaires.....	11
1.1.2.1 La première spécification : ICM et TE.....	11
1.1.2.2 La seconde spécification : EPI et PrE	13
1.1.3 De la difficulté et de la nécessité et d'étudier le développement préimplantatoire humain	17
1.1.3.1 Ethique et recherche sur l'embryon humain	17
1.1.3.2 Les modèles permettent d'étudier l'embryon humain en limitant leur utilisation	18
1.2 Etudier l'expression des gènes en masse : du transcriptome à la cellule unique	20
1.2.1 Estimer le transcriptome à partir de RNA-Seq	20
1.2.1.1 Le RNA-Seq est un ensemble hétérogène de méthodes de transcriptomiques.....	22
1.2.1.2 L'avènement du séquençage de transcriptome en cellule unique.....	25
1.2.1.3 L'analyse primaire permet de passer des reads à la table de séquençage	28
1.2.2 L'analyse de la table de comptage de l'expression est un défi mathématique	30
1.2.2.1 Aux origines de la complexité du RNA-Seq	30
1.2.2.2 La distribution de la table de comptage : de la loi de Poisson à ZINB .	33
1.2.2.3 La table de comptage brut est impropre à l'analyse de l'expression..	37
1.2.2.3.1 Filtrer la table de comptage.....	37
1.2.2.3.2 Normaliser des données de RNA-Seq	40
1.2.2.3.3 Corriger les effets de batch	45
1.2.2.4 Inférer une liste de gènes différentiellement exprimés.....	46
1.2.2.5 P-valeur et analyses multivariées.....	49
1.2.2.6 Sélectionner des gènes d'intérêt	51
1.2.2.7 Les méthodes de réductions de dimension	52
1.2.2.8 Résoudre le problème de classification par le partitionnement.....	56
1.2.2.8.1 Choisir une métrique de distance	58

1.2.2.8.2 Choisir un algorithme de partitionnement.....	59
1.2.2.9 L'enrichissement fonctionnel.....	62
1.2.2.10 Etudier les phénomènes de spécification à partir de scRNA-Seq	65
1.2.2.10.1 Inférer des trajectoires cellulaires à partir du transcriptome	65
1.2.2.10.2 Estimer la vélocité du transcriptome: une nouvelle façon de prédire le destin cellulaire.....	68
1.2.2.11 Le défi de la modélisation du transcriptome	68
2 Manuscrit #1 : Parallel derivation of isogenic human primed and naive induced pluripotent stem cells.....	70
2.1 Contexte	70
2.1.1 Pluripotence naïve et amorcée	70
2.1.2 Motivation de l'étude	71
2.2 Manuscrit	73
2.3 Discussion	98
2.4 Conclusion.....	99
2.5 Perspectives.....	100
3 Manuscrit #2 : Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophectoderm specification	101
3.1 Contexte	101
3.1.1 Le développement préimplantatoire humain par le prisme du scRNA-Seq	101
3.1.2 Motivation de l'étude	104
3.2 Manuscrit	106
3.3 Discussion.....	135
3.4 Conclusion.....	142
3.5 Perspectives.....	143
4 Références	145
Tableau annexe : résumé des outils utilisés	160
Liste des publications	162
Articles en premier auteur	162
Autres articles en co-auteur	162
Rapports préalables à la soutenance	164

Préambule : introduction vulgarisée

Durant les premières 24 h après la fécondation, l'embryon humain est constitué d'une cellule. Nous passons de cette cellule unique à 100 mille milliards à l'âge adulte, réparties entre plusieurs centaines de types cellulaires comme des neurones ou les globules blancs. La mise en place de tous ces types cellulaires ne se réalise pas en une étape, si bien qu'il existe un arbre « généalogique » des cellules : il existe une multitude de types cellulaire intermédiaires entre la cellule œuf et le neurone. Si l'on remonte cet arbre, nous tombons sur des cellules qui possèdent le potentiel pour se différencier en nouveaux types cellulaires : ce sont les cellules souches. Une des façons de classer ces cellules souches est d'examiner leur potentiel de différenciation. Ainsi certaines cellules souche ne peuvent se différencier qu'en un type cellulaire spécifique, on parle d'unipotence. D'autres cellules souches moins spécialisées peuvent se différencier en plusieurs types cellulaires : c'est la multipotence. C'est le cas par exemple des cellules souches neurales, qui peuvent se différencier en neurones ou astrocytes. Les cellules multipotentes peuvent-elles même se différencier en cellule souches multi ou unipotentes. Enfin, les cellules souches pluripotentes sont capables de se différencier en n'importe quel type cellulaire du de l'humain adulte. Les cellules souches pluripotentes sont à la base de l'arbre « généalogique » des types cellulaires du fœtus.

Les cellules souche pluripotentes se trouvent dans l'embryon quelques jours après la fécondation. L'embryon est alors dans un stade que l'on qualifie de blastocyste. C'est le blastocyste qui va s'accrocher à l'utérus de la mère lors de la nidation, à environ 7 jours après la fécondation chez l'Homme. Cette première période du développement, de la fécondation à la nidation, est qualifiée de développement préimplantatoire. Le blastocyste qui va s'implanter est composé d'une sphère de cellule à l'extérieur, et à l'intérieur, d'une cavité et d'une masse de cellule. Les cellules externes correspondent au trophoctoderme et contribueront au placenta, tandis que la masse cellulaire interne contient l'épiblaste qui est uniquement constitué de cellule souches pluripotentes. L'épiblaste donnera donc le fœtus.

L'épiblaste avant et l'épiblaste après la nidation sont pluripotents, cependant ils semblent être dans un état cellulaire différent. Par exemple, dans les cellules femelles adultes normales (qui possèdent deux chromosomes X), un des deux chromosomes X est inactif : il ne participe pas à la machinerie cellulaire. C'est le cas dans l'épiblaste

postimplantatoire (après la nidation). En revanche, dans l'épiblaste préimplantatoire, les deux chromosomes X sont encore actifs. Dans le cas de l'épiblaste préimplantatoire les cellules souches pluripotentes sont dites naïves, et pour l'épiblaste postimplantatoire on parle de pluripotence amorcée. La connaissance de l'existence de ces deux états de la pluripotence est récente : mieux comprendre leurs différences est un enjeu du domaine de recherche sur les cellules souches. Pour travailler sur ce problème nous avons à notre disposition plusieurs outils. Tout d'abord nous pouvons utiliser des cellules souches pluripotentes in vitro. Ces cellules ne proviennent pas directement d'embryons humains et permettent donc de ne pas détruire un embryon humain à chaque fois que l'on a besoin de cellules souches pluripotentes. Il existe deux types de cellules pluripotentes in vitro. Tout d'abord les cellules souches embryonnaires, qui proviennent d'un blastocyste dont la masse cellulaire interne a été mise en culture. Une fois que la culture cellulaire a démarrée, les cellules peuvent être multipliées indéfiniment. La deuxième méthode d'obtention de cellule pluripotente est plus récente. Dans ce cas on reprogramme des cellules adultes, par exemple de peau, en cellules souches pluripotentes. Ces cellules sont qualifiées de cellules souches pluripotentes induites. La reprogrammation s'effectue en forçant la cellule adulte à exprimer des gènes normalement spécifiques des cellules pluripotentes.

En effet, toutes les cellules de notre corps contiennent les mêmes gènes, mais tous ne sont pas exprimés en même temps. Nous pouvons prendre la métaphore d'un piano : toutes les cellules disposent d'un piano avec les mêmes notes, mais chaque type cellulaire a une mélodie qui lui propre. Un gène exprimé émet des molécules dans la cellule que l'on appelle des transcrits. Le transcriptome représente l'ensemble des transcrits et donc l'ensemble des gènes exprimés dans une cellule. Pour comparer les caractéristiques de la pluripotence naïve et amorcée, j'ai comparé les transcriptomes de cellules souches pluripotentes induites naïves et amorcée, et d'épiblaste préimplantatoire. Ce travail constitue le premier manuscrit présenté dans cette thèse.

Le deuxième manuscrit constitue mon projet principal, et porte de façon plus globale sur le développement préimplantatoire. Le but d'en comprendre la hiérarchie des événements, en partant des transcriptomes de cellules d'embryon à différent stade de cette période. Un des événements majeurs étudié est la première spécification cellulaire, c'est-à-dire la toute base de l'arbre « généalogique » des cellules. Les cellules de l'embryon vont alors perdre leur totipotence, qui est la capacité pour chacune de ces

cellules à pouvoir former un nouvel embryon à elle seule. Le trophoctoderme, que l'on peut qualifier de cellule souche de la lignée placentaire, et la masse cellulaire interne pluripotente vont émerger de cette première spécification. Chez la souris, cette première spécification a lieu pendant l'état qui précède le blastocyste, la morula. Cependant il existe des preuves que la première spécification ne se déroule pas de façon identique chez l'Homme.

Avant 2007, on ne pouvait étudier l'expression d'un nombre restreint de gène à la fois, nous n'avions pas accès au transcriptome complet. Le RNA-Seq (ou séquençage de transcriptome) est la méthode qui a résolu ce problème. Cependant, dans les premiers protocoles de RNA-Seq, il fallait plusieurs dizaines de milliers de cellules pour obtenir un seul transcriptome, ce qui correspond à plusieurs milliers d'embryons. Il a donc fallu l'apparition du single-cell RNA-Seq (séquençage de transcriptome en cellule unique) pour pouvoir obtenir un transcriptome par cellule. Le RNA-Seq, et particulièrement en cellule unique, génère une quantité phénoménale de données. L'analyse de ces données requiert donc des compétences en entre la biologie, les statistiques et l'informatique, qui sont incluses dans ma discipline : la bioinformatique.

L'apparition du single-cell RNA-Seq permis de faire avancer significativement l'étude des destins cellulaire, c'est-à-dire les mécanismes par lesquelles une cellule souche décide de sa différenciation. L'utilisation du single-cell RNA-Seq nous a ainsi permis de retracer les routes que prennent les cellules lors du développement préimplantatoire, et quels gènes étaient impliqués dans ces choix. Nous avons par exemple pu mettre en évidence que le trophoctoderme en contact avec la masse cellulaire interne évolue plus vite que le reste du trophoctoderme. Tous ces résultats sont disponibles sur une application web, ce qui permet à n'importe quelle équipe qui travaille sur le développement préimplantatoire de pouvoir tester ou formuler rapidement de nouvelles hypothèses. La compréhension du développement préimplantatoire est un enjeu majeur pour l'amélioration du taux de succès des fécondations in vitro (FIV). En effet, 2/3 des FIV n'amènent pas à une grossesse, et un mauvais développement préimplantatoire semble être en cause dans une partie des échecs.

1 Introduction

1.1 Du zygote à la nidation: le développement préimplantatoire

Le cycle de vie d'un individu chez les vertébrés commence généralement par la fécondation d'un ovule par un spermatozoïde. Chez les mammifères placentaires, une interface entre la mère et l'embryon élaborée se met en place peu après la fécondation et va permettre au fœtus de se développer jusqu'à la naissance. Le développement préimplantatoire correspond à la fenêtre de temps entre la fécondation et le début de la mise en place de cette interface materno-fœtale. Le développement préimplantatoire prend donc fin avec la nidation (*implantation*) de l'embryon dans la muqueuse utérine (**Figure 1a**).

Plusieurs événements moléculaires majeurs vont avoir lieu pendant cette période du développement préimplantatoire. Le premier est l'activation du génome embryonnaire (ou *Zygotic Genome Activation, ZGA*), qui s'achève au stade 8 cellules chez l'Homme (Newport and Kirschner, 1982; Schulz and Harrison, 2019). À la suite de la ZGA, plusieurs événements de différenciation de lignées cellulaires, ou spécifications vont avoir lieu. Le premier choix de destin cellulaire, appelé « première spécification », entraîne la ségrégation de la masse cellulaire interne (ICM) au centre de l'embryon et du trophoctoderme (TE) à sa périphérie. Un second événement de spécification au niveau de l'ICM aboutira à la formation de l'épiblaste (EPI) et de l'endoderme primitif (PrE) (**Figure 1b**).

Ainsi, trois lignées cellulaires distinctes sont spécifiées au cours du développement préimplantatoire humain et sont présentes au moment de l'implantation. Le TE est l'origine du placenta fœtal (**Figure 1c**), et de la partie externe de l'amnios (la membrane protectrice de la cavité amniotique). Le TE va aussi contribuer en grande partie au chorion (la membrane protectrice de l'embryon). Le PrE donnera la membrane qui délimite le sac vitellin, l'allantoïde. Il apporte également une contribution mineure aux tissus du fœtus (Gardner, 1982). L'EPI, le type cellulaire pluripotent, va donner les trois feuilletts embryonnaires et donc l'immense majorité des tissus du fœtus. Il est aussi à l'origine de la membrane interne de l'amnios.

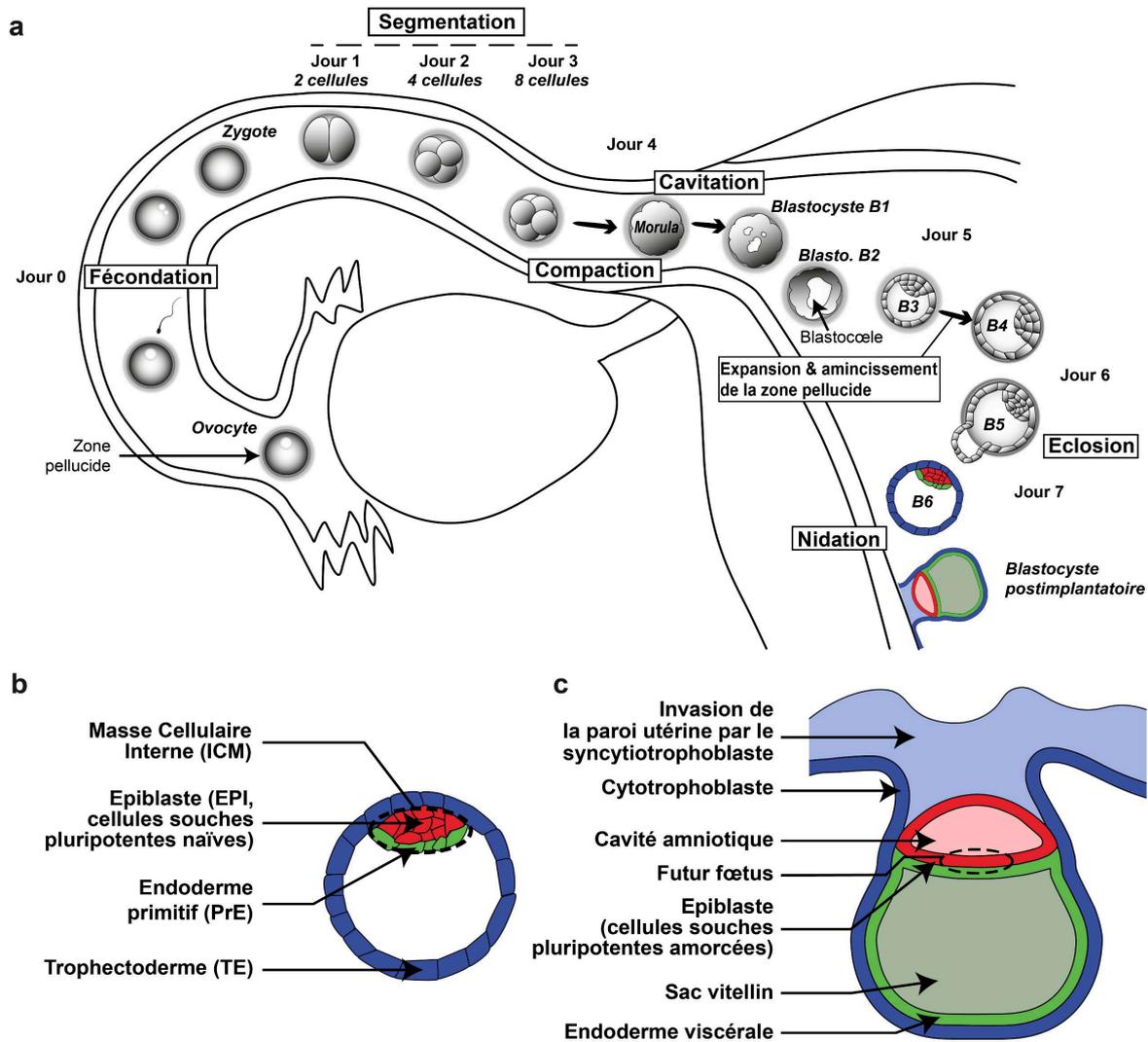


Figure 1 : Le développement préimplantatoire humain. a. Les différentes étapes du développement préimplantatoire : 1/ L'ovule est fécondé par un spermatozoïde dans les trompes de Fallope. 2/ Le zygote entre dans une phase de segmentation. Le nombre de cellules double alors toutes les 24h, tandis que la taille de l'embryon n'augmente pas. 3/ La compaction de l'embryon entraîne la formation de la morula, les cellules s'agglomèrent et deviennent difficiles à distinguer. 4/ Une cavité, le blastocœle, apparaît au sein de l'embryon. Cette étape de cavitation marque le début du stade blastocyste. 5/ L'expansion du blastocœle fait gonfler l'embryon, la zone pellucide s'amincit. 6/ Cet amincissement prépare l'éclosion du blastocyste, c'est-à-dire la sortie de l'embryon de la zone pellucide. 7/ La dernière étape du développement préimplantatoire est la nidation de l'embryon dans la muqueuse utérine. b-c. Lignées cellulaires présentes dans le blastocyste préimplantatoire (b) et postimplantatoire (c).

1.1.1 La fécondation *in vitro*: méthode principale pour permettre aux couples infertiles d'avoir un enfant

L'application clinique la plus immédiate de l'étude du développement préimplantatoire humain concerne la fécondation *in vitro* (FIV). La FIV offre la meilleure chance de conception pour de nombreux couples subfertiles. La FIV nécessite une ponction des ovocytes après une période de stimulation ovarienne., les ovocytes recueillis sont ensuite mis en présence des spermatozoïdes. Les ovocytes fécondés sont mis en culture, puis les embryons obtenus peuvent-être transférés dans l'utérus maternel ou bien être cryopréservés en vue d'un transfert différé. Malgré l'essor des techniques de FIV, les taux de réussite en FIV restent limités. Selon les enquêtes nationales, le taux moyen de fécondité par FIV varie entre 20 et 30% (Mansour et al., 2014). Ces résultats sont imputables à plusieurs facteurs :

- le faible niveau de fécondité de l'espèce humaine (Stephen and Chandra, 1998) ;
- la qualité des gamètes, en fonction des pathologies maternelles ou paternelles, de l'exposition parentale à des toxiques, de l'âge maternel, des protocoles de stimulation ovarienne ;
- la réceptivité utérine et le développement placentaire ;
- la qualité des embryons à transférer.

Améliorer ce dernier point est un enjeu majeur pour les laboratoires des centres clinico-biologiques de FIV. Actuellement et de façon consensuelle à travers le monde, la qualité embryonnaire repose sur l'évaluation de critères morphologiques au cours du développement des embryons. Pour ce faire, les embryons sont évalués, après observation sous microscope à des temps précis (The Istanbul consensus workshop on embryo assessment, 2011).

L'amélioration des conditions de culture (étuves tri-gaz, milieux) a permis de pousser les embryons en culture jusqu'au stade de blastocyste avec un transfert embryonnaire au 5^{ème} jour de développement (J5). Cette stratégie de transfert après une culture prolongée présente les avantages :

- de sélectionner l'embryon le plus viable (Gardner et al., 1998),

- de respecter la synergie entre l'utérus maternel et l'embryon, puisque physiologiquement l'embryon atteint la cavité utérine à partir du 4ème jour de développement (Croxatto et al., 1972) et que la pulsatilité utérine à J5 est plus favorable au maintien de l'embryon (Fanchin et al., 2001),
- de privilégier le transfert mono-embryonnaire (Glujovsky et al., 2016), limitant ainsi les risques inhérents aux grossesses multiples (fausses couches, prématurité, retard de croissance, syndrome transfuseur-transfusé).

À ce stade blastocyste, à J5, le référentiel utilisé pour l'évaluation des embryons est la classification de Gardner (Gardner and Schoolcraft, 1999). Les blastocystes sont évalués sur trois aspects différents :

1. le développement du blastocyste tenant compte de la cavitation, de l'expansion ou de l'éclosion du blastocyste. Gradé de 1 : blastocyste non expansé avec début de cavitation, à 6 : blastocyste totalement éclos (**Figure 1a**) ;
2. la qualité de la masse cellulaire interne (ICM). Gradée de A : ICM compacte contenant de nombreuses cellules, à C : quasi-absence d'ICM ;
3. la qualité du trophoctoderme (TE). Gradé de A : TE cellulaire formant un épithélium cohésif, à C : TE limité à quelques cellules larges.

L'objectif de cette classification est de sélectionner l'embryon ayant le meilleur potentiel de développement et donc la plus grande probabilité d'aboutir à une naissance. S'il est bien établi que le score combinant les trois paramètres est prédictif du devenir de l'embryon, la capacité de chaque paramètre à prédire de façon indépendante le devenir clinique des embryons est moins claire. Or, pour choisir l'embryon à transférer parmi plusieurs embryons ayant un score équivalent (par combinaison des trois paramètres), il est nécessaire de comprendre la contribution de chaque paramètre.

Plusieurs études portant sur des transferts mono-embryonnaires ont été publiées. Elles ont permis de mettre en évidence l'importance du grade de l'expansion du blastocyste au moment du transfert (Ahlström et al., 2013; Du et al., 2016; Subira et al., 2016; Thompson et al., 2013; Van den Abbeel et al., 2013), mais peinent à

identifier la contribution du grade de l'ICM ou du grade du TE comme marqueur prédictif de l'implantation des embryons, en raison de résultats contradictoires (Subira et al., 2016). Cependant des études récentes tendent à souligner le pouvoir prédictif de l'évaluation du TE pour le devenir des embryons transférés (Ahlström et al., 2011, 2011; Chen et al., 2014). La plus récente (Zhao et al., 2018) se restreint aux transferts mono-embryonnaires euploïdes. Cela permet d'éliminer, dans le classement des embryons, le facteur de confusion des anomalies chromosomiques. Les résultats montrent des taux de naissances similaires pour des blastocystes avec un TE de grade A ou B ou des blastocystes avec une ICM grade A. Ces considérations amènent à revoir les stratégies de sélection de l'embryon à transférer en insistant sur l'importance de l'évaluation morphologique du TE.

Il existe une variabilité inter et intra-observateurs dans la graduation des embryons en lecture « conventionnelle », rendant l'évaluation morphologique des embryons peu sensible et peu reproductible, avec une valeur prédictive positive modeste (Arce et al., 2006; Assín et al., 2009; Paternot et al., 2011). C'est pourquoi il est fondamental de rechercher d'autres marqueurs spécifiques de la qualité embryonnaire en FIV et plus particulièrement la qualité du TE des embryons. Cette étape nécessite de caractériser moléculairement le TE humain durant le développement embryonnaire pré-implantatoire. D'une façon plus globale, le manque de compréhension du développement préimplantatoire humain est une limitation à la performance des méthodes d'assistance à la reproduction.

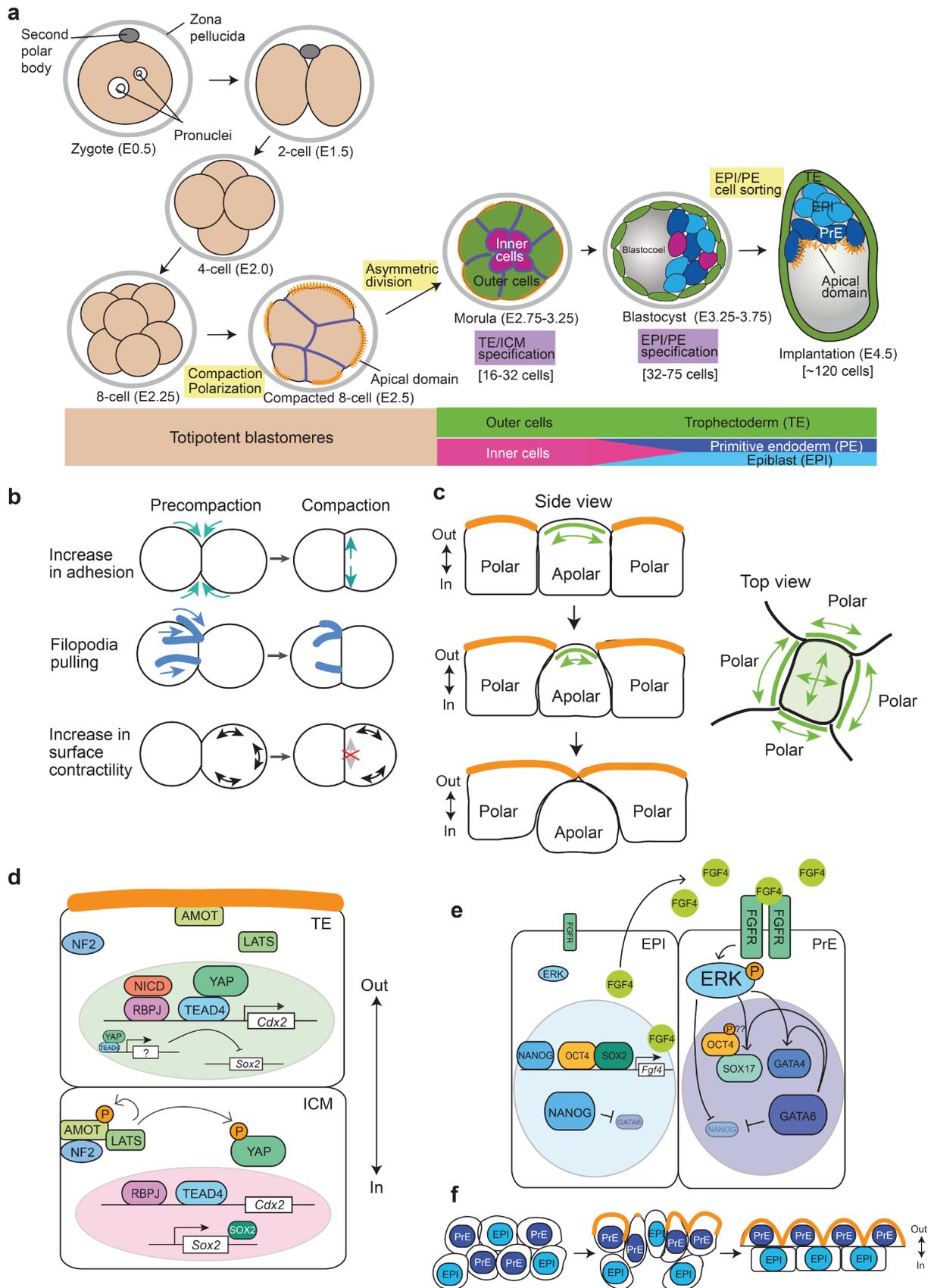


Figure 2 : Régulation de la première et de la deuxième spécification chez la souris. Figures issues de Chazaud et Yamanaka, 2016. **a.** Schéma des changements morphologiques et des étapes de différenciation des lignées cellulaires du développement embryonnaire préimplantatoire de la

souris. Les barres colorées montrent la progression séquentielle de la lignée depuis le blastomère totipotent jusqu'aux trois premières lignées : trophoctoderme (vert), épiblaste (bleu clair) et endoderme primitif (bleu foncé). Les événements morphogénétiques importants sont mis en évidence en jaune. Les événements de spécification des lignées sont surlignés en violet. Les lignes orange indiquent les domaines apicaux des cellules. **b.** Modèles de compaction au stade de 8 cellules. Dans le modèle classique (en haut), une augmentation de l'adhérence des cellules déclenche une augmentation des contacts entre les cellules (flèches vertes). Dans un deuxième modèle (au milieu), des filopodes dépendant de CDH1 (bleu) créent la force de traction (flèches bleues) pour l'aplatissement des cellules. Dans le dernier modèle (en bas), une augmentation de la contractilité de surface à l'interface cellule-milieu (flèches à double tête) contrôle le compactage. Dans ce modèle, les contacts cellule-cellule à médiation CDH1 empêchent une augmentation de la contractilité au niveau des contacts. **c.** La constriction apicale des cellules peut entraîner le positionnement extérieur/intérieur des cellules polaires/apolaires. Après des divisions asymétriques, certaines cellules apolaires sont présentes en position externe (vue de côté, en haut). La surface de ces cellules apolaires présente un niveau plus élevé de contractilité de l'actomyosine pour initier la constriction apicale (flèches vertes à double tête). Ceci augmente la pression intracellulaire de la cellule apolaire pour initier la formation de courbure aux contacts cellule-cellule avec les cellules polaires environnantes (vue de côté, milieu). La taille de la surface sans contact devient progressivement plus petite, à la fois par la constriction apicale et par le rétrécissement de l'espace entre les cellules polaires environnantes par une augmentation de la contractilité de l'actomyosine au bord des cellules polaires (vue de dessus). La cellule apolaire est alors complètement internalisée et l'espace entre les cellules polaires est fermé (vue de côté, en bas). Les domaines apicaux des cellules polaires sont représentés en orange. **d.** Régulation de la spécification ICM/TE. Dans une cellule TE, la protéine AMOT non phosphorylée est localisée au domaine apical (orange) et ne forme pas de complexe actif avec les kinases NF2 et LATS. La protéine YAP n'est alors pas phosphorylée et peut se transloquer dans le noyau pour se lier à TEAD4 et ainsi activer des gènes cibles tels que *Cdx2*. En parallèle, la voie Notch (NICD et RBPJ κ) est activée pour contribuer à l'activation de l'expression de *Cdx2*, tandis que l'expression de *Sox2* est supprimée par des mécanismes inconnus dépendant de la voie Hippo. Dans une cellule ICM, cependant, AMOT est phosphorylée et forme un complexe actif avec les kinases NF2 et LATS pour phosphoryler YAP. La protéine YAP phosphorylée est séquestrée dans le cytoplasme, ce qui empêche l'expression de *Cdx2* dans l'ICM. **e-f** Régulation de la spécification EPI/PrE. **e.** Dans les cellules EPI (à gauche), un niveau plus élevé de NANOG en association avec OCT4 et SOX2 régule l'expression de *Fgf4* et réprime l'expression de *Gata6* ainsi que l'activité de la voie FGF, comme l'indique le faible niveau d'ERK phosphorylé. Dans les cellules PrE (à droite), le FGF4 sécrété par les cellules EPI active la signalisation du FGF. Cette signalisation FGF plus élevée stimule l'expression de *Gata6* et réprime l'expression de *Nanog*. La signalisation FGF et GATA6 activent conjointement les gènes cibles, tels que *Sox17* et *Gata4*. **f.** Selon leur spécification, les cellules EPI (bleu clair) et PrE (bleu foncé) se séparent, et le PrE subit une épithélialisation. Une fois que les progéniteurs de PrE ont atteint la surface de l'ICM, ils amorcent la formation du domaine apical (orange) et un épithélium de PrE se forme autour des cellules EPI regroupées.

1.1.2 Les mécanismes à l'origine des premières spécifications de lignées cellulaires

Les événements à l'origine des spécifications de lignées cellulaires durant le développement préimplantatoire restent mal définis chez l'Homme. La plupart des connaissances concernant ces événements proviennent de travaux chez la souris. Les quelques travaux réalisés chez l'Homme ont néanmoins permis de mettre en lumière plusieurs différences entre le développement préimplantatoire de l'Homme et de la souris (Blakeley et al., 2015; Boroviak et al., 2018; Niakan and Eggan, 2013; Rossant and Tam, 2017).

1.1.2.1 La première spécification : ICM et TE

Chez la souris, la première spécification, la ségrégation du TE et de l'ICM, se produit au stade morula et est contrôlée par un axe YAP/TEAD4/CDX2 (**Figure 2d**) (Chazaud and Yamanaka, 2016; Rossant and Tam, 2017)

Les lignées ICM et TE sont établies par expression réciproque des facteurs de transcription spécifiques pour chaque lignée tel que *Oct4* (ICM) et *Cdx2* (TE). Des taux variables des protéines OCT4 et NANOG ont été observés dans tous les noyaux d'embryons de souris au stade de 8 cellules au début de la compaction. Dans les embryons compactés au stade 8 cellules, les protéines OCT4, NANOG et CDX2 sont présentes dans tous les noyaux à des niveaux variables. Après la blastulation au stade 32 cellules, OCT4 et NANOG sont restreints aux cellules internes et *Cdx2* est exprimé spécifiquement dans les cellules externes. OCT4, NANOG et SOX2 maintiennent le réseau de pluripotence dans les cellules de l'ICM, alors que dans les cellules externes, *Oct4* est inhibé par CDX2 ce qui contribue à engager les cellules vers un destin TE.

Chez les embryons de souris, l'expression de *Cdx2* est maintenue à la suite de l'interaction du facteur de transcription TEAD4 et de son coactivateur YAP1. TEAD4, détecté dès le stade 4 cellules, est exprimé de façon ubiquitaire tout au long du développement préimplantatoire. YAP1 est détecté (faiblement) dans les noyaux des cellules dès le stade 4 cellules. YAP1 devient fortement détecté dans les noyaux

de tous les blastomères au stade 8 cellules avant la compaction. Au cours de la division suivante (de 8 à 16 cellules), qui correspond à la compaction, YAP1 reste nucléaire dans les cellules externes, tandis que les cellules internes excluent YAP1 des noyaux par un processus impliquant sa phosphorylation. Ensuite, p-YAP1 est dégradé dans les protéasomes. Cela se traduit par une inhibition de l'expression de *Cdx2* dans les cellules internes et, en conséquence, par une induction de *Sox2*.

La voie de signalisation Hippo est régulée par la ségrégation dépendante de la polarité de facteurs tels que la kinase LATS1/2 (phosphorylation de YAP1) et l'AMOT (localisée différenciellement dans les cellules internes et externes) (Hirate and Sasaki, 2014; Hirate et al., 2013). Un grand complexe protéique composé de la E-cadhérine, de la Beta-caténine, de l'alpha-caténine et d'AMOT est formé au niveau des sites de contact dans les cellules internes et active probablement la kinase LATS1 / 2 qui phosphoryle YAP (Korotkevich et al., 2017). Certaines cellules apolaires présentant des taux élevés de p-YAP1 cytoplasmique ont été découvertes en position externe au moment de la compaction (Anani et al., 2014; Hirate et al., 2015). Elles sont intériorisées ultérieurement, ce qui indique que la voie de signalisation Hippo est activée avant que la position de la cellule dans l'embryon ne soit définie. Cette observation suggère que c'est la polarité cellulaire qui détermine l'activité de la voie Hippo et non la position des cellules dans l'embryon (Alarcon and Marikawa, 2018; Anani et al., 2014) (**Figure 2d**).

Chez l'Homme, nous avons peu de données mécanistiques disponibles sur la première spécification. Néanmoins, certaines différences ont été observées par rapport au développement murin. Par exemple, des expériences d'immunofluorescence ont révélé l'absence de CDX2 au stade morula chez l'homme. CDX2 n'est détecté que plus tard, au stade blastocyste, et est restreint aux cellules du TE (Niakan and Eggan, 2013). À ce jour, nous ne disposons d'aucun marqueur dont l'expression hétérogène au stade morula témoignerait d'une initiation de la spécification du TE chez l'homme.

1.1.2.2 La seconde spécification : EPI et PrE

Chez la souris, la seconde spécification (**Figure 2e-f**), qui entraîne la différenciation des cellules de l'ICM en EPI et PrE, a lieu au stade du blastocyste. Cette spécification est régulée par un axe NANOG/FGF4/GATA6 (Chazaud and Yamanaka, 2016; Rossant and Tam, 2017). Il s'agit d'un processus à plusieurs étapes commençant par la spécification binaire EPI / PrE suivie par la maturation de lignage (c'est-à-dire l'établissement de réseaux de gènes spécifiques et différenciation au sein de chaque lignée) et la migration des cellules qui forment deux compartiments distincts : un cluster d'EPI et un épithélium de PrE.

Les facteurs de transcription NANOG et GATA6 sont les premiers marqueurs de l'EPI et du PrE, respectivement (Chazaud et al., 2006; Kurimoto et al., 2006). Les deux sont présents dans tous les blastomères au stade 8 cellules mais, à partir du stade ~32 cellules, les cellules de l'ICM expriment soit *Nanog*, conduisant à un destin EPI, soit *Gata6*, conduisant à un destin PrE (Guo et al., 2010). Cela donne lieu à un modèle d'expression «sel et poivre» mutuellement exclusif de NANOG et de GATA6 au jour embryonnaire (E) 3,75. Des expériences de traçage de lignées cellulaires (Chazaud et al., 2006; Kurimoto et al., 2006; Meilhac et al., 2009; Plusa et al., 2008; Xenopoulos et al., 2015) montrent que ce processus se produit dans des cellules individuelles de l'ICM de manière asynchrone autour de E3.0-E3.75 (Bessonard et al., 2014; Gerbe et al., 2008; Plusa et al., 2008). Par conséquent, quelques cellules coexprimant les deux facteurs de transcription peuvent encore être identifiées à ~E3.75, et l'émergence de cellules EPI NANOG⁺/GATA6⁻ est le premier signe connu du processus de spécification. Étant donné que les niveaux d'expression de *Nanog* et *Gata6* sont relativement élevés dès le stade 8 cellules (Guo et al., 2010), le mécanisme contrôlant la spécification implique probablement un processus de régulation post-transcriptionnelle. Au niveau protéique, la spécification EPI/PrE est corrélée à l'expression accrue de l'un des facteurs de transcription et la disparition de l'autre (Bessonard et al., 2014; Guo et al., 2010), révélant une dynamique distincte entre les niveaux d'ARN et les niveaux protéiques. Il semble également exister une répression mutuelle entre les deux facteurs de transcription : chez les mutants *Gata6*^{-/-}, toutes les cellules de l'ICM expriment uniformément NANOG sans

aucun marqueur PrE, indiquant l'acquisition d'un destin EPI (Bessonard et al., 2014; Schrode et al., 2014), et inversement, chez les mutants *Nanog*^{-/-}, toutes les cellules ICM expriment GATA6 (Frankenberg et al., 2011). Cela suggère que NANOG et GATA6 pourraient se réprimer mutuellement, directement au niveau transcriptionnel, comme le suggèrent leurs sites de liaison identifiés dans les études d'immunoprécipitation de la chromatine de cellules souches embryonnaires pluripotentes (Singh et al., 2007) et de cellules d'endoderme extra-embryonnaire induit (iXEN) (Wamaitha et al., 2015).

Une série d'études a révélé que la signalisation FGF joue un rôle essentiel lors de la formation de la lignée PrE (**Figure 2e**). FGF4 marque spécifiquement les cellules EPI au stade blastocyste (Frankenberg et al., 2011; Guo et al., 2010; Kurimoto et al., 2006; Ohnishi et al., 2014) et n'est pas exprimé dans les mutants *Nanog*^{-/-} (Frankenberg et al., 2011). *Fgfr2*, le récepteur de FGF4, est exprimé dans toutes les cellules de l'ICM précoces à E3.25 avant d'être restreint aux cellules PrE à E3.5. Le blocage de la signalisation FGF est suffisant pour que toutes les cellules ICM acquièrent un destin EPI (Chazaud et al., 2006; Kang et al., 2013; Krawchuk et al., 2013; Nichols et al., 2009; Yamanaka et al., 2010). À l'inverse, l'addition de FGF4 en excès est suffisante pour différencier toutes les cellules ICM en PrE (Yamanaka et al., 2010). Il est intéressant de noter que *Gata6* est exprimé jusqu'au début du stade blastocyste chez les mutants *Fgf4*^{-/-} (Kang et al., 2013; Krawchuk et al., 2013), ce qui signifie que FGF4 est requis pour le motif sel et poivre, mais que d'autres facteurs régulent l'expression initiale de *Gata6*. De plus, l'ajout de FGF4 peut réprimer l'expression de *Nanog* en l'absence de GATA6 (Bessonard et al., 2014), tandis que le blocage de la signalisation FGF peut réprimer l'expression de *Gata6* en l'absence de NANOG (Frankenberg et al., 2011). Cependant, ces traitements doivent être initiés avant que *Nanog* et *Gata6* ne commencent à être exprimés, c'est-à-dire autour du stade de compaction, car leur expression devient insensible à ces traitements à des stades ultérieurs (Bessonard et al., 2014; Frankenberg et al., 2011; Schrode et al., 2014). Ensemble, ces résultats soulignent que la signalisation FGF, NANOG et GATA6 forment un réseau central qui pilote le processus de spécification. La proportion de cellules EPI / PrE dans l'ICM est donc régulée par les niveaux relatifs des composants de réseau dans les cellules individuelles (Bessonard et al., 2014; Kang et al., 2013;

Krawchuk et al., 2013; Schrode et al., 2014; Schröter et al., 2015; Yamanaka et al., 2010).

Le réseau de régulation des gènes EPI / PrE a récemment été transposé dans un modèle mathématique (Bessonnard et al., 2014) afin d'examiner de quelle manière le schéma sel-poivre mutuellement exclusif est établi dans l'ICM (**Figure 2f**). Cette modélisation prend en compte la répression mutuelle de NANOG et de GATA6 couplée à l'auto-activation, ainsi que la signalisation FGF régulant positivement *Gata6* tout en inhibant *Nanog*. Avec ces conditions de configuration initiales, le modèle est suffisant pour récapituler le processus de développement *in vivo*. Les simulations montrent qu'une ou plusieurs cellules de l'ICM promeuvent l'expression de NANOG, entraînant une augmentation de la sécrétion de FGF4. Cette concentration locale plus élevée de FGF4 extracellulaire induit le destin du PrE dans les cellules voisines, ce qui suggère que les cellules ICM individuelles peuvent adopter de manière asynchrone un destin de PrE ou d'EPI selon la concentration locale de FGF4. Les événements de spécification EPI/ PrE asynchrones observés chez les embryons peuvent donc être expliqués par la propagation hétérogène de l'activité paracrine de FGF4. De plus, les cellules ICM expriment *Fgfr2* de manière homogène à E3.25 (Ohnishi et al., 2014), ce qui suggère que les différences d'activité de ERK sont initiées par des différences de concentration locale de FGF4 et sont ensuite amplifiées par la diminution de l'expression de *Fgfr2* induite par NANOG dans les cellules de l'EPI.

Le modèle mathématique prédit qu'une fois leur choix effectué, les cellules de l'ICM ne changent pas leur destin dans les embryons intacts (Bessonnard et al., 2014; Schröter et al., 2015). Cela avait déjà été suggéré par des expériences de traçage de lignées cellulaires (Chazaud et al., 2006; Meilhac et al., 2009) et confirmé au moyen d'un reporter *Nanog-GFP* (Xenopoulos et al., 2015).

Cependant, bien que la plupart des cellules ICM soient spécifiées au jour E3.75, elles conservent une certaine plasticité et peuvent changer de destin cellulaire, comme cela a été montré par des expériences de modulation de la signalisation FGF (Nichols et al., 2009; Yamanaka et al., 2010) ainsi que par la formation de chimères (Grabarek et al., 2012). Cette plasticité disparaît de manière asynchrone à E4.0 (Grabarek et al.,

2012; Yamanaka et al., 2010). Elle est d'abord perdue dans les cellules de l'EPI (Grabarek et al., 2012), reflétant probablement leur spécification antérieure (Bessonnard et al., 2014). Malgré ces informations sur le réseau FGF-NANOG-GATA6, les événements qui déclenchent les différences précoces dans les cellules ICM ne sont pas connus. L'hétérogénéité la plus précoce observée dans les cellules ICM est l'expression bimodale de *Fgf4* à E3.25 (Guo et al., 2010; Kurimoto et al., 2006; Ohnishi et al., 2014). Plusieurs hypothèses ont été proposées pour expliquer cette induction du schéma sel-poivre (Hermitte and Chazaud, 2014): 1/ l'activation stochastique de l'expression génique renforcée par des différences entre les cellules dans la sécrétion et la signalisation de FGF (Dietrich and Hiragi, 2007; Ohnishi et al., 2014) ; 2/ une ERK phosphorylée réduite, la présence/absence d'un autre facteur, dans quelques cellules, favorisant l'expression de *Nanog* et de *Fgf4* (Bessonnard et al., 2014) ; 3/ la répartition asymétrique du FGFR2 au cours des divisions cellulaires successives (Mihajlović et al., 2015; Morris et al., 2010, 2013) ou l'accumulation de cellules internes exprimant le FGF4 (Krupa et al., 2014). Aucune de ces hypothèses n'a encore été rejetée ou privilégiée. Malgré tout, des simulations mathématiques imitant le bruit de transcription élevé de tous les facteurs clés indiquent qu'il est peu probable que ce bruit interne soit le mécanisme initiateur (Bessonnard et al., 2014; De Mot et al., 2016). Ainsi, bien que l'étape initiale soit connue (l'engagement d'une ou de plusieurs cellules ICM vers un destin EPI autour de E3.0 précédant toujours la spécification des cellules PrE), nous ne savons toujours pas pourquoi l'expression de *fgf4* et *Nanog* augmente dans quelques cellules ICM (Bessonnard et al., 2014; Xenopoulos et al., 2015).

Chez l'homme, quelques données documentent la spécification EPI/PrE. En étudiant SOX17 (PrE) et OCT4 (EPI), Kathy Niakan a montré que *OCT4* était exprimé dans toutes les cellules au stade morula, et n'était restreint à l'EPI qu'une fois le blastocyste expansé. La même étude a montré l'apparition de *SOX17* dans des cellules localisées dans l'ICM, à proximité du blastocœle, une fois l'expansion initiée (Niakan and Eggan, 2013). Pour GATA6, nous ne disposons pas de données d'immunofluorescence avant le stade blastocyste expansé. En revanche, GATA6 est limité aux cellules PrE à J8, après deux jours de culture "post-implantation" (Deglincerti et al., 2016).

1.1.3 De la difficulté et de la nécessité et d'étudier le développement préimplantatoire humain

1.1.3.1 Ethique et recherche sur l'embryon humain

Les embryons préimplantatoires contiennent peu de cellules ce qui rend leur étude difficile puisque beaucoup de technique d'expérimentation nécessite un nombre de cellule de l'ordre du millier, voir du million. Ils peuvent contenir plusieurs types cellulaires différents et sont donc hétérogènes. Ainsi, même avec un nombre infini d'embryon à disposition, nous ne pourrions pas en comprendre tous les mécanismes sans méthode qui opère à l'échelle de la cellule unique.

Chez l'Homme, les embryons sont difficiles d'accès : il est possible d'en obtenir mais en très faible quantité. Leur manipulation à des fins de recherche est aussi limitée. La recherche sur le développement préimplantatoire humain est donc complexe. Ceci est dû à l'aspect sensible de cette recherche, puisque le matériel étudié est un potentiel nouvel individu. Cet aspect sensible impose de travailler dans un cadre éthique clair et bien défini. Ce cadre diffère selon le pays, cependant la société internationale de recherche sur les cellules souches (ISSCR) émet des lignes de conduite pour tous les chercheurs travaillant sur des cellules souches ou embryons humains. Un article qui contiendrait des expérimentations contraires à ces lignes de conduite serait impubliable dans une revue prestigieuse. Les expérimentations suivantes sont prohibées :

- Culture de l'embryon au-delà de 14 jours ou formation de la ligne primitive, selon ce qui se produit en premier.
- Gestation d'un embryon humain ex utero ou dans l'utérus d'un animal non humain.
- Clonage à des visée de reproduction.
- Modification génétique d'embryon mis en gestations. Ce point est particulièrement sensible depuis l'affaire He Jiankui, dite des « CRISPR babies ».
- Chimères animales-humaines ayant le potentiel de former des gamètes.

En France, lors de mon travail de thèse, la loi cadrerait lors de la recherche sur l'embryon humain selon les directives suivantes

- Toute recherche sur l'embryon humain doit être approuvée par un comité de l'agence de la biomédecine.
- Les embryons ne peuvent pas être créés à des buts de recherche : ils sont issus de procédure de FIV et ne font plus l'objet d'un projet parental. Un accord explicite doit être donné par les deux parents, et ce don ne peut être fait contre rémunération.
- Les embryons peuvent être cultivés jusqu'à 7 jours après la fécondation.
- Aucune modification génétique ne peut être effectuée sur un embryon humain, même à des fins de recherche.

La loi française a évolué dans le cadre de la révision de la loi bioéthique de 2019, et se rapproche maintenant des lignes de conduite de l'ISSCR, notamment sur la durée de mise en culture des embryons (14 jours) et sur la possibilité de modifier le génome d'un embryon humain à des fins de recherche.

Les cellules souches pluripotentes cultivées sous forme de lignées cellulaires en dehors d'embryons sont dites *in vitro*. Ces cellules peuvent être générées de différentes manières (cf. 2.1.1 en page 70), et sont potentiellement capables de former des structures mimant l'embryon humain précoce (Pera et al., 2015; Rivron et al., 2018a, 2018b). L'existence de ces cellules floute la limite de la définition de l'embryon, puisqu'il n'est potentiellement plus simplement le résultat de la fécondation (Sagan and Singer, 2007). L'avancée de la biologie fondamentale amène ici de toute nouvelles questions philosophiques, éthiques et légales qui sont de véritable enjeu pour la recherche dans le domaine de l'embryologie et des cellules souches.

1.1.3.2 Les modèles permettent d'étudier l'embryon humain en limitant leur utilisation

Les études du développement préimplantatoire humain se sont essentiellement concentrées sur la réalisation d'immunofluorescence. Ceci permet à l'aide

d'anticorps spécifiques de révéler, en général, la présence de trois marqueurs protéiques à la fois par embryon. Peu de connaissances sont donc disponibles chez l'homme, et ces connaissances montrent de nettes différences avec le développement préimplantatoire murin (Blakeley et al., 2015; Boroviak et al., 2018; Niakan and Eggan, 2013; Rossant and Tam, 2017). Extrapoler les connaissances sur l'Homme à partir de modèles animaux, notamment la souris, ne semble donc pas être une alternative qui pourrait se substituer à elle seule à l'étude du développement préimplantatoire humain.

La première façon de contourner le problème est d'obtenir des modèles cellulaires *in vitro*. Dans ce cas le but est d'obtenir des cellules capables d'auto-renouvellement et capturant l'état cellulaire de la lignée voulue. L'épiblaste est pluripotent, c'est-à-dire qu'il possède un potentiel de différenciation en tout type cellulaire de l'organisme adulte. Les cellules souches pluripotentes *in vitro* sont donc un modèle de l'épiblaste qui permet d'étudier des concepts tel que les barrières épigénétiques qui limitent la plasticité des cellules. L'étude de ces barrières est d'autant plus importante que les blastocytes humains semblent avoir des caractères de plasticités particuliers par rapport à la souris. En effet le trophoctoderme semble être capable de se convertir en épiblaste (De Paepe et al., 2013). Dans ce cas, on parle de plénipotence, c'est-à-dire qu'une cellule de TE n'est pas capable de redonner un embryon entier comme le ferait un blastomère totipotent. Cependant dans un embryon, une cellule de TE humaine est suffisamment plastique pour se convertir en EPI, et peut donc théoriquement contribuer à tous les tissus intra et extra-embryonnaires.

Une autre façon d'étudier le développement préimplantatoire humain est d'établir des modèles à partir de données provenant de méthodes de séquençage en haut débit, telles que le RNA-Seq. Seulement, les méthodes classiques ne permettent pas de travailler sur l'embryon humain pour les raisons évoquées au début de cette section. Le développement des méthodes en cellule unique, notamment du single-cell RNA-Seq et de ses outils d'analyse a donc eu un impact majeur dans ce domaine de recherche (cf. 1.2.1.2 en page 25). Le va-et-vient entre les modèles cellulaires et les

embryons humains est une clef pour élucider les événements moléculaires du développement préimplantatoire humain.

1.2 Etudier l'expression des gènes en masse : du transcriptome à la cellule unique

Pour découvrir comment la spécification des différents destins cellulaires est régulée, il est primordial de développer des modèles *in vitro* et *in silico* pour les lignées cellulaires impliquées dans le développement préimplantatoire humain. Ceci est au cœur des thématiques de recherche de notre équipe. L'analyse de données transcriptomiques est centrale lorsqu'il s'agit de créer ou valider ces modèles. De façon plus générale, les analyses transcriptomiques ont pris une place importante dans l'étude du vivant. Ainsi, l'utilisation et la création de méthodes d'analyse de ce type de données sont devenues un enjeu majeur de la bioinformatique. Nous allons dans cette section passer en revue certaines de ces méthodes, liées aux différentes étapes de l'analyse de transcriptome.

1.2.1 Estimer le transcriptome à partir de RNA-Seq

Le RNA-Seq (Mortazavi et al., 2008) ou encore séquençage de l'ARN, est une technologie de séquençage haut débit permettant d'identifier et de quantifier les transcrits des gènes sur le génome entier ou tout du moins sur un ensemble conséquent de gènes, souvent plusieurs dizaines de milliers. Cet ensemble conséquent de transcrits est désigné par le terme de transcriptome. Nous pouvons donc aussi qualifier le RNA-Seq de séquençage de transcriptome. Ceci est rendu possible grâce à un ensemble d'avancées techniques et méthodologiques dans des domaines variés comme la biologie moléculaire, l'informatique et les statistiques. L'objectif de ces méthodes est de passer d'un échantillon biologique à une table de comptage de l'expression des gènes. Diverses stratégies ont été développées ces dernières années pour arriver à ce comptage.

Librairie



Séquençage ▼

FASTQ

@read1 *ligne 1 : Identifiant unique du read*
ATGCC... *ligne 2 : Séquence du read*
+ *ligne 3 : Séparateur +*
%%++)... *ligne 4 : Score de qualité de chaque nucléotide*
@read2
 ...

Index du génome de référence

Format spécifique à l'aligneur, fichier créé à partir du génome de référence en **FASTA**

Alignement ▼

SAM/BAM

SAM Flag : code décrivant l'alignement
 ID du read **read1** **256** *Feature sur laquelle le read a été aligné* **NM_002051.3** *Emplacement du nucléotide du read 1 sur le génome de référence* **613960** *Score d'alignement, signification propre à l'aligneur* **1** *Description du read «mate», utile seulement dans le cas d'un séquençage en «paired end»* **65M** *** 0 0**
CCAGCGCGAAC... *Séquence du read*
EDDEEDEE=EE... *Score de qualité de chaque nucléotide*
AS:i:0 XS:i:0 ... *Données spécifiques à l'aligneur (explication dans l'entête du fichier) au format TAG:TYPE:VALUE*
 read2 ...
 ...

Annotation du génome

Base de donnée des «features» (exon, gène...) le long du génome de référence, généralement au format **GTF/GFF**

Comptage ▼

Table de comptage de l'expression

Au format texte (CSV/TSV)

	Ech. 1	Ech. 2	Ech. 3	Autres échantillons
CDX2	0	46	0	
FHL2	0	0	0	
GATA2	0	140	26	→ 26 reads de GATA2 dans l'échantillon 3
GATA3	1	72	165	
Autres gènes				

Figure 3 : les acteurs du RNA-Seq. Les pointillés représentent des séquences pouvant ne pas être présentes selon le type de librairie. Par exemple, l'index permet d'identifier le batch d'expérience. FASTQ, SAM sont des formats de fichiers courant en bioinformatique. Leur structure est présentée en détail. Les BAM sont des fichiers SAM compressés par binarisation.

1.2.1.1 Le RNA-Seq est un ensemble hétérogène de méthodes de transcriptomiques

Les différences au sein des méthodes de RNA-Seq résident dans la construction de la collection d'ADN prête à passer dans le séquenceur, la librairie. La librairie est constituée de fragments courts d'ADN provenant de la rétrotranscription de l'ARN (ADN complémentaire), ainsi que de séquences artificielles. Chaque fragment correspondra à un « read » dans le fichier produit par le séquenceur. Ces fichiers sont habituellement de type FASTQ (**Figure 3**). Les fichiers FASTQ sont des fichiers texte renfermant la séquence des reads et des scores de qualité de séquençages associées à chaque nucléotide.

La nature des séquences artificielles ajoutées au fragment d'ADN complémentaire est propre à la stratégie de construction de la librairie (**Figure 3, Figure 4**). La plupart du temps sont au moins présentes des séquences qui permettant l'accroche et la réplification des fragments dans le séquenceur, les *adaptateurs*. Le *barcode* est un autre type de séquence ajoutée au fragment. Il permet d'identifier l'échantillon de provenance de l'ARN. Grâce aux barcodes, les échantillons peuvent être réunis pour être séquencés simultanément afin de réduire les coûts de séquençage. Les reads obtenus sont alors réattribués de façon adéquate selon la séquence du barcode durant l'analyse des résultats. Cette étape est appelée le démultiplexage (Wong et al., 2013).

Le séquençage de seconde génération, utilisé en majorité pour le RNA-Seq (Mortazavi et al., 2008), limite la longueur des fragments séquençables à quelques dizaines de paires de bases, or un transcrit est généralement long de quelques milliers de paires de bases (**Figure 5**). Nous pouvons alors classer les protocoles de RNA-Seq selon la façon dont cette limitation est gérée. Une façon de faire est de morceler les transcrits : un transcrit va donner plusieurs reads. Cette méthode fut la première stratégie employée dans le RNA-Seq. Le principal avantage du morcelage est de couvrir toute la longueur des transcrits, ce qui permet de modéliser les transcrits de départ et d'étudier par exemple la diversité des isoformes. L'inconvénient est que le nombre de reads par gène n'est pas seulement dépendant de l'abondance de ces transcrits, mais est aussi influencé par la longueur du gène.

Plus un gène est long plus il donne de reads. De nombreuses étapes d'amplification sont nécessaires pour séquencer une librairie de RNA-Seq. La nature exponentielle de l'amplification par PCR entraîne donc des risques de biais d'amplification, où des séquences vont être sur-représentées par rapport à leur quantité de départ. Pour résoudre ces problèmes de biais d'amplification, de nouvelles séquences appelées identifiant moléculaire unique (*Unique Molecular Identifier*, ou UMI), ont été introduites (Islam et al., 2014) dans les fragments de librairie de RNA-Seq. L'idée est de fixer une séquence générée aléatoirement de taille fixe avant amplification à chaque transcrit. Généralement, seule une partie du transcrit de départ est séquencée. Le nombre de reads d'un gène ne sera donc plus influencée par sa longueur. Néanmoins, cette étape mène à une perte d'information, rendant l'étude des isoformes limitée à la zone du transcrit séquencé. Après amplification, plusieurs fragments correspondent à un gène exprimé, ceux-ci provenant de plusieurs transcrits. En utilisant les UMI amplifiées avec le transcrit nous pouvons retrouver le nombre de transcrits capturées. Pour cela, il suffit de compter le nombre d'UMI différent par gène et par échantillon. La méthode de construction de librairie nommé *3' Digital Gene Expression Sequencing* ou DGE-Seq (Soumillon et al., 2014) fonctionne selon ce principe et est utilisée à Nantes, et dans le cadre de mes travaux de thèse. Le DGE-Seq est utilisé pour du RNA-Seq « bulk » (cf. 1.2.1.2, en page 25). Cette technologie utilise des barcodes et des UMI. Seul les 50 derniers nucléotides des transcrits sont séquencés, de façon qu'un transcrit puisse être associé à un UMI en un read (**Figure 4**). La longueur du gène n'est alors plus un facteur de confusion de l'estimation de l'expression.

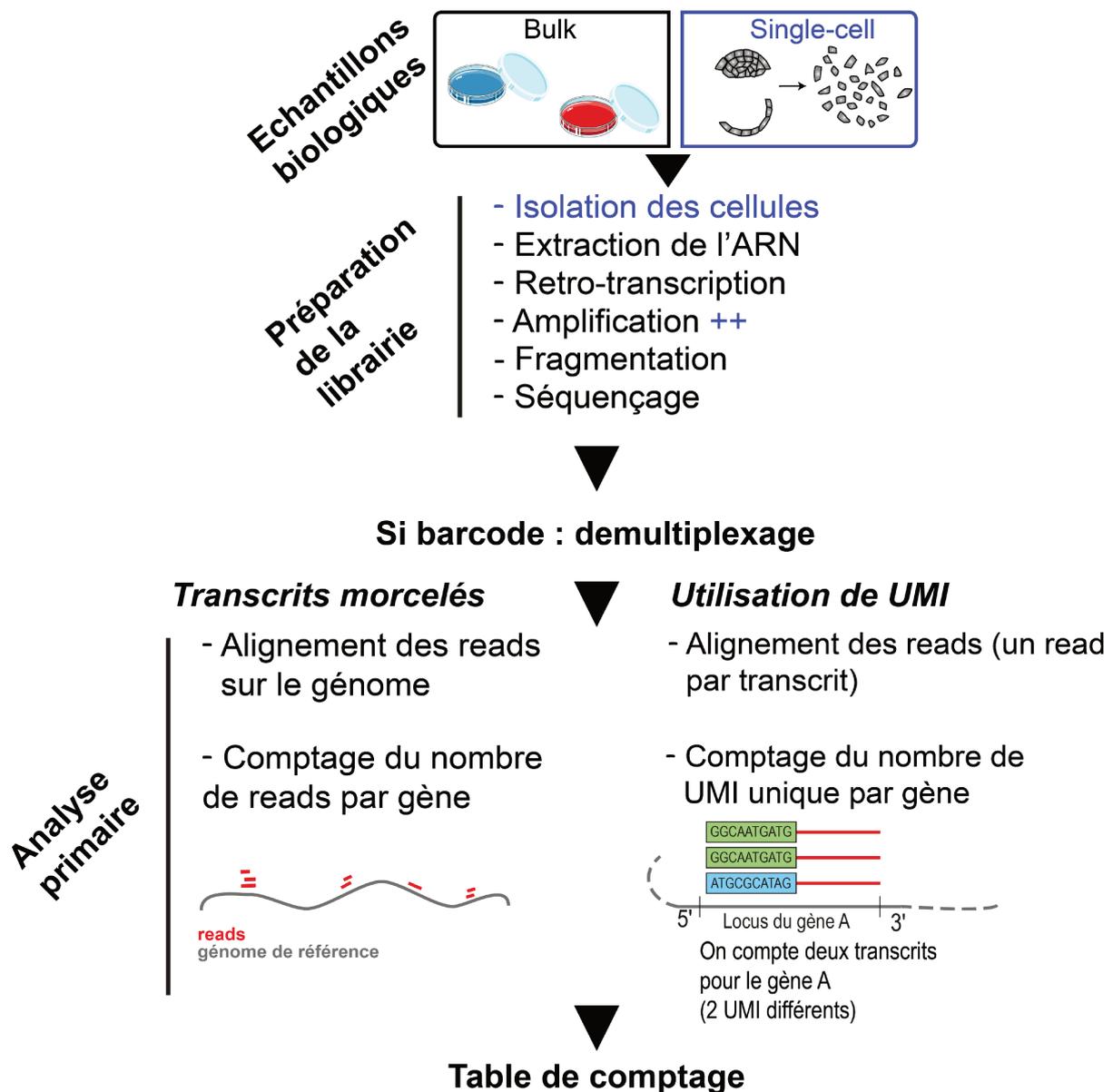


Figure 4: Les étapes du RNA-Seq amenant à la création de la table de comptage. Les étapes spécifiques au singe-cell sont notées en bleu. Le DGE-Seq sert d'exemple d'analyse primaire avec utilisation d'UMI.

Il est à noter que des technologies de RNA-Seq avec séquençages de troisième génération dits à « long reads » commencent à être utilisées. Le principal obstacle de ce type de méthode est le taux d'erreur du séquençage, cependant ce taux d'erreur s'améliore constamment. Pour l'instant l'un des principaux intérêts de cette technologie est l'étude de transcriptome dont le génome de référence de l'espèce n'est pas ou peu connu. En effet des reads plus longs permettent de reconstruire plus facilement les transcrits de départ sans génome de référence (Pollard et al., 2018) Des protocoles mixtes mélangeant seconde et troisième génération de séquençage sont utilisés dans le même cas de figure. Dans ce cas un transcriptome

de référence est créé avec les reads longs, puis les reads court provenant du séquençage deuxième génération sont alignés sur cette référence.

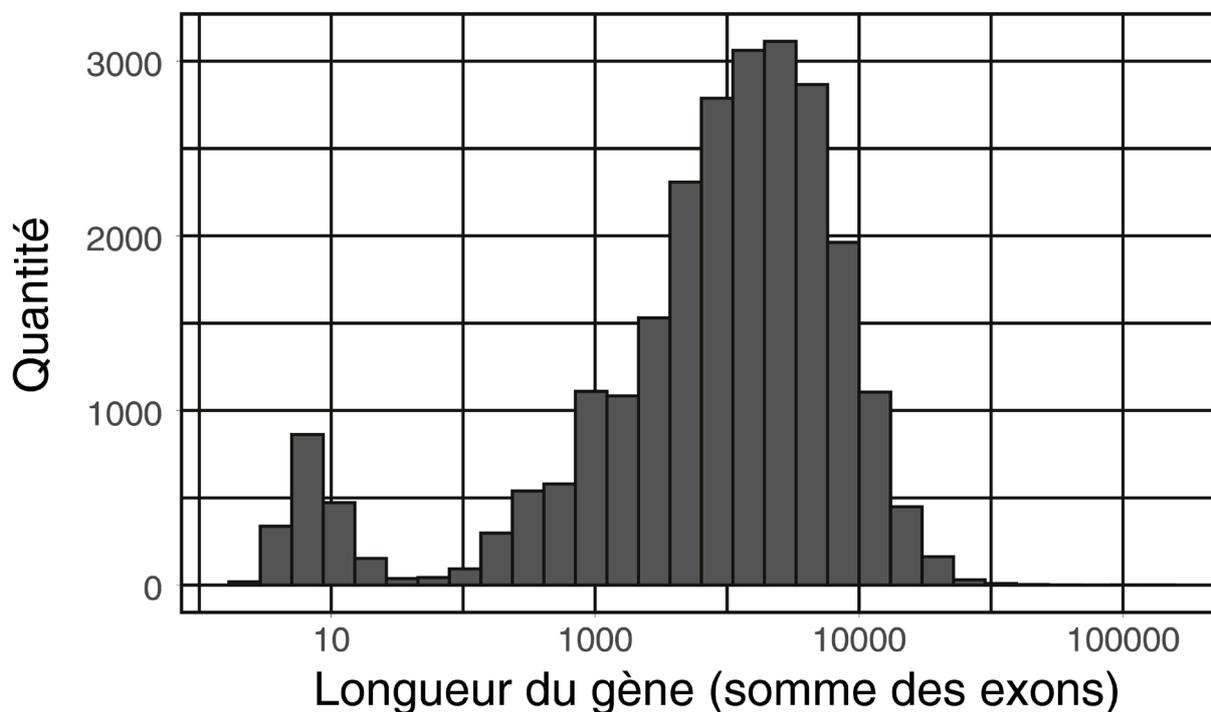


Figure 5 : Distribution de la longueur des gènes chez l'Homme. Ces données ont été calculées à partir du fichier d'annotation du génome humain (version GRCh38).

1.2.1.2 L'avènement du séquençage de transcriptome en cellule unique

Une autre façon de classer les méthodes de RNA-Seq est de s'intéresser à la nature des observations/échantillons au sens statistique du terme. Habituellement chaque échantillon de RNA-Seq est en réalité composé au minimum de plusieurs centaines de milliers de cellules, en effet il est nécessaire d'avoir un minimum de concentration en ARN pour un RNA-Seq efficace. Pour le DGE-Seq le minimum se trouve à 10 ng ce qui représente environ 1000 cellules dans le cas de cellules souches pluripotentes. Le fait de devoir moyenner l'expression d'une telle quantité de cellules à plusieurs conséquences. La première est d'obtenir un transcriptome ne correspondant à aucun type cellulaire, mais plutôt à un mélange des types cellulaires existant dans l'échantillon (Trapnell, 2015). On peut voir ce mélange comme un « smoothie » des transcriptomes de cellules uniques. Classiquement le RNA-Seq a été utilisé pour trouver des gènes ayant une réponse différente suivant une condition, souvent un traitement. Dans ce cas nous pourrions retrouver la liste des gènes différentiellement exprimés entre les conditions, mais pas les types

cellulaires responsables de cette expression différentielle. Pire ce problème peut résulter en l'implémentation d'un paradoxe de Simpson (**Figure 6**). Il est donc particulièrement difficile d'interpréter des résultats de RNA-Seq classique dans un contexte de type cellulaires hétérogènes, et il est impossible d'estimer cette hétérogénéité avec ces mêmes données.

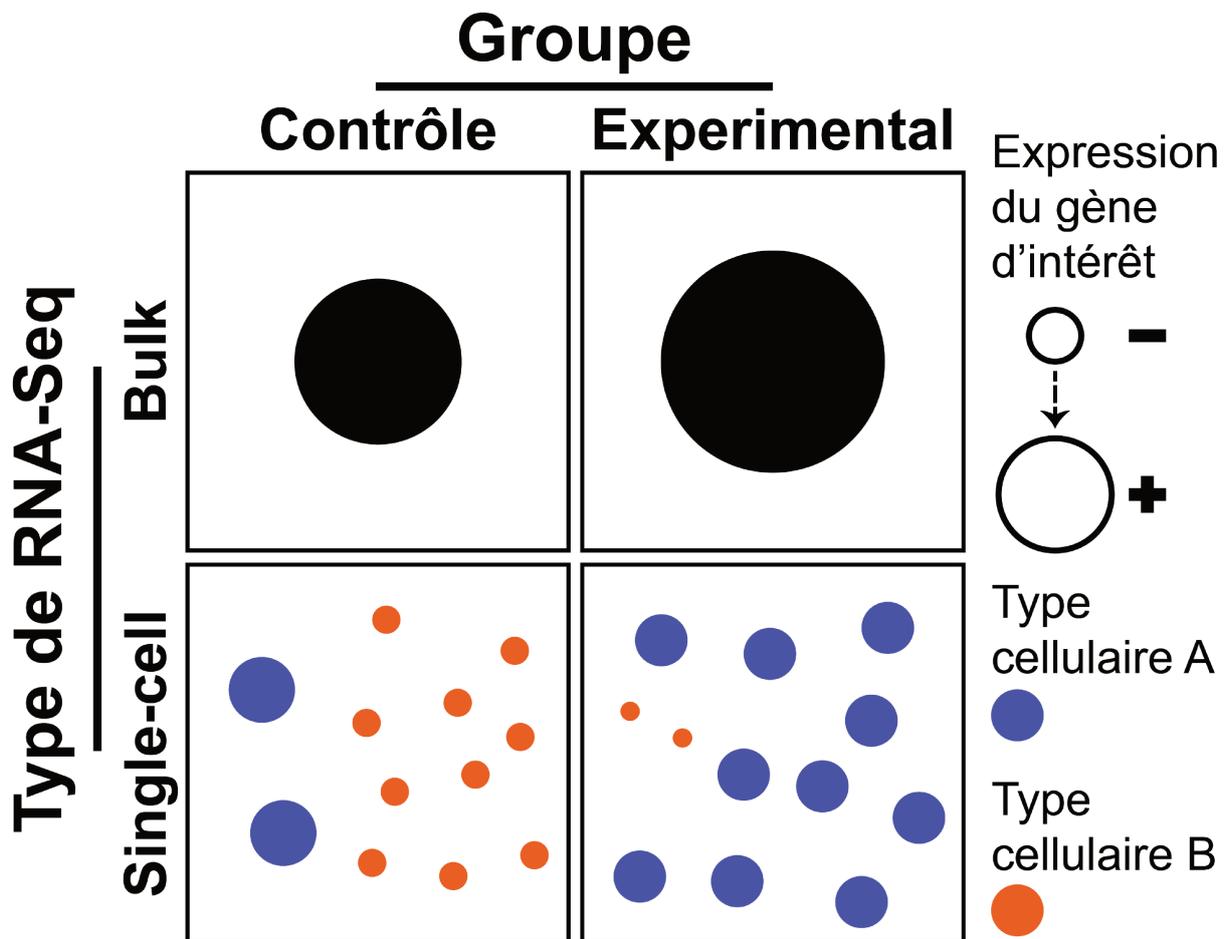


Figure 6 : Schéma de l'implémentation d'un paradoxe de Simpson. Prenons le cas de l'étude d'un unique gène, en RNA-Seq « classique » (bulk), le gène est plus exprimé dans les échantillons du groupe expérimentale. Lorsque que l'on séquence les mêmes échantillons en single-cell l'expression du gène a en revanche diminué dans le groupe expérimental pour tous les types cellulaires. La clef de l'explication se trouve dans le changement de fréquence des deux types cellulaires entre les deux groupes.

La deuxième conséquence est la disparition des transcriptomes de types cellulaires rares noyés dans le transcriptome de types cellulaires abondants, or les types cellulaires rares peuvent être la clef de la compréhension de processus biologiques. Enfin la quantité de cellules nécessaire à la réalisation du RNA-Seq classique le rends inaccessible à certains domaines de la biologie, tel que le développement

embryonnaire précoce, où un blastocyste contient quelques centaines de cellules de types différents.

Ces problèmes ont pu être résolus grâce à l'évolution du séquençage, des protocoles de construction de librairie et d'isolation des cellules qui permet de profiler le transcriptome d'une cellule unique (Tang et al., 2009). On parle de *single-cell RNA-Seq* (ou scRNA-Seq). Le terme *bulk RNA-Seq* est lui apparu en opposition au single-cell et désigne le RNA-Seq « classique ». En scRNA-Seq, l'unité de l'échantillon correspond à l'unité du vivant. Ainsi, l'apparition du scRNA-Seq a suscité un véritable engouement, et, est rapidement devenu un outil commun parmi l'arsenal des méthodes de séquençage en haut débit (dites *OMICS*). Cette effervescence a été le point de départ d'une évolution rapide des protocoles de construction de librairie, notamment grâce à l'apparition de plusieurs technologies clef (**Figure 7**).

D'autres méthodes de biologie moléculaire sont aussi adaptables en single-cell, tel que la PCR quantitative en cellule unique, plus vieille que le scRNA-Seq et qui reste une alternative viable au profilage d'expression, lorsque l'on s'intéresse à un panel de gène restreint (Bengtsson et al., 2008). D'autres méthodes *OMICS* sont aussi converties dans leur version « single-cell ». Voici quelques exemples :

- le séquençage de génome, ou DNA-Seq (Chen et al., 2016) ;
- du séquençage de marques épigénétique avec le séquençage au bisulfite (Hou et al., 2016) ;
- l'ATAC-Seq, qui représente des données d'ouverture et de fermeture de la chromatine (Buenrostro et al., 2015) ;
- de la mise en évidence de liaison chromatine-protéine avec le ChiP-Seq (Rotem et al., 2015) ;
- de la capture de conformation de l'ADN avec le single-cell Hi-C (Nagano et al., 2013).

L'ensemble de ces méthodes permet d'avoir une vision globale des différentes couches de régulations de la cellule, et nécessitent des analyses adaptées aux problématiques d'analyses propres aux données en cellule unique (cf. chapitres suivants).

1.2.1.3 L'analyse primaire permet de passer des reads à la table de séquençage

Le premier but du RNA-Seq, single-cell ou bulk, est la quantification de l'expression des gènes à l'échelle du transcriptome. Les FASTQ contiennent les reads, qui sont des données sous forme de chaîne de caractère. Il est donc nécessaire de passer des reads à des valeurs numériques qui seront agrégées dans une table de comptage. Ces étapes de transition des reads jusqu'à la table de comptage sont communément regroupées sous le terme d'analyse primaire (**Figure 4**). Les FASTQ sont tout d'abord inspectées grâce à des outils de contrôle qualité (QC) dont le plus connu est FASTQC (Andrews et al., 2012). Cette inspection peut amener à supprimer des reads ou des parties de reads problématiques, et améliorer les étapes ultérieures de l'analyse primaire.

La première étape de transformation des reads est l'alignement, généralement sur un génome de référence ou un transcriptome de référence. Le choix de l'aligneur, le logiciel qui procède à l'alignement, sera donc primordiale pour produire une table de comptage de bonne qualité (Grant et al., 2011). En effet, un aligneur doit-être capable d'aligner des millions de reads dans un temps raisonnable sur une référence comportant elle-même au minimum chez l'Homme environ trente millions de bases (cas de l'exome). De plus un aligneur de RNA-Seq doit-être capable de gérer les évènements d'épissage alternatifs et doit pouvoir aligner des reads comportant des erreurs de séquençage afin d'obtenir un taux d'alignement correct. Ces deux points augmentent de façon considérable les possibilités d'alignement. Le RNA-Seq n'a donc pas été seulement possible grâce aux avancées de la biologie moléculaire, mais aussi grâce aux travaux effectués en bioinformatique fondamentale. Parmi ces avancées, nous pouvons citer l'indexation du génome qui permet de stocker les positions de chaque motif existant d'une certaine longueur (ou *k-mer*) dans la référence (Lam et al., 2008), et la transformée de Burrows-Wheeler permettant de réorganiser la référence de façon à accélérer l'alignement (Li and Durbin, 2009). Ces procédés sont au cœur de l'algorithme d'alignement bowtie2 (Langmead and Salzberg, 2012) utilisé par l'aligneur HISAT2 (Kim et al., 2015). Les meilleurs aligneurs (Baruzzo et al., 2017), utilisées aujourd'hui en scRNA-Seq tel que HISAT2 et

STAR (Dobin et al., 2013) ont tout d'abord été développés pour du bulk RNA-Seq. Cependant la nature des reads n'est pas différente entre le bulk et le scRNA-Seq et ces aligneurs sont également adaptés pour des jeux de données de plusieurs milliers de cellules uniques ou moins.

Les fichiers produits par l'alignement sont habituellement des fichiers SAM ou leur version binaire moins volumineuse, les fichiers BAM. Ces fichiers contiennent entre autres la séquence de chaque read, et des informations sur l'alignement, notamment si l'alignement a bien été effectué et, le cas échéant, la position de l'alignement (**Figure 3**). Une fois les fichiers SAM/BAM obtenus il suffit de compter le nombre de reads alignés aux positions dans le génome correspondant à des gènes. Cette étape de comptage se fait généralement avec un fichier GTF/GFF contenant les positions des annotations dans le génome. Grâce à ce fichier il est possible par exemple de différencier les isoformes dans la table de comptage. Plusieurs outils sont utilisés pour le comptage, tel que le module count de HTSeq (Anders et al., 2015) ou Cufflinks (Trapnell et al., 2012). Ces dernières années des jeux de données contenant plusieurs centaines de milliers voire plusieurs millions de cellules sont apparus, notamment grâce à l'amélioration constante des technologies de capture et de barcoding des cellules (**Figure 7**). Des pseudo-aligneurs tel que Kallisto (Bray et al., 2016) ont été développés pour travailler à cette échelle, ceux-ci estiment la table de comptage plus rapidement que les aligneurs, sans alignement exact de reads mais plutôt en proposant une liste de transcrits d'appartenance possible.

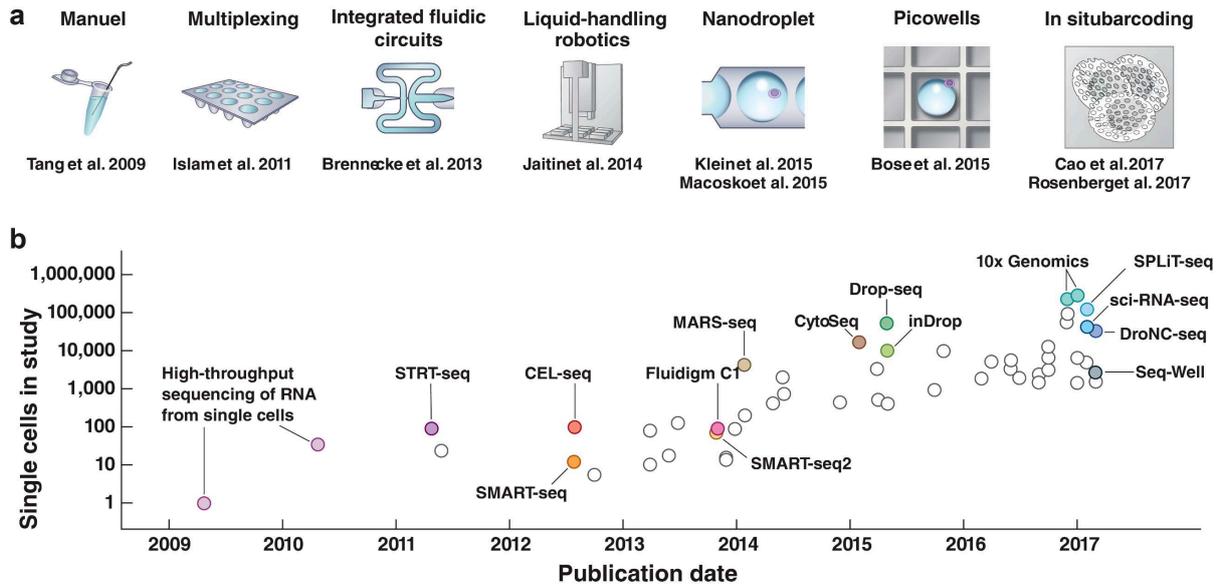


Figure 7 : Evolution du scRNA-Seq. Figure issue de Svensson et al., 2018. **a.** Principales technologies qui ont permis des sauts d'échantillonnage. Un saut à ~100 cellules a été permis par le multiplexage d'échantillons, puis un saut à ~1 000 cellules a été réalisé par des études à grande échelle utilisant des circuits fluidiques intégrés, suivi d'un saut à plusieurs milliers de cellules avec la robotique de manipulation des liquides. D'autres augmentations d'ordre de grandeur, à des dizaines de milliers de cellules, ont été rendues possibles par des technologies de capture aléatoire utilisant des nanogouttelettes ou nano-puits. Des études récentes ont utilisé le barcoding *in situ* pour atteindre à moindre coût prochain ordre de grandeur de centaines de milliers de cellules. **b.** Nombre de cellule utilisé dans des publication représentative par date de parution. Les principales technologies sont indiquées.

1.2.2 L'analyse de la table de comptage de l'expression est un défi mathématique

Dans cette partie, l'analyse de la table de comptage sera détaillée par le prisme des outils et méthodes utilisés ou envisagés lors de mon travail de thèse.

1.2.2.1 Aux origines de la complexité du RNA-Seq

Les techniques de quantification se basant sur du séquençage en haut débit ont pour point commun de générer une valeur par échantillon et par gène ou région chromosomique. Dans le cas de la transcriptomique, chaque gène est une variable. Pour l'Homme qui possède environ 50000 gènes et pseudogènes, cela représente donc 50000 variables. Si l'on veut placer un échantillon au sein d'un jeu de données de transcriptomique sans filtration de variables, il faut donc environ 50000

axes/dimensions. La taille de la table de comptage présente ainsi un véritable défi en termes d'analyse statistique.

Les statistiques classiques et notamment la P-valeur (Fisher, 1925) ont été pensées pour différencier deux type de variations :

1. Les variations dans les données dues aux fluctuations d'échantillonnage.
2. Les variations dues aux relations avec les variables étudiées (Poids, âge, groupe d'étude...).

Ainsi une p-valeur faible indique qu'il est peu probable que les variations observées dans les données soient seulement dues aux fluctuations d'échantillonnage.

Or avec un très grand nombre de variables, des phénomènes peu probables à l'échelle des statistiques classiques ont une forte probabilité d'apparition. L'ensemble des phénomènes délétères dus à un nombre élevé de variable constitue la « curse of dimensionality » (**Figure 8**) pouvant être traduit par « fléau de la dimension » (Bellman, 1961). L'analyse de données transcriptomiques requiert donc des statistiques adaptées à l'analyse en très haute dimension.

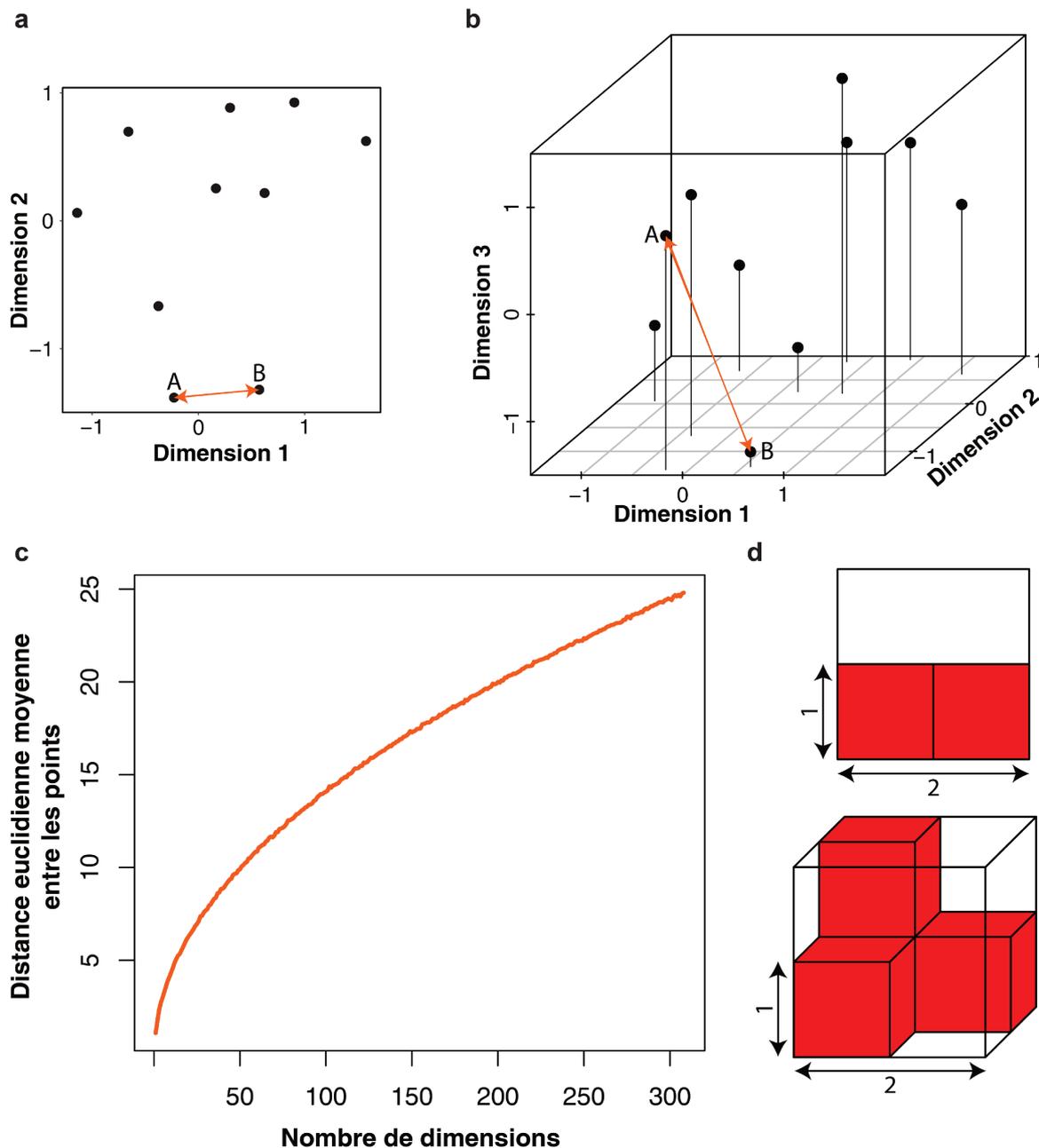


Figure 8 : Illustrations du fléau de la dimension. a-b. Dix points répartis au hasard selon une loi multinormale centrée réduite à deux dimensions (a) ou trois (b) dimensions. Les dimensions 1 et 2 sont conservés entre a et b. Le segment orange représente la distance euclidienne entre les points A et B. Il existe plus de « vide » dans la figure en trois dimensions. Ainsi la distance euclidienne a augmenté entre tous les points, y compris pour A et B. c. Distance moyenne entre dix points tirés selon une loi multinormale centrée réduite en fonction du nombre de dimensions de cette loi multinormale. Les points étant de plus en plus isolé, l'utilisation de certains algorithmes comme le clustering s'avèrent plus difficile. Cette courbe a été calculé à partir d'une simulation à 100 tirages pour chaque nombre de dimensions. d. Le volume d'un hypercube se calcule par l^d avec l le côté et d le nombre de dimensions. Une faible différence de côté entre deux hypercubes se traduit donc par une immense différence de volume à haut nombre de dimension. Ici en illustration, il faut quatre carrés de côté un pour remplir un carré de côté deux, alors qu'il faut huit cubes de côté un pour remplir un cube de côté deux.

Un défi supplémentaire du scRNA-Seq est lié à l'aspect « cellule unique » qui change la nature des données et les objectifs de leur analyse. La première difficulté provient du bruit des données, en effet la quantité de matériel de départ pour chaque échantillon est inférieure de plusieurs ordres de grandeur par rapport au bulk RNA-Seq, ce qui augmente les erreurs d'estimation de l'expression. D'autres phénomènes invisibles en bulk RNA-Seq viennent expliquer ce bruit, tel que le RNA bursting, le fait qu'à l'échelle de la cellule unique un gène exprimé n'est pas continuellement transcrit (Suter et al., 2011), et les phénomènes de transcriptions stochastiques (Raj and van Oudenaarden, 2008). Dans ce dernier cas des gènes non exprimés dans un type cellulaire peuvent tout de même être transcrit à l'échelle de la cellule. D'autres phénomènes apportent de l'hétérogénéité à un type cellulaire, comme le cycle cellulaire ou encore l'apoptose (Buettner et al., 2015).

Un autre défi est dû à la nature non supervisée des échantillons. En effet, le single-cell ne permet pas avoir d'a priori sur la nature des cellules séquencées. Toutes les analyses de single-cell commencent donc par une analyse non supervisée. Le choix de cette analyse est primordial puisque c'est elle qui va fixer les populations d'appartenance des cellules pour ensuite éventuellement retomber sur des analyses supervisées entre les types/groupes cellulaires déterminées. La notion d'échantillon et d'inférence pose aussi un problème, en effet dans les analyses en single cell la cellule est l'unité statistique mais n'est pas l'unité d'inférence. En effet si l'on veut inférer une conclusion sur des millions de cellules qui proviennent toutes d'un individu, la puissance statistique sera considérable. Néanmoins, il ne sera pas possible de généraliser l'inférence sur la population d'où provient cet individu. Le traitement de cette confusion entre l'échantillon biologique et l'échantillon statistique semble être pour l'instant peu implémenté dans les outils d'analyse en single-cell.

1.2.2.2 La distribution de la table de comptage : de la loi de Poisson à ZINB

La table de comptage est une matrice où les lignes correspondent aux gènes et les colonnes aux échantillons. Dans le cas particulier du single-cell, chaque colonne est

une cellule. Afin d'analyser cette matrice il faut approximer le processus derrière la répartition des données, c'est-à-dire trouver des lois de distribution qui permettent d'inférer des conclusions en minimisant les erreurs. Le problème du RNA-Seq peut être illustré avec des tirages aléatoires dans une urne. Prenons une urne contenant des billes de divers couleurs. Chaque couleur peut être représentée plusieurs fois : pour une urne nous pouvons établir l'abondance de billes par couleur. Dans ce modèle, une bille est un transcrit, une couleur un gène et une urne un échantillon.

Tout l'enjeu du RNA-Seq est de tirer suffisamment de billes pour estimer l'abondance des couleurs de départ des urnes. La quantité de billes d'une couleur provenant d'une urne lors d'un tirage est alors une variable aléatoire quantitative discrète dépendante du nombre de billes tirées dans l'urne et de l'abondance de la couleur. Cette variable répond à une loi binomiale de paramètre n, p avec n le nombre de billes tirées et p la probabilité d'attribution à la couleur. En RNA-Seq, nous avons un large éventail de gènes, dans ce cas l'urne contient des milliers de couleurs différentes. On peut donc considérer que la probabilité de tirage d'un couleur donnée est faible ; de plus nous tirons un grand nombre de billes. Ces propriétés nous permettent d'utiliser une loi de Poisson de paramètre λ correspondant aux nombres de billes tirées en moyenne dans un type d'urne avec un nombre total de billes tirées fixe. Autrement dit l'expression d'un gène dans un type cellulaire peut-être approximé pour une taille de librairie donnée par une loi de Poisson (Marioni et al., 2008; Wang et al., 2010). Les lois de Poisson sont fréquemment utilisées pour modéliser des processus de comptage. Dans cette distribution, la variance est égale à la moyenne. Cette approximation suffit pour décrire des données de RNA-Seq où chaque échantillon est issu d'une population de répliques techniques. Les répliques techniques correspondent à un même ARN mis en librairie et séquencées plusieurs fois, et sont utiles pour évaluer la fiabilité d'un protocole.

Cependant, dans la mesure où le RNA-Seq a pour but d'étudier des phénomènes biologiques, il faut utiliser des échantillons qui proviennent de populations de répliques biologiques. Dans le cas du bulk RNA-Seq, chaque échantillon doit donc provenir d'individus suffisamment différents pour pouvoir inférer les conclusions

de l'étude sur une population qui représente un intérêt. L'apport de la variabilité biologique change la distribution des données, la loi de Poisson n'est alors plus applicable (Marioni et al., 2008; Robinson and Smyth, 2007). Dans ce cas la dispersion (ou variance) de l'expression est plus grande qu'attendue. Cette surdispersion s'amplifie d'autant plus que la moyenne d'expression du gène est élevée (**Figure 9**). Dans l'exemple de notre urne, nous avons plus de mal que prévu à estimer l'abondance de départ d'une couleur, et plus l'abondance d'une couleur est élevée dans une urne, plus le nombre de billes obtenu va varier entre différent tirage. Nous pouvons combiner plusieurs lois de Poisson de différents paramètres λ pour tenter de modéliser la surdispersion. Nous pouvons généraliser cette approche en faisant de λ une variable aléatoire suivant une loi de distribution Gamma. Nous obtenons alors une distribution Poisson-Gamma. La loi de distribution Poisson-Gamma est strictement équivalente à une loi négative binomiale (Robinson and Smyth, 2007). Ainsi une des lois de distribution les plus utilisées en RNA-Seq est la négative binomiale. Les paramètres utilisés en RNA-Seq pour la négative binomiale se présentent sous la forme μ et ϕ , respectivement la valeur de centrage et de dispersion. Il est à noter qu'habituellement les paramètres de la loi négative binomiale se présentent sous la forme r, p . Néanmoins la forme μ, ϕ est plus intuitive pour décrire les données d'expression de gène. Par exemple, lorsque $\phi = 0$, la distribution est équivalente à une loi de Poisson (Robinson and Smyth, 2007).

D'autres pistes ont été étudiées pour modéliser la distribution des counts: la distribution log-normale, ou encore la transformation *voom*, qui fait adopter aux données une distribution normale. Cette dernière méthode a pour intérêt de rendre disponible au RNA-Seq tout l'arsenal statistique développé pour étudier les variables aléatoires suivant une distribution normale. Cependant les outils se basant sur ces deux approches se sont révélés moins efficaces que ceux qui se basent sur la loi négative binomiale (Costa-Silva et al., 2017; Froussios et al., 2019).

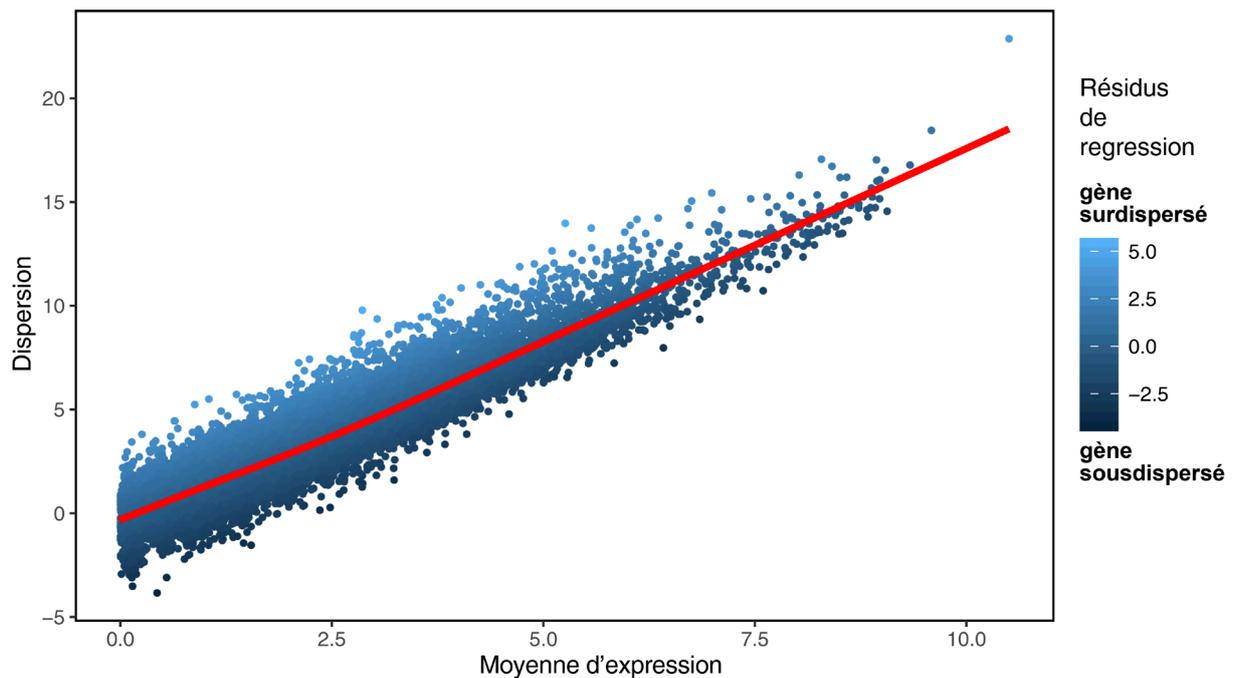


Figure 9 : Evolution de la dispersion en fonction de la moyenne d'expression pour chaque gène dans des données RNA-Seq. Les données de RNA-Seq suivent une loi négative binomiale, la dispersion augmente avec la moyenne d'expression. La moyenne d'expression et la dispersion sont à l'échelle logarithmique, et ont été calculées à partir de données de DGE-Seq normalisées, et utilisées dans Kilens et al., 2018. La dispersion est estimée par la variance. La courbe rouge représente une régression locale de la dispersion en fonction de la moyenne. La projection de chaque gène sur la courbe donne sa « dispersion théorique ». Si la dispersion empirique (mesurée) est supérieure à la dispersion théorique, le gène sera considéré comme surdispersé. Les valeurs de résidu à la régression donnent donc directement accès à cette notion de surdispersion. Généralement la dispersion est plutôt représentée par le coefficient de variation sur ce type de figure. La forme du nuage de points change, mais les résultats sont équivalents.

Les tables de comptages de scRNA-Seq contiennent plus de zéros que les tables de bulk RNA-Seq de tailles équivalentes. On parle de *dropout*. L'origine des dropout est multiple, et peut être aussi bien technique que biologique. Parmi les hypothèses les plus consensuelles nous pouvons citer les problèmes d'amplification dus au peu de matériel de départ, ainsi que le RNA bursting (Raj and van Oudenaarden, 2008). Les différentes stratégies mises en œuvre pour tenir compte (ou non) de l'inflation de zéros ont fondé de véritables « écoles » d'analyse en single-cell. Une de ces stratégies consiste à ajouter un paramètre supplémentaire à la distribution négative binomiale, qui va ajouter une probabilité pour que le count soit zéro, quel que soit sa valeur avec une négative binomiale seule. En reprenant l'exemple de l'urne, une fois le tirage effectué, certaines couleurs de billes disparaissent aléatoirement. Chaque couleur à une probabilité qui lui est propre de disparaître. Plus précisément,

une fois la valeur déterminée par la négative binomiale, l'expression est multipliée par un ou zéro selon une probabilité π . On parle dans ce cas de distribution ZINB (Vallejos et al., 2017), pour *Zero Inflated Negative Binomial*. Une autre stratégie est d'additionner les transcriptomes d'un ensemble de cellules proches, pour lisser l'impact des zéros (L. Lun et al., 2016). Dans ce cas nous nous retrouvons dans une table de comptage de pseudo-cellules ayant des propriétés proches d'échantillons de bulk RNA-Seq et pouvant donc être modélisées correctement par une loi négative binomiale. Une déconvolution est ensuite effectuée pour retomber sur une matrice de cellules uniques après traitement de la table de pseudo-cellules.

Quelle que soit la méthode de RNA-Seq, il est d'usage de réserver toutes les variables d'expression de gènes dans la table de comptage, et de stocker toutes les autres variables dans un tableau de données nommé table d'annotation des échantillons. Ce tableau contient par exemple des données de contrôle qualité ou bien encore des annotations qui renseignent sur le phénotype des échantillons.

1.2.2.3 La table de comptage brut est impropre à l'analyse de l'expression

Les tables de comptage de RNA-Seq ne sont pas statistiquement exploitables en sortie d'analyse primaire. En effet, elles contiennent de nombreuses sources de variations indésirables. Il est donc nécessaire d'identifier et d'éliminer ces sources de variation avant de pouvoir commencer à émettre des hypothèses biologiques à partir des analyses.

1.2.2.3.1 Filtrer la table de comptage

La première source de variation non désirable est due à la présence d'échantillons non comparables au reste de la collection. Le cas le plus fréquent est celui d'un échantillon contenant peu ou pas de reads alignés. La présence de ces échantillons peut provenir de problèmes techniques lors de leur manipulation : trop peu d'ARN au départ, ARN trop dégradé, ou encore un problème de réactif. En single-cell, les protocoles de type Drop-Seq (Macosko et al., 2015) utilisent un circuit microfluidique pour que des bulles contenant le matériel nécessaire à la

construction de la librairie encapsulent des cellules (**Figure 7a**). Les bulles ne rencontrant pas de cellules feront parties des échantillons vides. Quelle que soit la cause de leur présence, ces échantillons peuvent être écartés lors de l'analyse primaire, ou une fois la table de comptage obtenue. Dans ce cas la façon la plus simple de procéder est de se fier au nombre total de count (taille de librairie) de chaque échantillon. Une autre métrique utile est de calculer le nombre total de gènes exprimés. En effet à partir d'un certain nombre de count le nombre de gènes exprimés atteint un plateau (**Figure 10a**), signe que le transcriptome de l'échantillon a été correctement estimé par le RNA-Seq, il suffit alors de placer notre échantillon par rapport à ce plateau pour s'assurer de sa qualité. Peu importe les métriques utilisées, une méthode efficace est de placer chaque échantillon par rapport à la distribution de la métrique, ainsi il est possible de déterminer la partie de la distribution qui correspond à de mauvais échantillons (**Figure 10b-d**). Une façon plus simple consiste à fixer des valeurs seuils en dessous ou au-dessus desquelles un échantillon va être éliminé de l'analyse. Cette méthode peut être suffisante en bulk RNA-Seq. En effet selon la stratégie de librairie et le type d'échantillon, il existe des attentes a priori sur le nombre de count nécessaire à l'estimation du transcriptome. De plus il y a généralement moins d'échantillons en bulk qu'en single-cell, il est donc parfois difficile d'estimer la distribution des métriques de contrôle qualité en bulk RNA-Seq. Enfin nous pouvons effectuer du clustering ou de la réduction de dimension (cf. chapitres suivants), non pas à partir de l'expression des gènes mais à partir des métriques de contrôle qualité. Les mauvais échantillons seront facilement identifiables en un cluster et/ou une zone dans l'espace réduit. Les gènes peuvent aussi être filtrés. Ce filtrage se fait principalement avec des seuils relatifs à la moyenne d'expression et du nombre de fois où le gène est exprimé. Le but est de supprimer les gènes ne contenant pas ou très peu d'information, leur suppression permet principalement d'économiser du temps de calcul ou de gagner de la puissance statistique (cf. 1.2.2.4, en page 46).

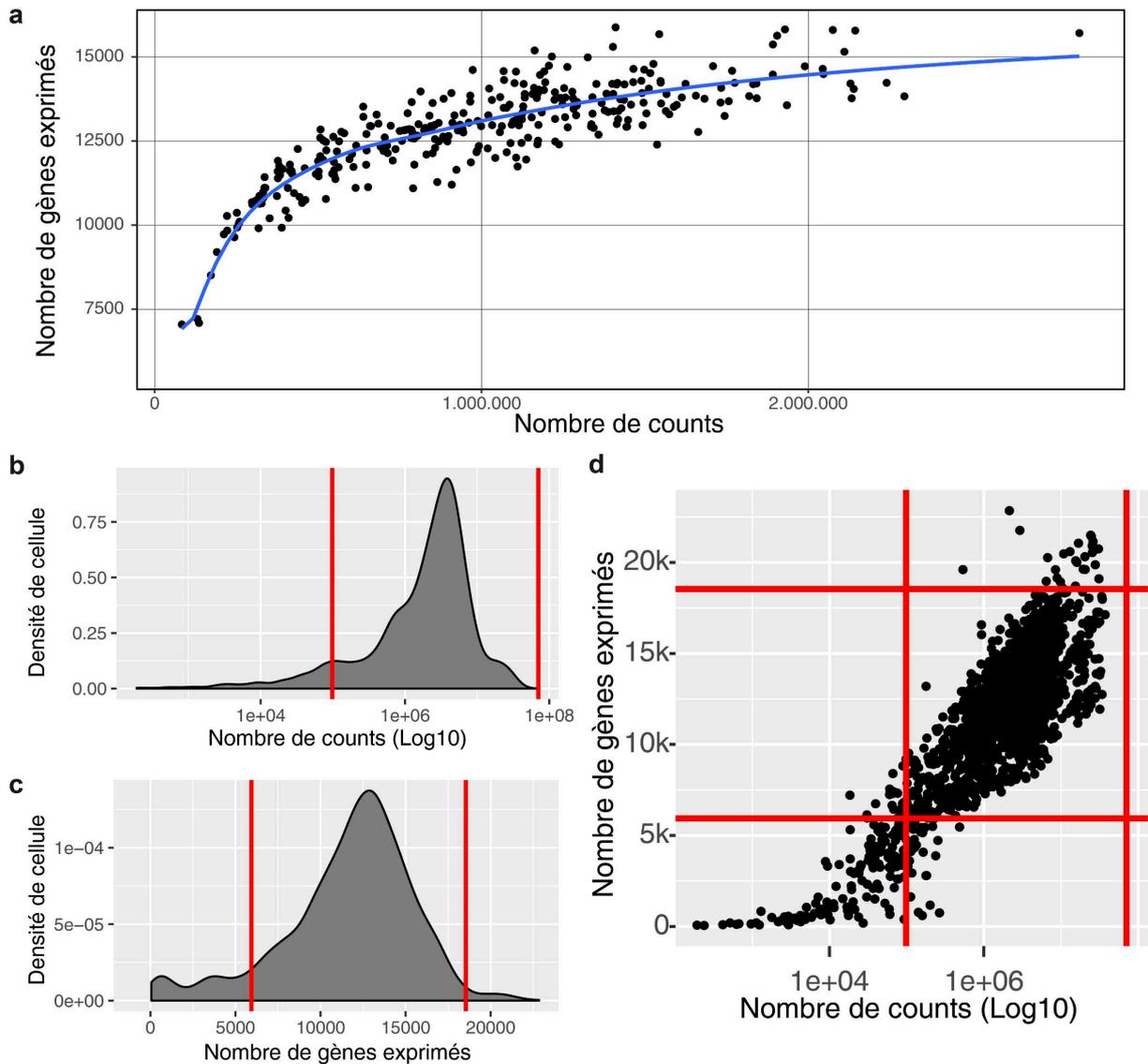


Figure 10 : Filtration des échantillons. **a.** Nombre de gènes exprimés en fonction du nombre de counts dans la table de comptage pour des échantillons de bulk RNA-Seq. La courbe bleue correspond à une régression locale **b-c.** Tracer la distribution d'une métrique de contrôle qualité, comme le nombre de count (**b**) ou le nombre de gènes exprimés (**c**) par cellule, permet d'identifier les modalités de la distribution et ainsi d'écartier les cellules qui ne seraient pas comparables aux autres. **d.** Nous pouvons aussi utiliser plusieurs métriques partitionner les données, avec des algorithmes de clustering ou encore manuellement avec des « gates », pour isoler les populations de qualité suffisante à la réalisation d'analyses. Les barres rouges représentent les seuils de filtrage. Les données sont celles utilisées et dans Kilens, Meistermann et al. 2018 (a) et Meistermann, Loubersac et al. (b-d).

1.2.2.3.2 Normaliser des données de RNA-Seq

Le deuxième problème avec la table de comptage est le fait que chaque échantillon a une taille de librairie qui lui est propre : le nombre de count d'un gène dans un échantillon n'est pas seulement dépendant de l'expression du gène mais aussi de la quantité de matériel séquencé dans cet échantillon (Bullard et al., 2010; Froussios et al., 2019). Nous avons donc besoin d'une étape de normalisation de la table de comptage.

La façon la plus simple de procéder est de diviser l'expression par la taille de la librairie puis de mettre à l'échelle l'expression par un facteur d'un million, on obtient des count par millions (CPM), aussi appelées reads par millions (RPM). Si nos counts correspondent à des UMI uniques on peut parler de UMI par million (UPM). Si chaque UMI représente un transcrit nous pouvons parler de transcrit par million (TPM). Dans les méthodes où un transcrit donne plusieurs reads il est nécessaire de prendre en compte la taille du gène si l'on veut passer des CPM aux TPM. En effet dans ce cas, plus un transcrit est long plus il va générer de fragments dans la librairie. Pour convertir les counts en TPM il faut alors diviser les counts bruts de chaque gène par un facteur consistant en la longueur de la somme totale des exons en kilobase, pour obtenir des reads par kilobase (RPK). Ensuite il suffit de diviser l'expression obtenue par la somme totale des expressions de l'échantillon et de multiplier par un million pour obtenir des TPM. Enfin nous pouvons inverser les deux dernières étapes, c'est-à-dire dire partir de CPM et appliquer la division par la somme des exons en kilobases, dans ce cas nous obtenons des reads par kilobase par million (RPKM) (Mortazavi et al., 2008).

Pour un gène i , un échantillon j et des counts brut C , dans les méthodes par UMI où un count est spécifique d'un transcrit nous avons donc :

$$CPM_{ij} = RPM_{ij} = UPM_{ij} = TPM_{ij} = \frac{C_{ij}}{\sum C_j} \times 10^6$$

Dans les méthodes où un transcrit donne plusieurs reads, avec L la longueur des exons en kilobases nous avons :

$$CPM_{ij} = RPM_{ij} = \frac{C_{ij}}{\sum C_j} \times 10^6$$

$$RPKM_{ij} = \frac{RPM_{ij}}{L_i}$$

$$TPM_{ij} = \frac{C_{ij}/L_i}{\sum \frac{C_j}{L_j}} \times 10^6 = \frac{RPKM_{ij}}{\sum RPKM_j} \times 10^6$$

Ces types de normalisation « par million » posent un problème majeur : plus un gène est exprimé plus il va écraser l'expression des autres gènes. Les normalisations par million se basent sur l'hypothèse fautive que tous les échantillons d'un même jeu de données possèdent le même nombre de transcrit au total. Ces normalisations rendent difficiles la comparaison d'expression entre des échantillons provenant de populations statistiques différentes. En revanche la normalisation d'un échantillon n'est pas dépendante du reste du jeu de données, ce caractère universel la rend utile lorsque l'on veut établir un profil de type cellulaire (Petryszak et al., 2014). Son caractère intuitif est aussi un avantage : il est facile de se représenter ce qu'est un transcrit pour un million. Ainsi les représentations graphiques d'expression provenant de RNA-Seq se font souvent en RPKM, TPM ou CPM. L'utilisation de la division par la taille du gène pour les bibliothèques où les transcrits sont morcelés entre plusieurs reads pose aussi un problème. Il a été montré que l'expression normalisée par TPM/RPKM dans ce cas n'est pas beaucoup plus utile, voir plus biaisée, que l'expression normalisée par d'autres stratégies (Bullard et al., 2010). Dans les faits, comparer les valeurs d'expression entre des gènes au sein d'un échantillon est une analyse peu courante, il est préférable de s'intéresser aux variations d'expression

entre deux ensembles d'échantillons (*fold-change*) ou de travailler sur des valeurs de co-expression.

Une autre façon de normaliser est d'ajouter des ARN exogènes en quantité connue dans la librairie, qui seront incorporés avec les ARNm endogènes dans la librairie. On parle alors de *spike-in* (Fardin et al., 2007). Il suffit ensuite de calculer un facteur de division de l'expression par échantillon de façon à ce que les valeurs de *spike-in* soient les mêmes pour tous les échantillons. Ce facteur de division de l'expression est appelé facteur de normalisation. Par exemple, nous pouvons calculer la valeur de spike in moyenne entre tous les échantillons, puis faire le ratio entre cette valeur moyenne et chaque valeur de spike in. Ce ratio nous donnera le facteur de normalisation. Cependant les spike-in reposent sur des hypothèses fortes (Robinson and Oshlack, 2010)., ainsi au moindre problème d'amplification par rapport au reste de librairie l'utilisation des *spike in* peut amener à calculer des facteurs de normalisation faux.

a

Matrice de comptage	Ech. 1	Ech. 2	Ech. 3	Pseudo-Ech.
Gène 1	1	2	1	$\sqrt[3]{1 \times 2 \times 1} \approx 1.26$
Gène 2	6	12	6	7.56
Gène 3	12	24	100	30.65
Gène 4	0	0	10	0.00

Matrice des ratios	Ech. 1	Ech. 2	Ech. 3	Pseudo-Ech.
Gène 1	$1/1.26 \approx 0.79$	1.59	0.79	1
Gène 2	0.79	1.59	0.79	1
Gène 3	0.39	0.78	3.26	1
Facteurs de normalisation	0.79	1.59	0.79	1

Matrice normalisée	Ech. 1	Ech. 2	Ech. 3
Gène 1	$1/0.79 \approx 1.26$	1.26	1.26
Gène 2	7.56	7.56	7.56
Gène 3	15.12	15.12	125.99
Gène 4	0	0	12.60

b

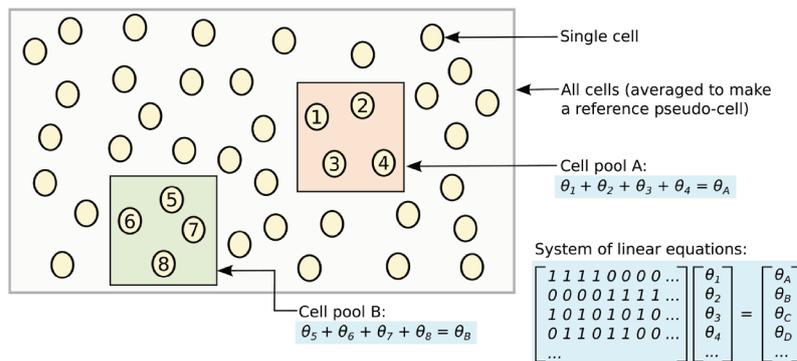


Figure 11 : Normalisations de RNA-Seq principalement utilisées durant ma thèse. a. Les étapes de normalisation par DESeq2. Dans ce cas nous calculons d’abord un pseudo-échantillon dont l’expression consiste en la moyenne géométrique de chaque gène. La moyenne géométrique a plusieurs intérêts dans ce cas : elle est plus robuste que la moyenne arithmétique à l’influence des valeurs extrêmes, et n’est pas sensible aux changements d’échelle. Autrement dit chaque échantillon va contribuer de façon égale dans le calcul du pseudo-échantillon. Le calcul de la moyenne géométrique contient un produit, ainsi au moins zéro l’expression dans le pseudo-échantillon du gène correspondant sera de zéro. Les gènes à zéro seront ensuite retirés du pseudo-échantillon. Pour les gènes restants nous calculons pour chaque échantillon un vecteur de ratio d’expression entre l’échantillon et le pseudo-échantillon, nous obtenons ainsi une matrice de ratio. Pour chaque échantillon nous prenons la médiane des ratios, cette médiane sera le facteur de normalisation de l’échantillon. **b.** Figure issu de L. Lun et al., 2016. Schéma de la méthode de déconvolution utilisée par la librairie *scraper*. La moyenne de toutes les cellules est calculée pour tout le jeu de données, on obtient ainsi une pseudo-cellule de référence. Ensuite, les cellules corrélées sont regroupées par pool. Les valeurs d’expression pour les cellules dans les données du pool A sont additionnées et normalisées par rapport à la référence pour obtenir un facteur de normalisation de pool (θ_A). Ce facteur est égal à la somme des facteurs θ_j pour les cellules j de 1 à 4. En répétant ceci pour de multiples pools, par exemple, le pool B, on aboutit à la construction d’un système linéaire qui peut être résolu pour estimer θ_j pour chaque cellule j .

Une des normalisations les plus populaires en bulk RNA-Seq est l'utilisation de la médiane des ratios comme facteur de normalisation, on parle de normalisation *Relative Log Expression* (RLE), c'est notamment l'approche utilisée par DESeq2 (Love et al., 2014) (**Figure 11a**). L'intuition derrière cette méthode est de trouver le « meilleur gène de ménage » par échantillons qui est au centre de la distribution des potentiels gènes de ménage. Ainsi le fait de supprimer les gènes contenant au moins un zéro n'est pas un problème puisque ces gènes ne peuvent pas être des candidats à la liste de gènes de ménage, cela rend l'étape de filtrage primordiale. En effet les mauvais échantillons contiennent plus de zéros, ce qui restreint la liste de gènes utilisées dans la normalisation. Cette normalisation est peu robuste dans le contexte du single-cell, en effet l'inflation de zéro et le grand nombre d'échantillons limitent drastiquement la liste des gènes ne contenant pas de zéros. Il existe d'autres façons de normaliser des données de bulk RNA-Seq, nous pouvons citer les *Trimmed Mean of M-values* (TMM) (Robinson and Oshlack, 2010), qui offrent des performances équivalentes à la normalisation RLE (Maza, 2016).

Les méthodes de normalisation développées pour le bulk RNA-Seq ont été largement utilisées pour les données de scRNA-Seq, durant les premières années après l'émergence du scRNA-Seq. Cependant, ces normalisations, à l'instar de la normalisation RLE de DESeq2, ne permettent pas de normaliser les données de façon robuste (Vallejos et al., 2017). Plusieurs approches de normalisation spécifiques au single-cell ont ainsi été développées, et se basent sur les approches développées pour prendre en compte l'inflation de zéros et /ou la nature hétérogène des données. Ainsi la normalisation de scran se base sur une addition de transcriptomes de cellules corrélées (**Figure 11b**) (L. Lun et al., 2016). Un facteur de normalisation est ainsi calculé pour chaque pseudo-cellule puis déconvolué pour repasser à l'échelle de la cellule unique. L'approche de diviser l'expression de tous les gènes d'une cellule par un facteur de normalisation se fait sous l'hypothèse que tous les gènes sont affectés de la même façon par les variations d'expression indésirables entre échantillons. SCnorm (Bacher et al., 2017) utilise un algorithme de régression par quantile pour prendre en compte les cas de figures où cette hypothèse est fautive. La librairie R scone (Cole et al., 2019) est particulièrement utile pour normaliser des données de scRNA-Seq. En effet elle permet de tester plusieurs

types de normalisation spécifiques ou non au single-cell, et fourni des métriques pour choisir la plus adaptée au jeu de données étudié.

Une transformation des données est souvent effectuée après la normalisation des données. Celle-ci permet de mettre tous les gènes sur une même gamme dynamique d'expression. Sans cette transformation les gènes les plus exprimés sont les responsables majeurs des variations dans les données, puisque la dispersion de l'expression augmente avec la moyenne d'expression dans une loi négative binomiale. Cependant cette variation n'est pas forcément due à des effets biologiques. La transformation la plus courante pour mettre tous les gènes sur une même gamme dynamique d'expression est la suivante :

$$x' = \log_2(x + 1)$$

Le logarithme « aplati » l'expression de plus en plus au fur et à mesure qu'elle augmente. La moyenne d'expression n'est alors plus intrinsèquement liée à la dispersion, et la distribution s'approche de la loi normale. Le choix de la base 2 pour le logarithme est consensuel, l'utilisation d'une autre base ne changerait pas les propriétés de la transformation. Nous ajoutons un avant la transformation pour éviter les valeurs d'expression négatives voire infiniment négatives si $x = 0$. Néanmoins, cette transformation biaise l'expression en faveur des valeurs faibles. Ainsi la *variance stabilized transformation* ou VST de la librairie DESeq2 (Love et al., 2014) et la transformation *voom* de la librairie limma (Ritchie et al., 2015) permettent de transformer la variance de façon plus robuste que la transformation logarithmique.

1.2.2.3.3 Corriger les effets de batch

Même une fois normalisées les données peuvent être biaisées par des conditions expérimentales différentes entre groupes d'échantillons. On parle alors d'effet de batch. Le cas le plus courant est un jeu de données constituées de plusieurs préparations de librairie. Ces effets de batch sont inévitables lors de la fusion de différents jeux de données, maîtriser les effets de batch est donc indispensable à la méta-analyse de données de RNA-Seq.

Une première méthode pour supprimer les effets de batch est d'effectuer une régression linéaire de l'expression en fonction des batch connus, qui correspondent à une variable qualitative dans la table d'annotation des échantillons. Grâce aux propriétés de la régression linéaire nous pouvons obtenir l'expression théorique à batch égal (Ritchie et al., 2015). Une deuxième méthode est d'effectuer une réduction de dimension linéaire et de trouver les axes de variation qui discriminent les batch, pour ensuite supprimer la contribution de ces axes aux données. Enfin une méthode populaire, ComBat, estime les effets de Batch par une régression, et les régularise par une approche bayésienne empirique (Johnson et al., 2007) (cf. le détail du fonctionnement de DESeq2 ci-dessous pour en savoir plus sur la régularisation par une approche bayésienne empirique). Un point commun entre toutes ces méthodes est qu'il doit exister au moins une population commune connue entre chaque batch. Plus il existe de populations en commun, mieux l'effet de batch sera estimé, et plus la correction sera efficace. Nous fournissons en réalité à ces méthodes une matrice de design, et nous pouvons donc indiquer les variations d'intérêt que l'on souhaite conserver lors de la correction.

Ces méthodes sont cependant peu performantes en single-cell, ainsi de nouvelles approches ont fait leur apparition. La librairie Seurat utilise la corrélation canonique pour identifier des sources de variations communes aux batches (Satija et al., 2015). Une autre méthode est d'identifier les paires de cellules les plus corrélées entre deux jeux de données pour calculer l'effet de batch. Cette méthode est nommée *mutual nearest neighbors correction* (Haghverdi et al., 2018) et se base sur l'hypothèse qu'il existe une probabilité forte pour avoir deux cellules quasiment identiques entre deux jeux de données si le nombre d'échantillon est suffisamment élevé. Si ces deux cellules sont identiques alors les seules variations restantes sont dues aux variations indésirables. Il est à noter que théoriquement il n'y a pas besoin d'y avoir de population en commun pour que ces méthodes soient efficaces.

1.2.2.4 Inférer une liste de gènes différentiellement exprimés

Le profilage de l'expression des gènes a souvent pour but de savoir si des gènes ont une valeur centrale d'expression différente entre deux populations d'échantillons. Ces gènes sont les DEG (gènes différentiellement exprimés ou *differentially*

expressed genes), et représentent souvent l'enjeu de l'analyse en bulk RNA-Seq. Usuellement les jeux de données de bulk RNA-Seq contiennent relativement peu d'échantillons par population étudiée, pour un nombre très élevé de variables. Ce constat rend crucial le choix du test statistique utilisé pour déterminer les DEG. Parmi les méthodes les plus performantes en bulk RNA-Seq se trouve DESeq2 (Love et al., 2014) qui se base sur distribution négative binomiale. En plus de la modélisation par négative binomiale, la particularité de DESeq2 pour diminuer le risque de faux négatif se trouve dans sa façon de contrôler le problème de surajustement (*overfitting*) (Chicco, 2017). Le surajustement est un problème fondamental de statistique, et consiste en la question suivante : « Est-ce que notre modèle est utile en dehors du jeu de données à partir duquel il a été établi ? » Plusieurs méthodes existent pour minimiser le risque d'ajustement, celle utilisé par DESeq2 est la régularisation des paramètres. En effet, à deux étapes du calcul de la p-valeur d'expression différentielle, des paramètres sont régularisés par rapport à leur première estimation, en se basant sur une loi de distribution estimée par les données, et non a priori comme une régularisation par une approche bayésienne classique. On parle alors de régularisation par une approche bayésienne empirique. La première de ces étapes est l'estimation du paramètre de dispersion α de la loi négative binomiale pour chaque gène. Ce paramètre est estimé une première fois par maximum de vraisemblance : plusieurs valeurs de paramètres sont évaluées et on retient celui qui est le plus conforme aux données. Cette première estimation est qualifiée de dispersion empirique. Dans une distribution négative binomiale la dispersion est reliée à la moyenne d'expression, ainsi deux gènes avec des moyennes d'expression proches devraient avoir une dispersion proche. En se basant sur ce principe, une régression de la dispersion empirique sur la moyenne d'expression est effectuée (**Figure 9**), les valeurs de dispersion ajustées à la régression sont dénommées dispersions théoriques. Enfin, le paramètre de dispersion retenu pour chaque gène est la dispersion théorique, sauf dans le cas où la dispersion empirique est trop éloignée de la dispersion théorique, le gène est alors sur-dispersé ou sous-dispersé.

La deuxième régularisation de paramètre se trouve au niveau de l'estimation des *log2 fold-change* (LFC). Le *fold-change* d'un gène pour deux populations de cellules

est un ratio des moyennes d'expression de chaque population. Utiliser un logarithme de base 2 sur un ratio a pour conséquence de pouvoir inverser le numérateur et le dénominateur du ratio sans changer la valeur absolue du résultat, seul le signe se verra affecté. Les valeurs de LFC sont tout d'abord modélisées grâce à une régression suivant le modèle linéaire généralisé (GLM) à partir de la matrice de design contenant au moins la variable d'intérêt, qui peut correspondre aux populations biologiques ou à une variable quantitative. Le coefficient directeur de la droite sert à estimer le LFC. D'autres variables telles que les batchs peuvent être renseignés dans la matrice de design de la régression pour corriger les LFC. Une fois le LFC de chaque gène estimé, une loi normale est ajustée aux LFC de l'ensemble des gènes. Cette distribution normale servira d'a priori pour une nouvelle régression GLM. Cette régression va permettre d'estimer le LFC finale et d'obtenir une valeur d'erreur standard associée à chaque LFC. Avec cette procédure, moins il y a d'information disponible, plus la LFC sera concentrée autour de zéro. Ainsi les LFC estimés par DESeq2 ont moins tendance à être extrêmes pour des gènes faiblement exprimés qu'une estimation basée sur les ratios de moyenne d'expression des échantillons. Enfin le LFC régularisé et l'erreur standard associée serviront à calculer la p-valeur de chaque gène à partir d'un test de Wald.

Les outils de mises en évidence de DEG spécifiques au single-cell ne semblent pour l'instant pas offrir un gain élevé en termes de statistique par rapport aux approches développées pour le bulk RNA-Seq (Wang et al., 2019). Néanmoins, ces approches sont souvent bien plus lentes que celles développées spécifiquement pour le scRNA-Seq, ce qui les rend inutilisables pour analyser des jeux de données de plusieurs milliers de cellules. Beaucoup de ces outils ont comme particularité de prendre en compte les dropout, c'est le cas pour SCDE (Raj and van Oudenaarden, 2008). Il semble cependant que retrouver les DEG est rarement une analyse pertinente en single-cell. Ainsi un des buts les plus communs au single-cell est de définir de nouvelles populations de cellules et de définir les gènes qui en sont les marqueurs. Dans ce cas on préférera des analyses de spécificité de gène à la population face à toutes les autres, par exemple à l'aide de courbes ROC.

1.2.2.5 P-valeur et analyses multivariées

Pour toutes les analyses où l'on obtient une p-valeur par gène, telles que l'analyse d'expression différentielle, il se pose un problème d'inflation du risque α de 1^{ère} espèce. Supposons un jeu de données où l'expression différentielle de chaque gène est sous l'hypothèse nulle. Autrement dit, tous les échantillons proviennent d'une seule population, et où les différences d'expression que l'on observe sont dues à des fluctuations d'échantillonnage. Ce jeu de donnée contient 1000 gènes, à chacun est attribué une p-valeur d'expression différentielle entre deux groupes d'échantillons tirés aléatoirement. Dans ce cas si l'on prend un risque α de 1% par gène, la probabilité de ne pas avoir de faux positif est de 99% ($1 - 0.01$). La probabilité de ne s'être jamais trompé pour 1000 gènes indépendants est alors de $(0.99)^{1000} \approx 4.32 \times 10^{-5}$. Autrement dit le risque alpha « global » de s'être trompé au moins une fois est quasiment de 100% ($1 - 4.32 \times 10^{-5}$). Une première façon de limiter ce risque est de diminuer le nombre de tests statistiques en sélectionnant les gènes les plus susceptibles d'être intéressants (cf. 1.2.2.6, en page 51). Une seconde méthode est d'ajuster les p-valeurs en fonction du nombre de tests effectués. L'ajustement le plus simple est celui de Bonferroni et consiste à multiplier chaque p-valeur par le nombre de test effectués. Néanmoins, cette approche est trop conservatrice pour le nombre de tests effectués en RNA-Seq. Les p-valeurs ajustées par la procédure de Bonferroni sont trop élevés pour garantir un taux de faux négatif satisfaisant. Ainsi un des ajustements les plus utilisés en RNA-Seq est la procédure de Benjamini-Hochberg (Benjamini and Hochberg, 1995) et se calcule de la façon suivante : soit p la p-valeur du gène, k le rang de la p-valeur et n le nombre de tests effectués.

$$p. \text{ajustée} = \frac{p \times n}{k}$$

Ainsi plus la p-valeur du gène est significative, plus elle va être pénalisée. Cette procédure permet de contrôler le risque α tout en ayant un taux de rejet moins élevé que l'ajustement de Bonferroni.

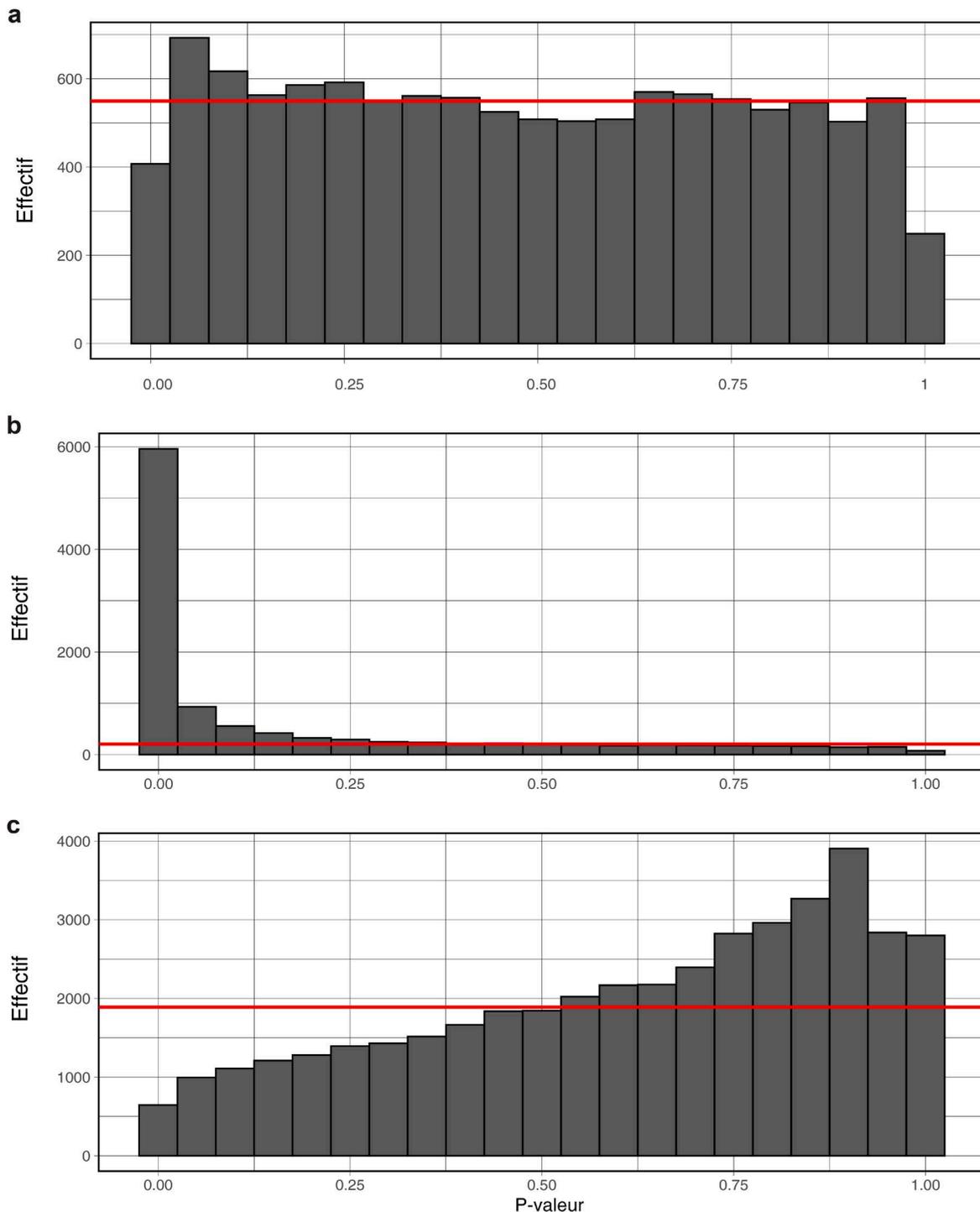


Figure 12: Distributions typiques d'un grand nombre de p-valeurs. Les valeurs proviennent de différentes analyses d'expression différentielle provenant de données de DGE-Seq. La barre rouge indique la valeur médiane des p-valeurs. **a.** La p-valeur suit une loi de distribution uniforme sous l'hypothèse nulle. Une distribution proche de la distribution uniforme comme présenté ici est donc un indice que nous sommes peut-être dans un cas proche de H_0 , c'est-à-dire que nos échantillons proviennent de la même population: il n'existe pas de gènes différentiellement exprimés. **b.** Distribution typique de la p-valeur lorsqu'il existe des gènes différentiellement exprimés: les processus biologiques entre nos deux groupes d'échantillons sont vraisemblablement différents. **c.** Distribution des p-valeurs lorsque le modèle statistique n'est pas fiable. Dans ce cas précis, le nombre d'échantillon était trop faible pour que DESeq2 obtienne un taux de faux négatif satisfaisant.

Le grand nombre de p-valeur en RNA-Seq peut aussi être une force. En effet nous pouvons nous en servir pour estimer la distribution des p-valeurs non ajustées. Sous l'hypothèse nulle et si notre modèle est correct alors la distribution des p-valeurs doit suivre une loi de distribution uniforme (**Figure 12**). Si la distribution des p-valeurs est uniforme nous pouvons suspecter que les différences observées entre tous les gènes sont en grande majorité dues aux fluctuations d'échantillonnages. Si la distribution n'est pas uniforme, et la densité de probabilité augmente autour de zéro alors nous pouvons conclure qu'il existe des variations qui ne peuvent pas s'expliquer seulement par les fluctuations d'échantillonnage. Il existe alors deux cas de figures : soit le modèle utilisé ne convient pas aux données, car le taux de faux positif est trop élevé, soit il existe réellement des variations significatives. Toutes les autres formes de distribution de p-valeurs mettent en évidence un mauvais choix de modèle.

1.2.2.6 Sélectionner des gènes d'intérêt

Une façon de résoudre les problèmes liés à un haut nombre de dimensions est de réduire ce nombre de dimension. La façon la plus simple d'y parvenir est de retenir un nombre restreint de variables. Cette sélection peut s'effectuer a priori de l'analyse et correspondre à des gènes d'intérêts mis en évidence dans la littérature, généralement une liste de marqueurs reliés à la question biologique. Cependant cette méthodologie peut s'avérer laborieuse, en effet il est nécessaire de sélectionner au minimum plusieurs dizaines de gènes afin de couvrir le transcriptome de façon suffisante. Un nombre suffisant de gènes est aussi nécessaire pour éviter l'influence de phénomènes spécifiques au single-cell, tel que les dropouts et les phénomènes d'expression stochastiques. Enfin une sélection manuelle entraîne inévitablement un risque de biais de confirmation des hypothèses déjà connues.

Les méthodes de sélection de variables non ou semi-supervisées permettent d'éviter ces risques. Le principe général est d'isoler la variation des gènes due à une expression différentielle entre les populations d'échantillon. Plus ce type de variation est élevé, plus le gène est potentiellement intéressant pour l'analyse. Un principe courant pour mettre en évidence cette variance fait intervenir l'étape de

DESeq qui consiste à effectuer une régression de la dispersion empirique sur la moyenne d'expression (**Figure 9**). Le rapport dispersion empirique sur dispersion théorique, ou encore le résidu de la régression sera un score direct de surdispersion du gène.

Des méthodes alternatives de sélection de gènes peuvent être utilisées telles que l'analyse des gènes contribuant le plus aux axes des composantes principales (cf. 1.2.2.7 ci-dessous) ; ou des méthodes basées sur la mise en évidence de gènes différentiellement exprimés entre clusters d'échantillons comme l'algorithme *dpFeature* (Qiu et al., 2017).

1.2.2.7 Les méthodes de réductions de dimension

Une autre solution pour réduire le nombre de dimension du problème est de compresser l'information, grâce aux méthodes dites de réduction de dimension. La première utilité de la réduction de dimension est d'offrir un point de vue global sur le nuage de points constitué par les données dans une seule projection, généralement en dimension 2.

La méthode de réduction de dimension linéaire la plus utilisée en biostatistique est l'analyse en composante principale ou ACP (Pearson, 1901). Les réductions de dimensions linéaires ont comme particularité de ne pas déformer l'espace de départ, mais d'en offrir un point de vue spécifique. Ce point de vue est pour l'ACP les axes (ou composantes) où le nuage de point est le plus dispersé (**Figure 13b**). L'espace calculé par ACP est composée de nouvelles variables, appelées composantes. Les coordonnées des échantillons dans les composantes principales consistent en un mélange des anciennes variables. La « proportion » de ces mélanges est donnée par la matrice des contributions des gènes aux axes (*gene eigenvectors*). L'existence de cette matrice de contribution représente un avantage certain face à d'autres méthodes de réduction de dimensions. Elle permet d'ajouter des échantillons sans recalculer l'espace réduit, et de passer à volonté des coordonnées dans l'espace de départ à celle de l'espace réduit. Chaque composante est décorrélée des autres et représente un axe de variation unique dans les données. Ainsi une composante qui représente un axe de variation indésirable peut

simplement être retirée du jeu de données grâce à une ACP. Les gènes contribuant le plus à des composantes intéressantes peuvent être aussi sélectionnés, dans ce cas l'ACP peut faire office de méthode de sélection de variables.

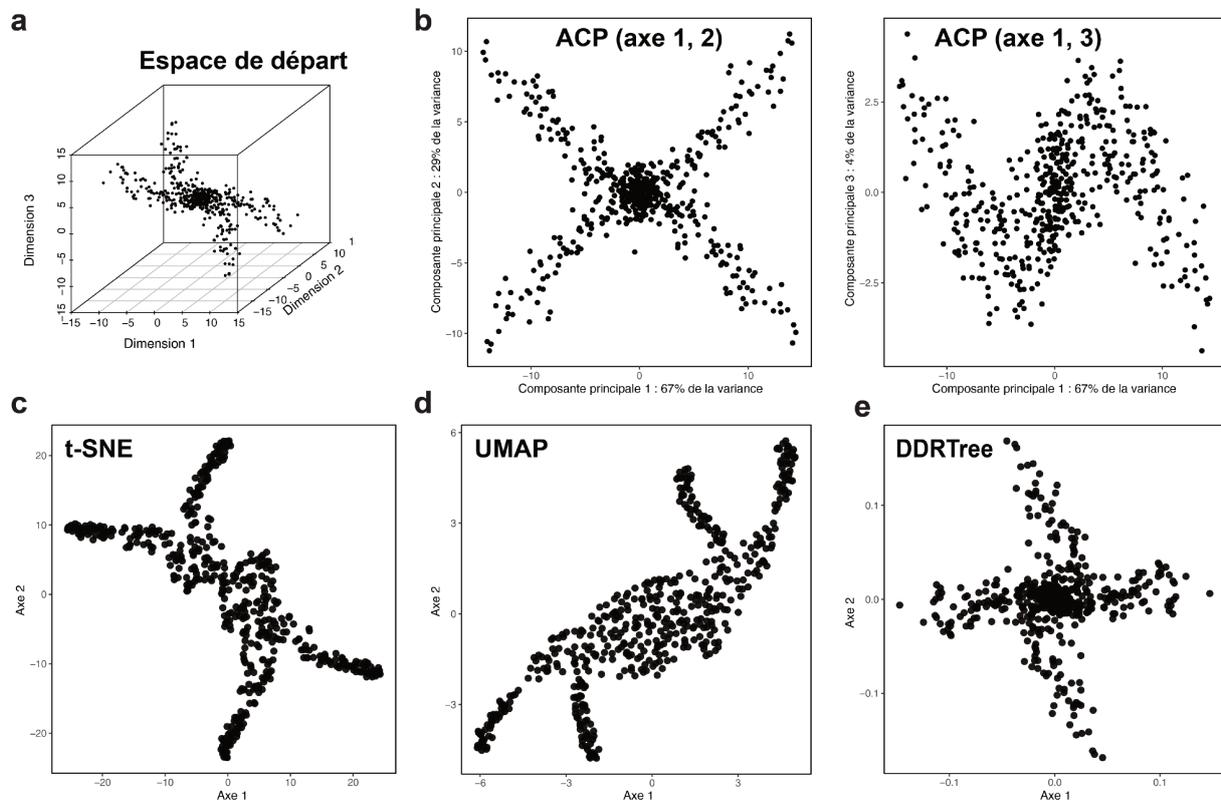


Figure 13 : Test de méthodes de réduction de dimension à l'aide d'une simulation. Les meilleurs paramètres de chaque algorithme ont été retenus après plusieurs essais. **a.** Espace de départ à trois dimensions, Le nuage de points forme un « X » dont les branches sont « pliées ». Le centre du X est plus dense en échantillons. Le bruit suit une loi normale. **b.** Résultat de l'ACP pour le couple de composante principale 1, 2 (à gauche) et 1, 3 (à droite). **c.** Réduction en deux dimensions par t-SNE (perplexity = 30). **d.** Réduction en deux dimensions par UMAP (n_neighbors = 200, min_dist = 0.3). **e.** Réduction en deux dimensions par DDRTree (paramètres par défaut).

Néanmoins, la complexité des jeux de données de transcriptomique a particulièrement augmenté depuis l'apparition du scRNA-Seq (Svensson et al., 2018), si bien que l'aspect linéaire de l'ACP n'est plus suffisant pour décrire les données avec une restriction en dimension 2 ou 3 (Andrews and Hemberg, 2018). L'ACP reste un algorithme central de l'analyse en single-cell. En effet, elle permet de simplifier l'espace des données en retirant les axes de variations contribuant le moins à la dispersion du nuage de points. Généralement environ quelques dizaines de variables sont retenues dans l'espace de l'ACP ; ce qui a pour conséquence de réduire drastiquement le temps de calcul des analyses ultérieures

qui se baseront sur ce nouvel espace. Cette utilisation indirecte de l'ACP est adoptée par Seurat (Satija et al., 2015). Cependant, les composantes principales sont liées à l'espace de départ par un ensemble de coefficients de contribution de gènes difficiles à interpréter. Il est donc nécessaire de revenir à l'espace de départ lors des analyses centrées sur les gènes, par exemple lors de la mise en évidence des gènes différentiellement exprimés. Des stratégies alternatives autour de l'ACP ont aussi permis de se servir à nouveau de cette méthode en tant que projection globale de données de scRNA-Seq. Dans ce cas, la matrice des contributions est calculée avec une sélection d'échantillons la plus hétérogène possible, puis tous les échantillons sont projetés (Lall et al., 2018). Ainsi le type cellulaire le plus fréquent ne monopolise pas la variation des données lors de la création de la matrice des contributions.

La t-SNE (t-Distributed Stochastic Neighbor Embedding) (Maaten and Hinton, 2008) a été une des méthodes de réduction la plus utilisée pour projeter des jeux de données de scRNA-Seq. Cette méthode déforme l'espace de départ, et est donc classée dans la catégorie des méthodes de réduction de dimension non linéaires. La construction des t-SNE est itérative et la transformation est non réversible. L'algorithme calcule tout d'abord une matrice de similarité entre tous les échantillons. Ce score de similarité décroît rapidement entre deux échantillons en fonction de la distance euclidienne, selon la fonction densité d'une loi normale. Plus une région de l'espace de départ est dense en échantillon, plus la gaussienne est aplatie (**Figure 14**), et la similarité décroît donc moins vite en fonction de la distance. Ainsi les échantillons au sein d'ensembles dispersés vont se regrouper et les échantillons dans les régions plus denses vont s'espacer dans la projection finale. La recherche de gradient optimale de similarité pour un échantillon est directement liée au paramètre de perplexité de la t-SNE. Plus ce paramètre est élevé, moins le gradient décroît vite pour l'ensemble des échantillons. Ainsi une perplexité plus élevée va apporter une meilleure préservation de la structure globale du nuage de points. Les échantillons sont ensuite projetés avec des coordonnées aléatoires dans l'espace de la projection et une matrice de similarité est calculée pour cet espace. Cette fois, le gradient ne décroît pas selon une densité loi normale mais selon une densité de loi de distribution de Student (*t-distribution*, à l'origine du « t » de t-SNE). À chaque itération, les coordonnées des échantillons dans la projection vont être

progressivement modifiés, de façon à ce que les deux matrices de similarités se ressemblent de plus en plus. La t-SNE « apprend » comment placer les échantillons dans l'espace de la projection (**Figure 13c**). La t-SNE n'est néanmoins pas exempte de défauts. L'algorithme est lent pour des jeux de données de la dimension du scRNA-Seq, ainsi une ACP ou une sélection de variables sont nécessaires au préalable de la t-SNE. Enfin même avec une perplexité élevée le t-SNE a tendance à ne pas préserver les structures globales de l'espace de départ. Ainsi les groupes (ou cluster) de cellules sont préservés mais pas l'agencement de ces groupes de cellules. Ce dernier point est bien souvent à l'origine de mauvaises interprétations concernant la t-SNE.

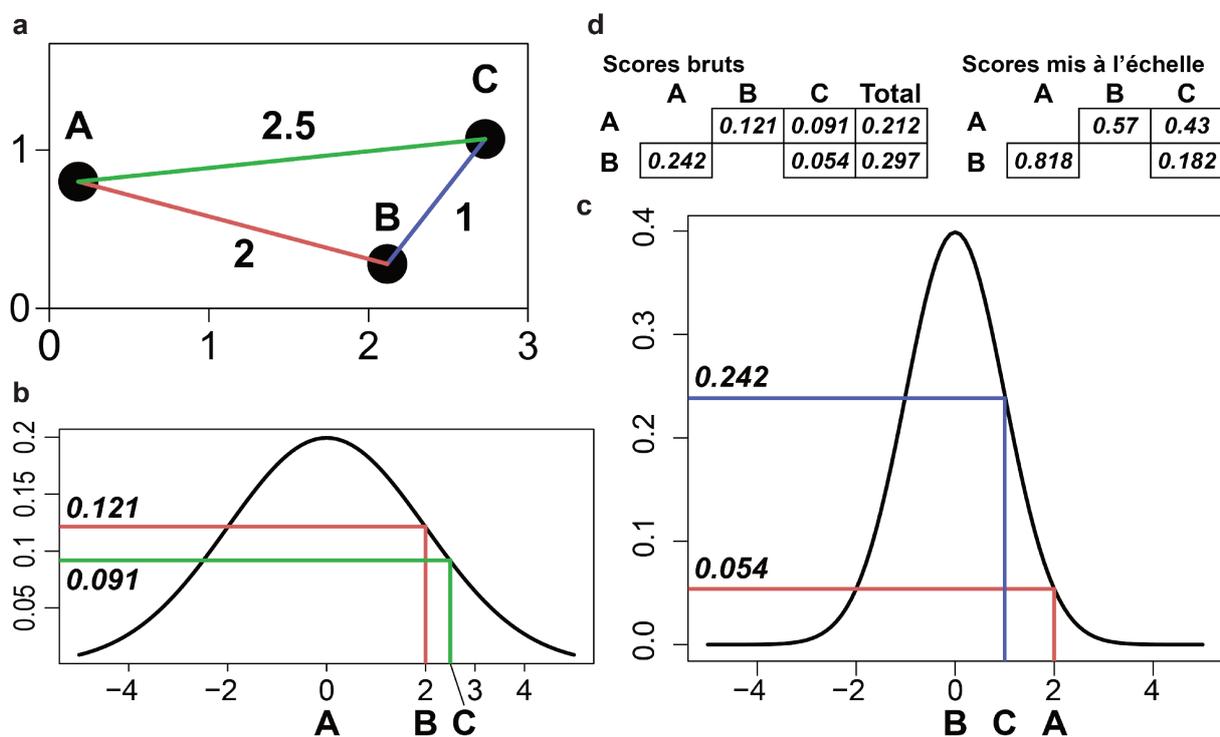


Figure 14 : Calcul du score de similarité de la t-SNE dans l'espace de départ. **a.** Espace de départ de dimension deux contenant 3 points. La distance euclidienne est indiquée. **b-c.** Mesure du score de similarité pour le point A (**b**) et le point B (**c**). Le score décroît suivant une fonction de densité de la loi normale, en fonction de la distance euclidienne. La région autour de A est moins dense que la région autour de B, la gaussienne est donc plus « aplatie » (la dispersion de la loi normale augmente). **d.** Extrait de la matrice de similarité. Le score calculé est divisé par la somme des scores mesuré pour le point, de façon à ce que la somme des scores du point fasse 1. Cela va permettre de comparer les scores de régions de densité différentes.

La UMAP (Uniform Manifold Approximation and Projection) (McInnes et al., 2018) est une méthode proche de la t-SNE mais qui corrige certains de ses défauts. La UMAP est plus rapide et conserve mieux les structures globales (**Figure 13d**). Cette

méthode de réduction de dimension semble remplacer la t-SNE en termes de popularité. La UMAP diffère de la t-SNE notamment par l'utilisation de fonctions différentes pour estimer la matrice de similarité et apprendre la meilleure représentation en basse dimension. La première étape de la UMAP est de construire un graphe des k plus proches voisins dans l'espace de départ, où chaque nœud est un échantillon et chaque arrête représente un lien de proximité. L'existence ou non de ces arrêtes et leur poids va directement définir la matrice de similarité. Le principal argument de la UMAP est donc le réglage du nombre de plus proches voisins lors de la création de ce graphe. Plus cette valeur est élevée plus les échantillons sont influencés par des échantillons distants. Ainsi une valeur de plus proche voisin élevée va mieux préserver les structures globales de l'espace de départ.

La méthode de réduction de dimension adoptée par l'un des principaux outils que j'ai utilisé, Monocle2 (Qiu et al., 2017), se nomme DDRTree (**Figure 13e**). Cette méthode est une réduction de dimension non linéaire de la famille des *reversed graph embedding*. La première étape de cette méthode consiste en l'apprentissage d'un graphe principal au sein de l'espace de départ. Ce graphe principal est un arbre qui ne passe pas par les échantillons, mais plutôt par les barycentres d'échantillons proches. Cet arbre est ensuite transposé dans l'espace de basse dimension ; et les échantillons sont projetés autour de l'arbre de façon à ce que leur position par rapport à l'arbre dans l'espace en haute dimension soit respectée.

De nombreuses autres approches de réduction de dimension sont utilisées en single-cell. Parmi elles se trouvent les auto-encodeur (Wang and Gu, 2018), qui permettent d'entraîner des réseaux de neurones à décrire des données de single-cell.

1.2.2.8 Résoudre le problème de classification par le partitionnement

Réduire la complexité d'un jeu de données peut aussi passer par son partitionnement. Dans ce cas les données semblables seront regroupées en sous-ensembles, dénommées classes ou clusters. Le but est d'avoir une partition avec un nombre de clusters à la fois suffisant pour décrire les variations dans le jeu de

données, mais aussi assez restreint pour que la partition soit facilement interprétable. Dans un partitionnement optimal, Les variations au sein des clusters (inertie intraclasse) doivent être minimisées et les variations de données entre les clusters (inertie interclasse) doivent être maximisées.

Les algorithmes de partitionnement (ou clustering) sont largement utilisés en RNA-Seq, notamment pour partitionner les variables ou les échantillons. Dans le cas des variables, le but est de les regrouper en clusters de gènes co-exprimés, souvent appelées modules de gènes. Résumer l'expression du transcriptome par module de gènes permet une compréhension intuitive du jeu de données (**Meistermann, Loubersac et al. Figure 4A**). Le clustering en bulk RNA-Seq sur les échantillons est souvent utilisé comme contrôle que les populations d'origine a priori des échantillons correspondent bien à des clusters. L'importance du clustering est plus centrale en scRNA-Seq, en effet les clusters de cellules correspondront directement aux populations de cellules identifiées. Une grande partie de la recherche de nouvelles méthodes d'analyse de scRNA-Seq se concentre donc sur le clustering des cellules ; le but visé est la création d'algorithmes qui permettent d'identifier les types et sous types cellulaires présents dans les données.

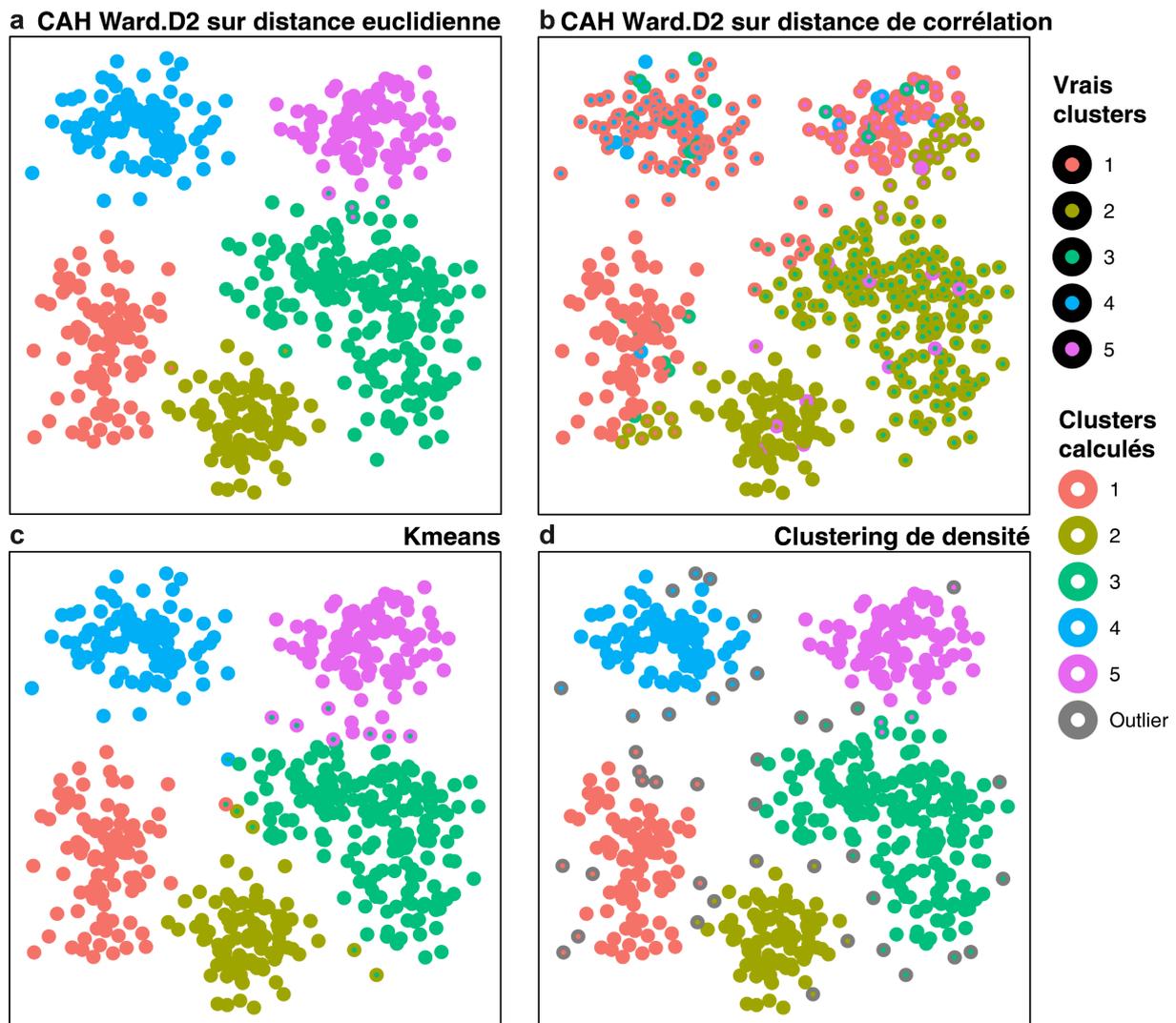


Figure 15 : Test de méthode de partitionnement à l'aide d'une simulation. Chaque cluster a été généré par un processus en dimension 2 : les échantillons d'un cluster sont placés aux mêmes coordonnées ou en suivant une droite. Puis un bruit aléatoire suivant une loi normale a été ajouté à chaque échantillon. **a.** Partitionnement à $k = 5$ à partir d'une classification ascendante hiérarchique sur une distance euclidienne. L'algorithme de CAH est le critère de Ward. **b.** Partitionnement à $k = 5$ à partir d'une CAH sur une distance de corrélation de Pearson. L'algorithme de CAH est le critère de Ward. A noter que la distance de corrélation de Pearson n'est pas adaptée à un clustering sur deux dimensions. **c.** Kmeans à $k = 5$. **d.** Clustering basé sur la densité, la distance de « saut » ε de l'algorithme a été paramétrée pour faire apparaître 5 clusters. Les échantillons isolés sont marqués comme *outliers*.

1.2.2.8.1 Choisir une métrique de distance

Partitionner les données se fait en deux étapes : calculer une matrice carrée de distance (ou de similarité) des observations (variables ou échantillons), puis utiliser un algorithme pour ordonner et/ou détecter les sous éléments semblables dans cette matrice. Dans ce cas particulier, une matrice de similarité peut aussi être vue comme une matrice d'adjacence d'un graphe décrivant la proximité entre les

observations. Le choix de la métrique utilisée dans la matrice de distance/similarité a une incidence primordiale dans les résultats du clustering. Une des métriques de distance les plus communes est la distance euclidienne (**Figure 15a**), Elle peut cependant donner de mauvais résultats dans un espace à grand nombre de dimension, en raison de propriétés géométriques contre intuitives d'un tel espace (Clarke et al., 2008). Il est donc nécessaire de réduire le nombre de dimension avant d'utiliser une distance euclidienne. Il est à noter que le transcriptome contient peu de variables indépendantes : de très nombreux gènes sont coexprimés. Dans ce cas précis la distance euclidienne reste relativement efficace par rapport à un cas où toutes les variables seraient indépendantes. La distance de Manhattan est proche de la distance euclidienne d'un point de vue mathématique. Il a été montré qu'elle peut surpasser la distance euclidienne dans les espaces à grand nombre de dimensions (Aggarwal et al., 2001). D'autres distances restent efficaces dans l'espace du transcriptome, notamment celles basées sur des coefficients de corrélation. Ces distances ont aussi pour avantage de fonctionner avec des données non normalisées. Néanmoins, les distances de corrélation ont des propriétés contre intuitives, surtout à un faible nombre de dimensions (**Figure 15b**). Malgré cela, les coefficients de corrélations sont une mesure qui permet d'approcher le concept de co-expression, et ils sont donc particulièrement utilisés dans le clustering de gènes. L'information mutuelle est une autre mesure qui permet d'approcher le concept de co-expression. Les coefficients de corrélation et d'information mutuelle s'avèrent souvent complémentaires selon la nature des données et le type de clustering choisi (Song et al., 2012).

1.2.2.8.2 Choisir un algorithme de partitionnement

Un des algorithmes les plus classiques est le clustering par *k-means* (ou *k-moyennes*), qui fait partie de la famille des partitionnements basées sur les barycentres. L'idée est de choisir *k* points dans l'espace, chaque point sera le barycentre d'un cluster. Chaque point est ensuite attribué à un cluster en fonction du barycentre le plus proche. Dans le cas des *K-means*, l'algorithme est itératif. Les barycentres sont au départ placés aléatoirement, puis déplacés jusqu'à minimiser la somme des distances entre chaque point et leur barycentre. Il existe au moins deux

défauts majeurs dans cette méthode. Le premier est que chaque cluster s'inscrit dans une ellipse, certaines configurations de clusters peuvent donc ne pas être estimés correctement (**Figure 15c**). Le deuxième problème est de déterminer le meilleur k , ce qui nécessite par exemple d'exécuter l'algorithme avec différents k pour retenir le meilleur choix.

Les algorithmes de classification ascendante hiérarchique (CAH) résolvent en partie ces problèmes. Le résultat de l'algorithme n'est en effet pas une partition, mais un dendrogramme. Ce dendrogramme fournit une représentation intuitive de la distance entre tous les échantillons du jeu de données et peut lui-même être partitionné pour retrouver des clusters. L'algorithme de CAH que j'ai utilisé durant cette thèse est celui basé sur le critère de Ward (Jr, 1963). Il n'existe pas de consensus pour ou contre l'utilisation de cette méthode (Duò et al., 2018). Cependant cet algorithme a pour principe de fonctionnement une minimisation de l'inertie intraclasse et une maximisation de l'inertie interclasse, ce qui me semble être une approche cohérente avec un partitionnement optimal. Une autre force de la CAH est de pouvoir construire un dendrogramme consensus entre plusieurs sous-échantillonnages du jeu de données, c'est-à-dire une procédure de *bootstrap*. (Suzuki and Shimodaira, 2006). Une fois le dendrogramme calculé, la partition optimale peut être estimée à l'aide du dendrogramme. Par exemple, la perte d'information peut être évaluée entre la partition à k clusters et la partition à $k + 1$ clusters, il suffira de choisir la partition avec un k raisonnable et qui minimisera la perte d'information. Il existe d'autres procédures de partitionnement de CAH, tel que *dynamicCutree* (Langfelder et al., 2008) qui analyse la forme du dendrogramme pour déterminer les clusters. Cet algorithme est notamment utilisé par une méthode que j'ai employé pendant ma thèse pour le clustering de gènes : WGCNA (Langfelder and Horvath, 2008).

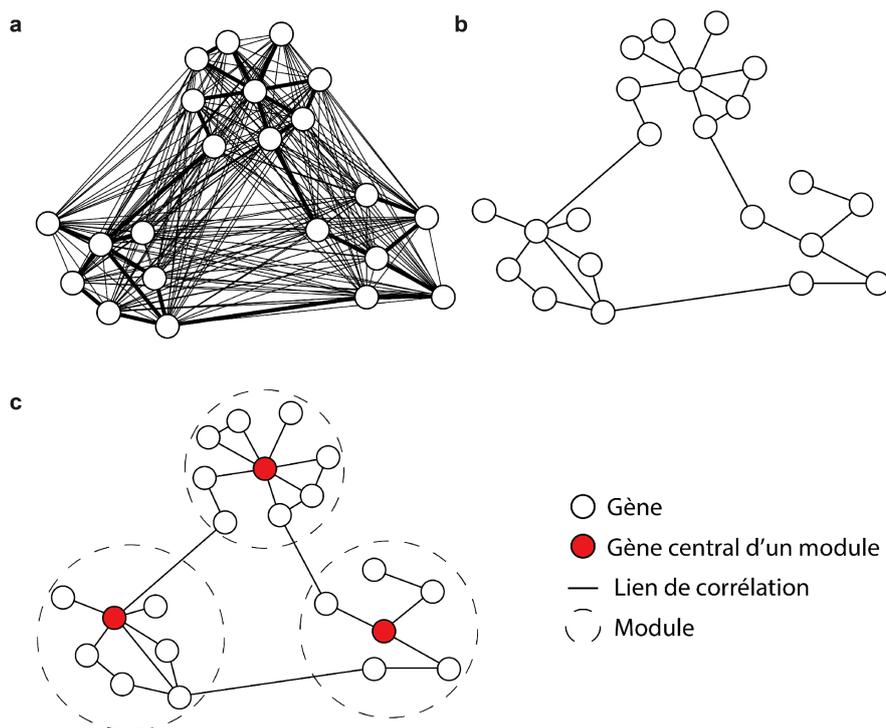


Figure 16 : Schéma de l'algorithme WGCNA. **a.** La matrice de corrélation entre les gènes peut être vue comme une matrice d'adjacence. L'intensité de la corrélation donnera le poids de l'arête entre deux gènes. **b.** Les valeurs de corrélation sous un seuil β sont ignorées, le réseau est grandement simplifié. **c.** Un partitionnement du graphe est effectué à partir d'une matrice de distance basé sur la topological overlap measure (TOM). Les clusters mis en évidence seront les modules de gènes. Diverses statistiques provenant de la théorie des graphes, comme la centralité seront calculées pour chaque gène.

WGCNA (*Weighted Gene Correlation Network Analysis*) est un algorithme de clustering faisant partie des méthodes basées sur les graphes. Une matrice d'adjacence est construite à partir d'une mesure de similarité, comme le coefficient de corrélation de Pearson. Cette matrice va être binarisée selon un seuil β (**Figure 16**). Ce seuil est choisi de façon à ce que le graphe obtenu ressemble le plus possible à un graphe invariant à l'échelle. Les auteurs de l'algorithme considèrent en effet que le graphe de co-expression du transcriptome forme un réseau invariant à l'échelle. C'est-à-dire que la topologie du graphe est la même quelle que soit la partie du graphe examinée, à l'image d'une figure fractale. Une fois le graphe estimé, une matrice de distance entre les nœuds du graphe (les gènes) est calculée à partir d'une métrique de graphe dénommée *Topological Overlap Measure* (TOM). Une CAH va être effectuée à partir de cette mesure pour calculer un dendrogramme, qui sera partitionné par une procédure de *dynamicCutree*. Les méthodes de clustering

basées sur les graphes sont aujourd'hui largement utilisées pour partitionner les cellules dans les données de scRNA-Seq. L'algorithme de Seurat (Satija et al., 2015) est notamment basé sur ce type de clustering. Le but de ces algorithmes est de détecter les communautés au sein du graphe de proximité des cellules, à l'aide d'algorithmes comme celui de Louvain (Blondel et al., 2008) ou de Leiden (Traag et al., 2019).

Il existe encore bien d'autres familles de méthodes de partitionnement, par exemple le clustering basé sur la densité (Ester et al., 1996). Ce type d'algorithme est rapide et permet de détecter des clusters avec des topologies particulières (**Figure 15d**) et même de détecter les échantillons aberrants. Dans cette procédure, on « saute » d'échantillon en échantillon selon une distance maximale nommée voisinage ε prédéfinie par l'utilisateur. Un cluster est défini par un ensemble d'échantillons accessibles entre eux par « saut », peu importe l'échantillon qui est choisi comme point de départ. Il est donc nécessaire de choisir le paramètre de voisinage ε , autrement dit la distance de « saut » avec attention. La contrepartie de cette méthode est que des échantillons très éloignés dans l'espace de départ peuvent être attribués au même cluster.

1.2.2.9 L'enrichissement fonctionnel

Les différentes étapes de l'analyse de données de RNA-Seq amènent la constitution de listes de centaines voire de milliers de gènes d'intérêt en sortie d'analyse. Cette quantité de résultats implique un travail fastidieux d'investigation gène par gène. C'est ici qu'intervient l'enrichissement fonctionnel, qui permet d'attribuer à des groupes de gènes des fonctions biologiques communes. Cette étape permet de générer des hypothèses sur le lien entre le profil d'expression et le phénotype. L'enrichissement fonctionnel se fait au minimum à partir d'une base de données d'annotations fonctionnelles et d'une méthode d'enrichissement. Dans ces bases de données, les gènes sont regroupés par terme. Un terme est un ensemble de gènes qui participent à une même fonction. Il existe deux grands types de base de données d'annotation fonctionnelle : les bases de données où chaque terme correspond à une liste de gènes, c'est le cas de *Gene Ontology*; et les bases de données dont les gènes sont organisés en réseau au sein de chaque terme. Ce dernier cas est

représenté par *KEGG* et *Reactome*. Il existe deux grandes catégories d'algorithmes d'enrichissement, le modèle ORA (*Over-representation Analysis*) et FCS (*Functional Class Scoring*). Dans le Modèle ORA, les gènes sont classés en deux groupes, souvent en différentiellement exprimés et non différentiellement exprimés. Pour chaque terme, on construit une table de contingence avec ces deux groupes et l'appartenance ou non des gènes au terme (**Figure 17a**). La p-valeur d'enrichissement correspondra au test statistique utilisé sur cette table de contingence pour établir un lien entre la variable qualitative « appartient au terme » et la variable qualitative « est différentiellement exprimé ». Un des tests les plus communs dans ce cas est le test exact de Fisher. La méthode ORA est chronologiquement la plus vieille (Boyle et al., 2004).

Dans la méthode FCS (Subramanian et al., 2005), tous les gènes sont classés à partir d'un score. Ce score est bien souvent au choix défini par l'utilisateur lui-même, et doit refléter une façon d'ordonner les gènes par importance. Ce score peut par exemple être la p-valeur, le fold-change, ou bien encore un composite des deux. Chaque gène va recevoir un rang en fonction de ce score (**Figure 17b**). Le but est de savoir si les gènes appartenant à une fonction biologique ont une distribution de leur rang statistiquement différente des autres gènes. Cette méthode est utilisée par GAGE (Luo et al., 2009), et fgSEA (Sergushichev, 2016). Les méthodes FCS sont plus efficaces que les méthodes ORA (Bayerlová et al., 2015), mais nécessitent de pouvoir classer les gènes ce qui peut être impossible dans certains cas.

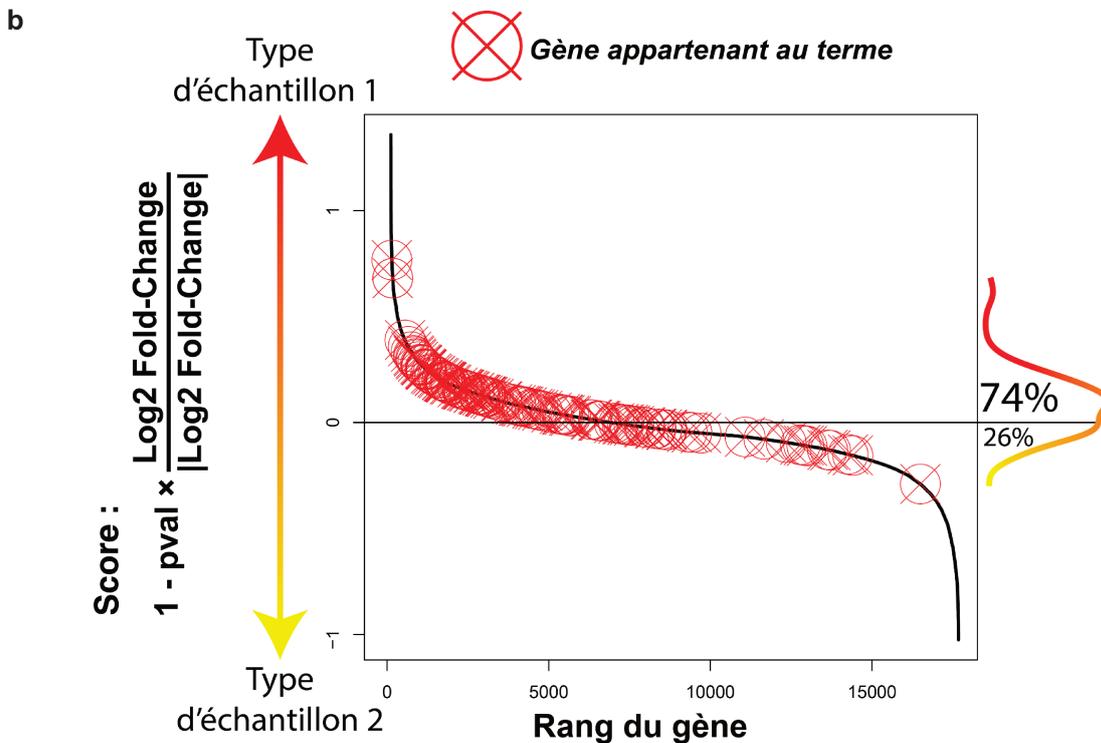
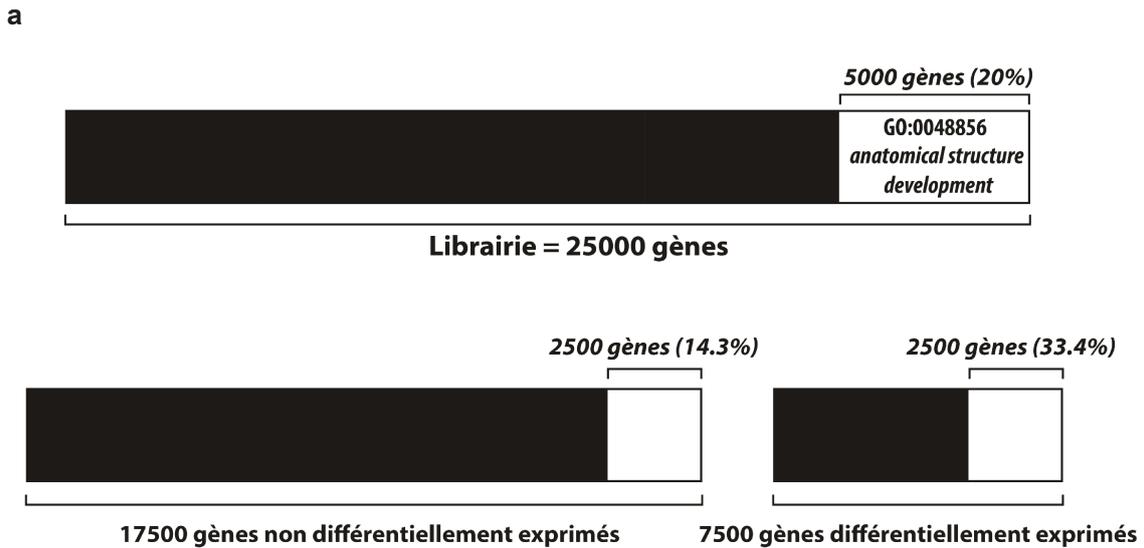


Figure 17 : Algorithmes d'enrichissement fonctionnel. **a.** L'Over Representation Analysis (ORA), se base sur la classification binaire des gènes par l'utilisateur. Par exemple les gènes peuvent être classés en différentiellement ou non différentiellement exprimés en fonction de la significativité du test d'expression différentielle. Ensuite pour chaque terme d'une base de données nous pourrions calculer le taux d'enrichissement du terme, en gènes différentiellement exprimés. **b.** Le Functional Class Scoring (FCS) nécessite un score qui va permettre de classer les gènes, ici ce score est composite entre la p-valeur et le Log₂ Fold-Change. Le but est ensuite de savoir si pour chaque terme, la distribution des rangs des gènes appartenant au terme est aléatoire ou non. Dans le cas présenté il semble y avoir plus de gènes du terme dans les rangs élevés.

1.2.2.10 Etudier les phénomènes de spécification à partir de scRNA-Seq

Le scRNA-Seq a ouvert la voie à la reconstruction des trajectoires de destin cellulaire à l'échelle du transcriptome. La biologie du développement a particulièrement bénéficié de cette avancée dans la compréhension des phénomènes de spécifications cellulaires. Il existe donc un fort besoin d'algorithmes performants de prédiction du devenir des cellules.

1.2.2.10.1 Inférer des trajectoires cellulaires à partir du transcriptome

Reconstruire les trajectoires des cellules présentes dans le jeu de données est une méthode efficace pour étudier les processus dynamiques. On parle d'analyse par pseudo-temps (Cannoodt et al., 2016). Le pseudo-temps est une métrique arbitraire qui mesure l'avancement d'une cellule particulière dans le processus dynamique étudié. Le pseudo-temps peut-être aussi vu comme la quantité totale de changement transcriptomique d'une cellule à partir de la cellule la moins « mature » (ou cellule racine) de la trajectoire cellulaire. Le pseudo-temps au sens strict est donc une variable quantitative qui possède plusieurs propriétés intéressantes. Tout d'abord, si l'on effectue une régression de l'expression en fonction du pseudo-temps, nous obtenons les expressions théoriques aux différentes étapes du processus étudié (Qiu et al., 2017). L'expression au cours du pseudo-temps peut ne pas être monotone, c'est-à-dire qu'un gène peut être activé puis réprimé au cours d'un seul et même processus. La régression linéaire n'est donc pas suffisante pour modéliser l'expression au cours du pseudo-temps, il faut donc utiliser des modèles de régression polynomiale, de degré correspondant à la complexité du processus étudié, généralement deux ou trois. Une autre approche est d'utiliser une régression non paramétrique, par exemple une régression locale, pour modéliser des processus de complexité inconnue, en contrepartie ce type de régression limite la compréhension du phénomène étudié. Les mesures de qualité de la régression, tel que le coefficient de R^2 et la p-valeur associée, vont permettre de mettre en évidence les gènes dont l'expression est liée aux variations du pseudo-temps, et qui ont donc un rapport avec le processus dynamique étudié. Enfin grâce à cette régression nous pouvons examiner l'ordre d'apparition et de disparition des transcrits. Cette information pourrait s'avérer déterminante pour modéliser les réseaux de

régulation des gènes. Cependant les méthodes de pseudo-temps semblent pour l'instant moins précises que les mesures à différents intervalles de temps pour établir des réseaux de régulations causaux (Qiu et al., 2018). Une autre propriété intéressante du pseudo-temps est sa comparaison avec des données temporelles lorsque celles-ci sont disponibles : plus le pseudo-temps avance vite par rapport au temps, plus les changements transcriptomiques sont rapides au cours du processus étudié. Ce dernier cas est observable dans nos données de scRNA-Seq. En effet il existe un « saut » transcriptomique au jour 5 entre les cellules non spécifiées et les cellules de la branche épiblaste. Une fois les cellules dans la branche de l'épiblaste, le pseudo-temps « ralentit » et les cellules subissent peu de changements transcriptomiques aux jours 6 et 7 (Meistermann, Loubersac et al. Figure 2G).

La plupart des méthodes d'inférence de trajectoire travaillent en trois temps : 1/ une réduction de dimension 2/ l'apprentissage des trajectoires sous une forme de composantes, d'arbres ou de graphes. 3/ l'assignement d'une position sur une trajectoire pour chaque cellule, l'avancement de la cellule sur la trajectoire nous donnera le pseudo-temps associé. Dans le cas de Monocle2 (Qiu et al., 2017) la méthode de réduction utilisée (DDRTree) a comme particularité d'estimer intrinsèquement des trajectoires. Il existe une myriade d'autres stratégies pour faire de l'inférence de trajectoire (Cannoodt et al., 2016; Saelens et al., 2019). La stratégie la plus simple est de réduire le jeu de données en une dimension, la position de chaque cellule par rapport à cette dimension est alors considérée comme la valeur de pseudo-temps. Cette approche ne permet pas d'observer des phénomènes de bifurcation et donc ne permet pas l'étude de processus de spécification cellulaire semblables à ceux du développement préimplantatoire. Les algorithmes utilisés par une méthode d'inférence de trajectoire vont donc directement déterminer les possibilités de la méthode en termes de topologie des trajectoires. Par exemple PAGA (Wolf et al., 2019) est une méthode basée sur l'analyse du graphe de proximité des cellules et permet d'inférer des trajectoires indépendantes ou comportant des cycles. Il est donc nécessaire de connaître la nature du processus étudié pour choisir un algorithme de d'inférence de trajectoire de façon optimale (Figure 18).

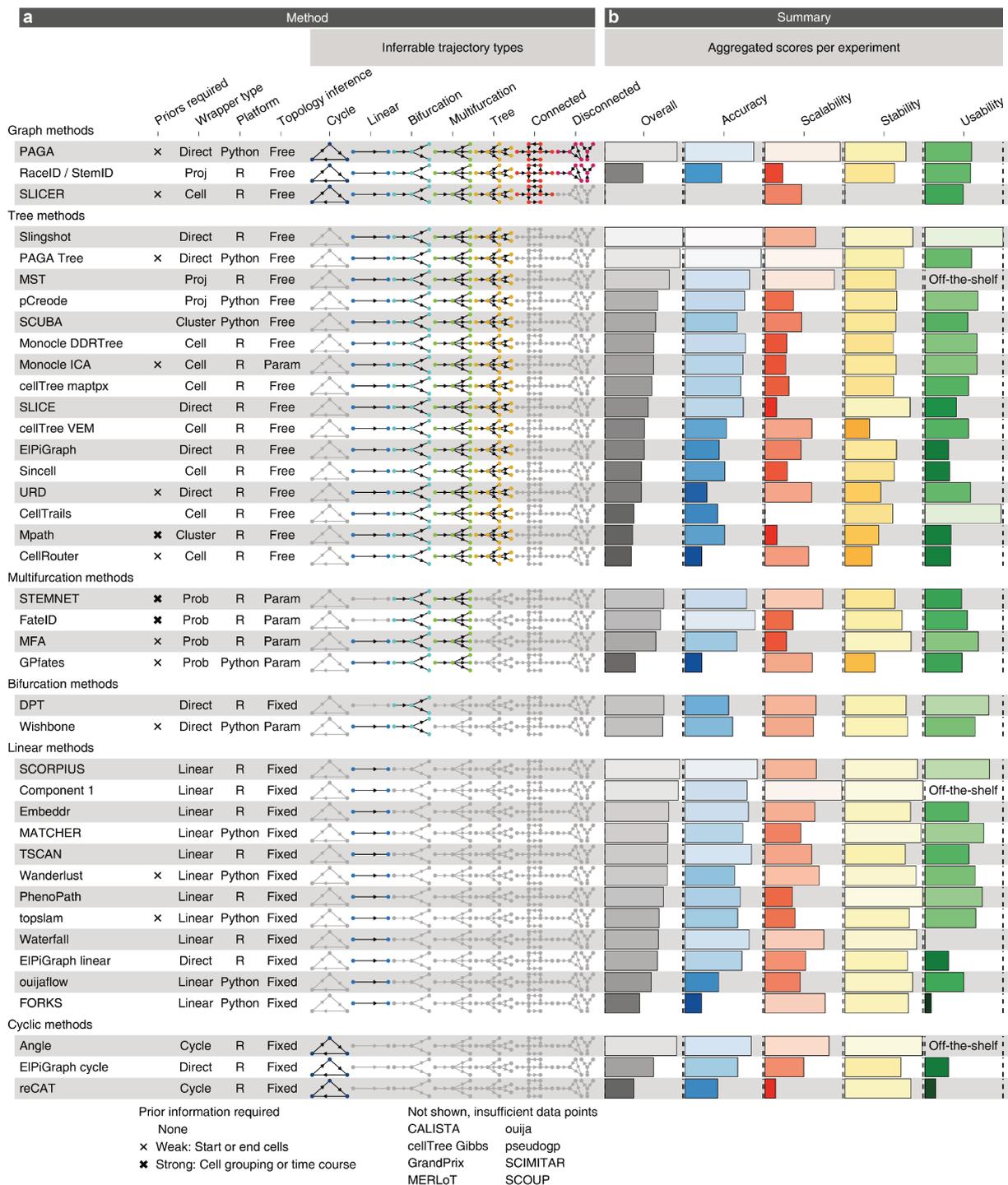


Figure 18: Evaluation de 45 méthodes d'inférence de trajectoire / pseudo-temps. Figure issue de Saelens et al., 2019. **a.** Caractérisation des méthodes selon le type de wrapper, leurs prérequis, si la topologie déduite est contrainte par l'algorithme (fixed) ou un paramètre (param), et les types de topologies déductibles. Les méthodes sont regroupées verticalement en fonction du type de trajectoire le plus complexe qu'elles peuvent déduire. **b.** Les résultats globaux de l'évaluation comportent quatre critères : précision en utilisant une trajectoire de référence sur des données réelles et synthétiques, évolutivité avec un nombre croissant de cellules et de caractéristiques, stabilité à travers les sous-échantillons de l'ensemble des données et la qualité de la mise en œuvre.

1.2.2.10.2 Estimer la vélocité du transcriptome : une nouvelle façon de prédire le destin cellulaire

Le pseudo-temps est une des méthodes utilisées pour prévoir des trajectoires de destin cellulaire, et il existe d'autres stratégies de prédiction du destin cellulaire à partir de l'estimation de l'expression des gènes, telles que l'estimation du « paysage » de Waddington (Wang et al., 2011). Une approche récente a cependant ouvert un nouveau champ dans l'analyse de données de scRNA-Seq : l'estimation de la vélocité de l'expression (La Manno et al., 2018). Cette méthode se base sur la présence de données introniques dans les fichiers de séquences provenant de scRNA-Seq. Ces données introniques permettent d'estimer la proportion d'ARN immatures (non épissés). En prenant en compte la demi-vie des ARNm matures il est alors possible d'estimer la vélocité de l'expression, c'est-à-dire si le nombre de transcrits d'un gène est en cours d'augmentation ou de diminution. Une fois la vélocité d'expression de chaque gène calculé, il est possible d'inférer l'état de la cellule plusieurs heures après le séquençage. Il est aussi possible d'estimer une matrice de probabilité de transition entre chaque cellule. Cette matrice permet de retrouver les populations ancestrales et descendantes probables pour chaque cellule présente dans le jeu de données. Un des plus gros points forts de cette technologie est son caractère rétroactif sur tous les jeux de données de scRNA-Seq. Enfin ce type d'analyse et l'inférence de trajectoire forment un ensemble de méthodes que l'on peut qualifier sous le terme de temporalomique (Lederer and La Manno, 2020).

1.2.2.11 Le défi de la modélisation du transcriptome

Un des buts possibles de l'analyse de transcriptome est la compréhension des liens de corrélation voir de causalité qui existent entre les gènes. Les algorithmes comme WGCNA ne permettent pas d'avoir une vision causale de la cascade de régulation de l'expression. Pour entrer dans ce type de raisonnement il est nécessaire de faire de la modélisation de réseau plus avancée à l'aide d'outils provenant de la biologie des systèmes. Ces outils peuvent être de différentes natures et modéliser les régulations de gènes de façons discrètes (réseaux booléens) ou continu (modèle à équation différentielles). De plus le modèle peut être déterministe ou stochastique, dans ce

dernier cas la mise à jour de l'état du réseau se fait en fonction de la probabilité de chaque interaction. Ces types de modélisation me semblent être très efficaces pour acquérir des connaissances fiables sur le vivant, cependant la construction du réseau peut s'avérer fastidieuse voire impossible à l'échelle du transcriptome entier. D'autres méthodes prennent en considération la particularité des données d'expression en cellule unique pour modéliser le réseau de régulation de gènes, notamment la nature discontinue et stochastique de l'expression (Bonnaffoux et al., 2019).

Les données transcriptomiques seules ne permettent pas d'avoir une vision des autres couches de régulations de l'expression, comme l'épigénome où les caractéristiques de la chromatine. Or celles-ci s'avèrent indispensable à la compréhension de la régulation de l'expression à l'échelle du transcriptome (Trapnell, 2015). De nombreuses méthodes permettant de caractériser ces autres couches de régulations apparaissent, ou se voient adaptées à l'échelle de la cellule unique. Ces nouvelles données représentent autant de nouvelles variables à combiner avec les données transcriptomiques, ce qui représente un véritable potentiel dans la compréhension du vivant, mais aussi un immense défi en termes d'analyse.

2 Manuscrit #1 : Parallel derivation of isogenic human primed and naive induced pluripotent stem cells

2.1 Contexte

2.1.1 Pluripotence naïve et amorcée

En 1998 les premières cellules souches embryonnaires humaines (hESC) furent obtenues en dérivant de la masse cellulaire interne préimplantatoire (Thomson et al., 1998). Les hESC sont pluripotentes et possèdent une capacité d'auto-renouvellement. Cet auto-renouvellement permet de cultiver les hESC indéfiniment, les hESCs ont donc ouvert de nouvelles pistes de recherche en thérapie cellulaire et en recherche sur le développement humain.

L'étude des hESCs comparées aux ESC murines (mESC) a révélé les spécificités des lignées de cellules souches pluripotentes (PSC) en fonction des espèces : ces différences incluent le milieu de culture permettant de les maintenir, la morphologie, ainsi que leur profil transcriptomique (Nichols and Smith, 2009). Leur situation épigénétique est aussi différente, la plus importante illustration est la situation du chromosome X : dans les ESC murines, les deux X des cellules femelles sont actifs, contre un X inactivé chez l'Homme. Enfin elles n'ont pas le même aspect fonctionnel, ainsi les ESC de primates ne peuvent pas être intégrées dans des blastocystes préimplantatoires pour réaliser des chimères, ce qui est possible chez la souris, et utilisé pour créer des modèles murins génétiquement modifiés. L'observation de ces différences entre ces espèces a conduit à deux hypothèses : 1/ les cellules pluripotentes sont fondamentalement différentes entre l'Homme et la souris ; 2/ Il existe deux états de la pluripotence, et il est plus facile de capturer un des états selon l'espèce.

En 2007 une partie de l'énigme fut résolu, lorsque des cellules murines semblables aux hESCs furent obtenues (Brons et al., 2007; Tesar et al., 2007). Ces cellules sont dénommées EpiSC et obtenues à partir d'épiblaste postimplantatoire par dérivation dans un milieu de culture proche de celui utilisé pour les hESC. Cette différence

fondamentale a abouti à la distinguer de deux types de pluripotence: la pluripotence naïve, préimplantatoire, représentée par les mESC ; et la pluripotence amorcée, postimplantatoire et représenté par les EpiSC et les hESC (Nichols and Smith, 2009). La question qui se pose alors porte sur l'existence et la capture de l'état naïf chez l'Homme. Ces questions furent résolues en 2014, (Takashima et al., 2014; Theunissen et al., 2014) lorsque des équipes parvinrent à convertir des hESC amorcées en hESC naïves, par l'utilisation de milieux de cultures spécifiques. La création de ces milieux de culture, notamment le t2iLGö a permis en 2016 la dérivation d'hESC naïves à partir de masse cellulaire interne humaine (Guo et al., 2016).

En parallèle de l'étude de la pluripotence, une autre découverte a eu un impact fondamental sur notre compréhension du destin cellulaire et en particulier la pluripotence. En 2006, l'équipe de Shinya Yamanaka a reprogrammé des cellules somatiques en cellules souches pluripotentes induites, d'abord chez la souris (Takahashi and Yamanaka, 2006) puis chez l'Homme (Takahashi et al., 2007). Cette reprogrammation de cellules différenciées en cellules pluripotentes est possible par l'expression forcée de quatre facteurs de transcription clefs : les facteurs OSKM (OCT4, SOX2, KLF4, c-MYC). Ceux-ci mettent en place une boucle de rétroaction positive entraînant leur expression de façon endogène et à terme un changement de l'identité de la cellule vers la pluripotence amorcée chez l'Homme. L'intérêt des iPSC réside dans la facilité à générer des nouvelles lignées avec différents fonds génétiques. Cette capacité permet par exemple d'utiliser les iPSC comme modèle d'étude de l'influence du fond génétique sur la pluripotence (Schnabel et al., 2012). Les iPSC permettent aussi de garantir un fond génétique identique au patient dans le cas d'une utilisation en clinique, ce qui limite les phénomènes de rejet en cas de thérapie à base de ces cellules.

2.1.2 Motivation de l'étude

Notre étude s'inscrit à la croisée de ces dernières grandes avancées dans le domaine de la pluripotence, et vise à obtenir des cellules souches pluripotentes naïves induites chez l'Homme (hNIPSC). Il a été montré que lors de la reprogrammation que des caractéristiques naïves apparaissent de façon transitoire (Cacchiarelli et al.,

2015) et que l'on pouvait donc espérer capturer l'état naïf en utilisant simplement la transfection par OSKM et en utilisant un milieu de culture adéquat. La reprogrammation directe pouvait permettre de comparer les cellules provenant du même fond génétique, dans différent milieu, permettant de réévaluer le niveau de pluripotence obtenu en fonction des conditions de culture. Nous nous sommes servis du milieu utilisé lors de l'obtention des hNESC (t2iLGö) et d'un milieu commercial (RSet).

Le premier objectif de ce travail était de confirmer que ces lignées soient dans un état de pluripotence naïve. Un autre objectif de ce travail était de mieux caractériser la pluripotence naïve chez l'Homme. Dans cette optique nous avons aussi utilisé notre protocole de reprogrammation avec des milieux habituels aboutissant à une pluripotence amorcée (mTeSR1, KSR + FGF2). Nous avons de même inclus des hESC et de l'ARN des hNESC (Guo et al., 2016). Toutes ces lignées ont été séquencées par DGE-Seq en « bulk », c'est-à-dire que l'ARN provient du mélange de l'ensemble des cellules présentes dans le puits de culture au moment de la lyse (cf. 1.2.1.2, en page 25). Un des objectifs de ma thèse a été de transformer cette information transcriptomique en hypothèses biologiques telles que :

- Quels sont les facteurs de transcription spécifique d'un état particulier de pluripotence ?
- Comment trouver de manière fiable les fonctions biologiques modifiées d'un état à l'autre ?
- Quels sont les nouveaux marqueurs des différents états de pluripotence ?

Un autre défi concerne la comparaison des données transcriptomiques de natures différentes : des lignées *in vitro* (données de bulk RNA-Seq, chaque transcrit séquencé est compté une fois) aux données provenant de l'embryon humain (données de single-cell RNA-Seq, transcrits morcelés en plusieurs reads). La comparaison des cellules à l'embryon préimplantatoire est essentielle pour évaluer les caractéristiques de chaque lignées pluripotente et son milieu de culture avec la situation lors du développement préimplantatoire.

2.2 Manuscrit

ARTICLE

DOI: 10.1038/s41467-017-02107-w

OPEN

Parallel derivation of isogenic human primed and naive induced pluripotent stem cells

Stéphanie Kilens et al.[#]

Induced pluripotent stem cells (iPSCs) have considerably impacted human developmental biology and regenerative medicine, notably because they circumvent the use of cells of embryonic origin and offer the potential to generate patient-specific pluripotent stem cells. However, conventional reprogramming protocols produce developmentally advanced, or primed, human iPSCs (hiPSCs), restricting their use to post-implantation human development modeling. Hence, there is a need for hiPSCs resembling preimplantation naive epiblast. Here, we develop a method to generate naive hiPSCs directly from somatic cells, using OKMS overexpression and specific culture conditions, further enabling parallel generation of their isogenic primed counterparts. We benchmark naive hiPSCs against human preimplantation epiblast and reveal remarkable concordance in their transcriptome, dependency on mitochondrial respiration and X-chromosome status. Collectively, our results are essential for the understanding of pluripotency regulation throughout preimplantation development and generate new opportunities for disease modeling and regenerative medicine.

Correspondence and requests for materials should be addressed to L.D. (email: Laurent.david@univ-nantes.fr).

[#]A full list of authors and their affiliations appears at the end of the paper

Pluripotent stem cells (PSCs) possess the unique ability to self-renew and differentiate into all cell types of a fully functional adult, making them invaluable tools to study human development, model diseases and design new regenerative medicine approaches. In mammals, pluripotency exists in at least two states: naive pluripotency that represents the ground state of pluripotency found in the preimplantation epiblast and primed pluripotency that corresponds to cells poised for differentiation found in the post-implantation epiblast^{1,2}. To date, the majority of human embryonic stem cell (hESC) lines have been derived and maintained in the primed state, and identifying culture conditions supporting human naive pluripotency has been a major goal for the past decade. Since 2013, several studies have yielded multiple, distinct conditions to induce and maintain naive pluripotency^{3–9}. In parallel, significant progresses have been made to characterize the molecular signature of human pre-implantation epiblast cells^{10–15}, establishing guidelines to assess human naive pluripotency¹⁶. Collectively, those studies showed that two media supported naive pluripotent stem cells converted from primed cells or derived directly from human embryos, demonstrating hallmarks of human epiblast cells: 5i/L/AF^{8,17,18} and T2iLGö^{7,15,19,20}. However, it remains unknown whether naive pluripotency can be induced from somatic cells directly without a primed intermediate, and if so, with sole expression of OKMS (Oct4, Klf4, cMyc and Sox2), like in mouse^{21–23}.

Here we present a protocol enabling the parallel derivation of isogenic human induced primed (hiPSCs) and naive (hiNPSCs) pluripotent stem cells. hiNPSCs are reprogrammed using T2iLGö^{7,19} or RSeT. hiNPSCs are benchmarked against the human preimplantation epiblast, the gold standard of human naive pluripotency, at the transcriptomic, metabolic and epigenetic levels. Overall, hiNPSCs derived in T2iLGö medium display remarkable similarities to preimplantation epiblast. Thus, direct somatic cell reprogramming to human naive pluripotency complements the array of assays enabling in-depth analysis of human pluripotency.

Results

Reprogramming somatic cells into naive hiPSCs. We aimed to develop a direct reprogramming method to simultaneously generate isogenic naive and primed human PSCs. We overexpressed *OCT4*, *KLF4*, *MYC* and *SOX2* in human fibroblasts from 5 healthy donors, using a non-integrative Sendai virus. At day 7, cells were split to 3 tissue culture dishes, enabling to induce multiple pluripotent states directly from the same parental cells. At day 9, we cultured emerging colonies in primed pluripotency medium (KSR+FGF2) and in media supporting human naive pluripotency (RSeT and T2iLGö) (Fig. 1a). Both media contain 2i, inhibitors of MEK and GSK3 β which are essential for mouse PSCs maintenance²⁴, and LIF. Besides 2i and LIF, T2iLGö medium contains a PKC inhibitor^{7,19,25}, while the RSeT is a medium derived from the NHSM⁵, composed of inhibitors of JNK and p38, FGF2 and TGF β 1, which supports interspecies chimeras. RSeT medium was chosen due to accessibility and apparent low genomic abnormality rate, and T2iLGö because it was reported to yield cells with more stable genome over 5i/L/AF^{7,8,17}. In order to broaden our analysis, we switched some KSR+FGF2 hiPSC lines to mTeSR1 feeder-free medium. In total, we generated 25 cell lines (Fig. 1b and Supplementary Table 1), of which cells grown in RSeT or T2iLGö formed dome-shaped colonies resembling mouse embryonic stem cells (mESCs). We controlled Sendai expression and confirmed transgene independency of hiNPSCs, but at higher passages than in hiPSCs (Supplementary Fig. 1 and Supplementary Table 1). hiPSCs and hiNPSCs display karyotype identical to the parental fibroblasts; however, hiNPSCs tend to acquire chromosomal abnormalities,

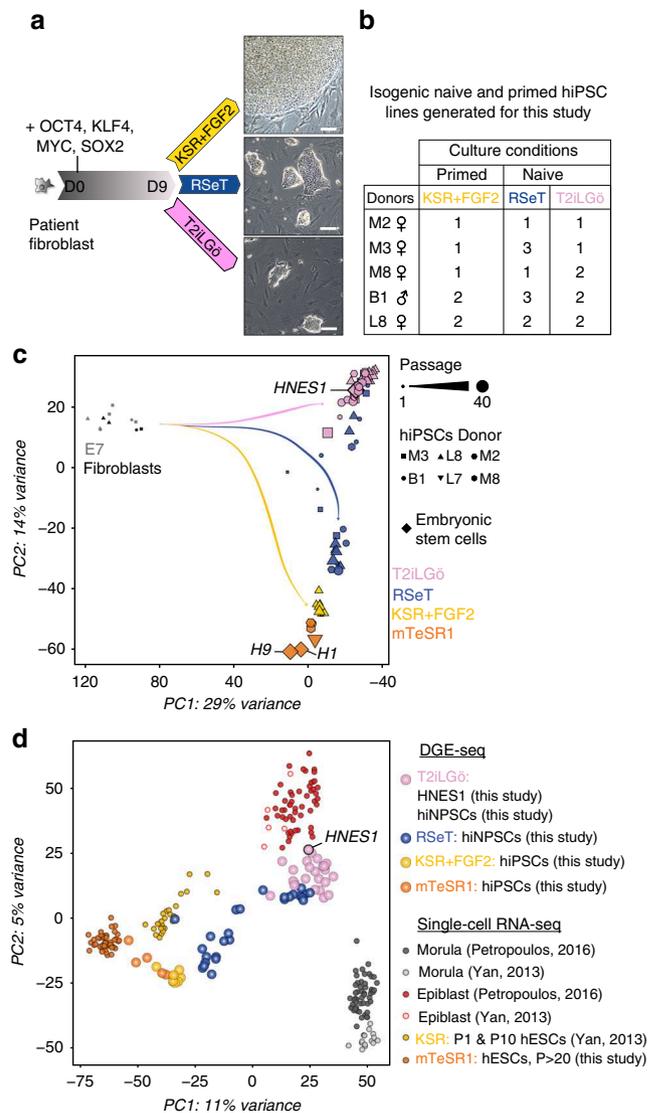


Fig. 1 Direct reprogramming of somatic cells into hiNPSCs. **a** Direct generation of isogenic naive and primed hiPSCs. Fibroblasts were transduced with 3 Sendai viruses expressing a polycistron KLF4/OCT4/SOX2, MYC and KLF4 at a ratio of 5:5:3, respectively. Cells were split on feeders at day 7, and placed in the indicated media at day 9. Scale bar = 100 μ m. **b** Summary of lines generated for this study in primed (KSR+FGF2, yellow) or naive culture media (RSeT, blue or T2iLGö, pink) originated from 5 different donors. **c** Different pluripotent states are induced depending on culture media. Transcriptomes of hiPSCs and hiNPSCs, control primed hESC lines H1 and H9 or the naive hESC line HNES1¹⁹ were analyzed by PCA. Symbols represent donor lines, and size of the symbols represents the passage. Arrows have been drawn to highlight the reprogramming trajectories. **d** T2iLGö hiNPSCs are the closest to human epiblast cells. PCA of single-cell RNA-seq data sets from preimplantation embryo samples^{11,12} compared to primed hPSCs from ref.¹¹ and to primed/naive hiPSCs/hESCs from this study

as previously reported for human naive embryonic stem cells (hNESCs)^{8,17,19} (Supplementary Table 1). These genomic alterations have recently been associated with the inhibition of MEK through PD0325901, one major component of most media supporting human naive pluripotency²⁶. We limited the diploid/tetraploid ratio by reprogramming and growing cells under hypoxic conditions and constant rock inhibition (Y27632) (Supplementary Table 1), and by subcloning T2iLGö hiNPSCs.

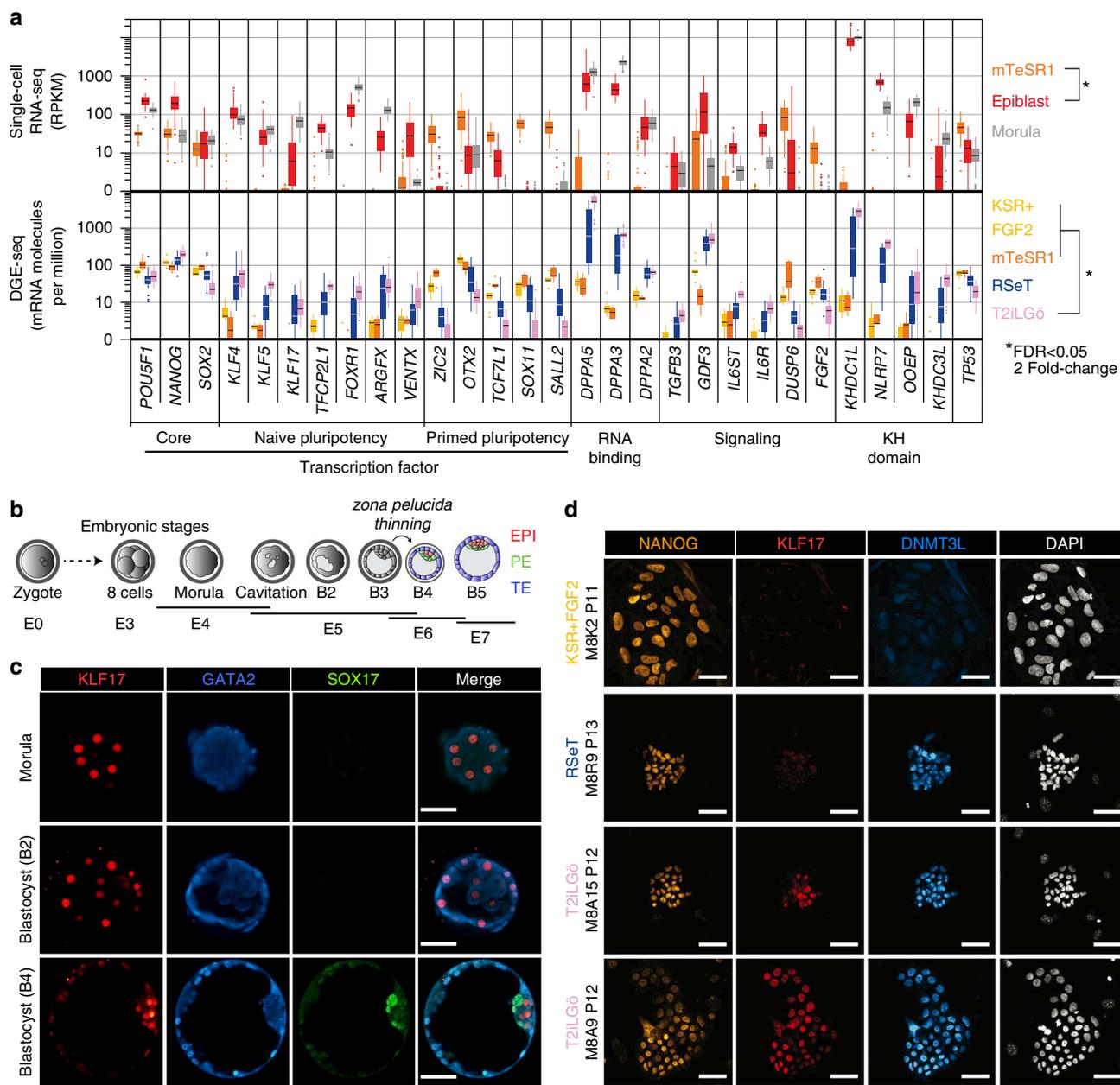


Fig. 2 hiNPCs express markers specific to human epiblast cells. **a** Specific naive pluripotency markers display identical profiles in T2iLG6 hiNPCs and preimplantation epiblast cells. Individual differentially expressed genes plotted as RPKM for single-cell RNA-seq or mRNA molecules per million of total mRNA molecules for DGE-seq. Upper panel: all genes are differentially expressed (Epi vs primed), except *SOX2*; lower panel: all genes are differentially expressed (T2iLG6 vs primed), except *POU5F1*(*OCT4*) and *NANOG*. Error bars are defined as s.e.m. Statistical tests used to compute differentially expressed genes are defined in the “Differential Expression profiling” section of the Methods. **b** Schematic representation of the human preimplantation development comparing clinical staging (Morula, B2, B3, B4 and B5) with corresponding embryonic days (E). EPI epiblast cells in red, PE primitive endoderm cells in green; TE trophoblast cells in blue. **c** KLF17 protein is expressed in all morula cells before being restricted to epiblast cells in the blastocyst. Human embryos were cultivated in a time-lapse microscope and fixed at indicated stages (morula, B2 or B4 blastocysts). Immunofluorescence for KLF17 (red), GATA2 (blue) and SOX17 (green) was performed. For each indicated embryonic stage, immunofluorescence was performed on 3 biological replicates. Scale bar = 50 μ m. **d** KLF17 protein is specifically expressed in T2iLG6 hiNPCs (pink) and not in isogenic lines cultivated in RSeT (blue) and KSR-FGF2 (yellow). Indicated cell lines were analyzed by immunofluorescence for NANOG (yellow), KLF17 (red) and DNMT3L (cyan). This figure is representative of 8 biological replicates. Scale bar = 50 μ m

We analyzed our hiNPC and hiPSC lines, at different passages, by 3' digital gene expression RNA-sequencing (DGE-seq), a quantitative method based on molecular indexing of messenger RNA (mRNA) molecules^{27,28}. Controls used in this analysis are primed (H1 and H9) and naive (HNES1) hESC lines¹⁹. Principal component analysis (PCA) revealed two major components: the first discriminating between parental cells and

pluripotent stem cells, the second discriminating between H1 and H9 on one side and HNES1 on the other side (Fig. 1c). T2iLG6 and RSeT hiNPC lines were separated along the second principal component. The majority of T2iLG6 hiNPCs clustered with HNES1, regardless of passage number, while cells exposed to RSeT formed two intermediate clusters and tended to become more similar to primed human PSCs (hPSCs) at later passages

(trajectory drawn in Fig. 1c). To assess the development stage associated with hiNPSCs in different culture media, we compared them by PCA to the single-cell RNA-seq data sets of hESCs upon derivation (P1 and P10), high-passage hESCs, human epiblast cells and human morula cells^{11,12} (Fig. 1d). The first component classified samples from the morula cells, the cellular fate preceding pluripotency, to high-passage primed hESCs. Strikingly, T2iLGö hiNPSCs cluster together with epiblast cells while RSeT hiNPSCs sit between primed and naive PSCs. Altogether, these data reveal that our protocol can reprogram somatic cells directly to a state resembling the human epiblast, without an intermediate passage in primed media.

hiNPSCs express markers specific to human epiblast cells. To further characterize hiNPSCs, we compared in depth their transcriptomes (obtained by DGE-seq) to that of human preimplantation epiblast (generated by single-cell RNA-seq)^{11,12}. To reduce influence of sequencing protocols, we performed a two-step analysis. We first compared the human epiblast signatures¹² with those we obtained by performing single-cell RNA-seq from 52 H1 and H9 primed hESCs (Supplementary Fig. 2a). This yielded 6628 differentially expressed (DE) genes when a cutoff of twofold and false discovery rate (FDR) of <0.05 was applied. Second, we compared transcriptomes obtained by DGE-seq of primed hESCs and hiPSCs in KSR+FGF2 or mTeSR1 to T2iLGö hiNPSCs, as they clustered with HNES1 (Supplementary Fig. 2a). Using the same cutoff, we found 3003 DE genes between hiPSC and T2iLGö hiNPSC populations, among which 1980 are common DE genes between hESCs and epiblast cells. Among the top DE gene candidates overexpressed in epiblast and hiNPSCs, we found genes related to RNA binding, such as the *DPPA* family or the KH-domain proteins *KHDC1L*, *NLRP7*, *OOEP* and *KHDC3L* (Fig. 2a). Quantitative DGE-seq showed that those genes represent 1% of the transcriptome in T2iLGö hiNPSCs and HNES1, suggesting that mRNA processing regulated by those genes might play a prime role in naive pluripotency. Our analysis further identified transcription factors whose expression was specifically elevated in naive cells, such as *KLF4* and *KLF5*, and a cohort of specific naive pluripotency regulators: *KLF17*, *FOXR1*, *VENTX* and *ARGFX*; other transcription factors, such as *OTX2* and *SOX11*, were in contrast elevated in hiPSCs (Fig. 2a). Specific signaling pathways were also linked to pluripotency status. The transforming growth factor- β (TGFB) pathway ligands *TGFB3* and *GDF3* as well as the interleukin-6 (IL6) receptors *IL6R* and *IL6ST* were overexpressed in naive PSCs, while *FGF2* was distinctly overexpressed in primed PSCs.

Among the genes differentially expressed between naive and primed PSCs, we further investigated the extremely high expression of *DPPA5* in T2iLGö hiNPSCs (which represent around 0.2% of total mRNAs). We confirmed the expression of *DPPA5* at the protein level by western blot, only in T2iLGö hiNPSCs but not in the RSeT hiNPSCs or hiPSCs (Supplementary Fig. 3a). To our knowledge, it is the first time that *DPPA5* protein has been reported as a marker of human naive PSCs. We also investigated the expression of *KLF17* at the protein level, in hiPSCs, hiNPSCs and human embryos. During human preimplantation development, pluripotent cells emerge after the morula stage and are restricted to epiblast cells at the blastocyst stage²⁹ (Fig. 2b). Immunofluorescence (IF) analysis of human preimplantation embryos revealed that *KLF17* is strongly expressed in all cells of the morula (E4/E4.5). *KLF17* is also present at the B2 stages (E4.5/E5), and rapidly becomes restricted to the epiblast at the B4 blastocyst stage (E5.5/E6) (Fig. 2c). *KLF17* expression is distinct from that of *GATA2*, a marker of human trophoderm, expressed after B2 stage, and *SOX17*, a marker of primitive

endoderm, only expressed at the B4 stage. This supports an important role of *KLF17* during establishment of pluripotency in vivo, and is in line with single-cell RNA-seq analysis¹² (Supplementary Fig. 3b). IF analysis of *KLF17* and *NANOG* in primed and naive hiPSCs showed that while all PSCs expressed *NANOG*, only the T2iLGö hiNPSCs expressed *KLF17* (Fig. 2d and Supplementary Fig. 3c). This was further confirmed by flow imaging, highlighting a striking difference in signal intensity and in the number of *KLF17*-positive cells with nuclear localization between hiPSCs, RSeT and T2iLGö hiNPSCs (Supplementary Fig. 4). In particular, the intensity median of nuclear *KLF17* is 2.2 for the T2iLGö hiNPSCs, 1.08 for the RSeT hiNPSCs and 0.79 for the hiPSCs, confirming that the *KLF17* profile for RSeT hiNPSCs is closer to hiPSCs (Supplementary Fig. 4B). In contrast, *DNMT3L* is upregulated in both T2iLGö and RSeT hiNPSCs at the protein level, revealing the intermediate pluripotent state of RSeT hiNPSCs.

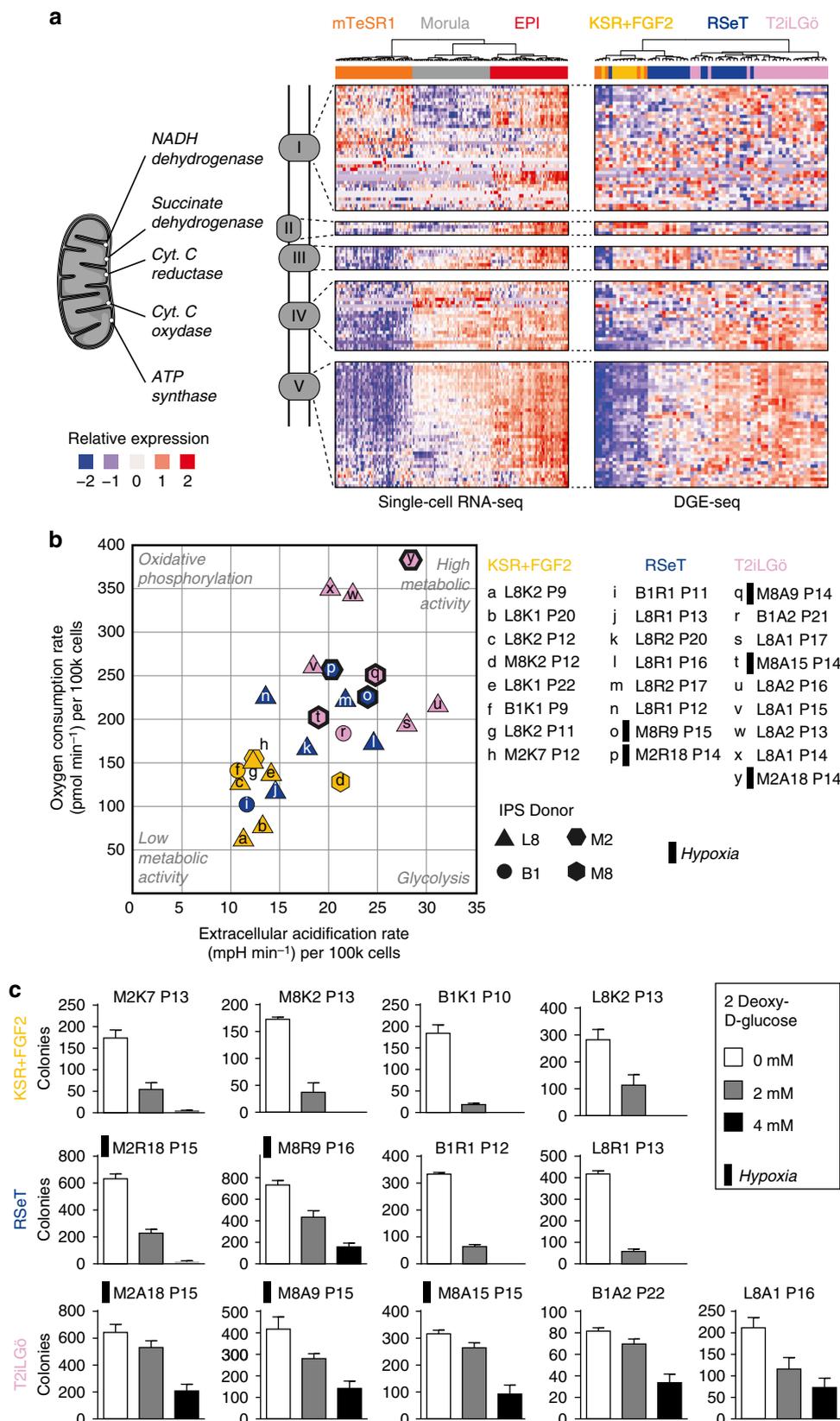
In addition to specific individual markers, we identified a strong correlation between naive pluripotency and pathways related to metabolism (Val-Iso-Leu degradation, purine and pyrimidine metabolism), cell cycle (p53 pathway, cell cycle, apoptosis) and cell junctions (adherent and tight junctions, focal adhesion) (Supplementary Fig. 2b). As those pathways are enriched in both epiblast cells and T2iLGö hiNPSCs, it suggests potential cross-talks between metabolic pathways and human naive pluripotency. Nonetheless, characterization of specific active pathways would be complementary to individual markers or to recently proposed transcriptome profile³⁰ to assess the naive nature of human PSCs. We tested the discriminative power of metabolism, cell cycle or cell junction pathways signatures to classify hiNPSCs, and observed that the result was depending on the pathway (Supplementary Fig. 2b). To improve the predictive power of our pathway-based approach, we performed PCA for each pathway and combined the first components to create a three-dimension space, in which we projected our samples. We observed that all the samples were aligned along a common axis, delimited by primed hESCs on one end and naive preimplantation epiblast cells on the other end. Suitably, each sample is clearly classified along this axis, including the RSeT hiNPSCs which have only partially gained expression of components of those pathways and are sitting between primed hiPSCs (in KSR+FGF2 or mTeSR1) and T2iLGö hiNPSCs (Supplementary Fig. 2c).

Altogether, our thorough analysis highlighted specific markers and pathways that characterize naive pluripotency. Moreover, we uncovered a hierarchy of markers distinguishing between T2iLGö hiNPSCs and intermediate RSeT PSCs.

Metabolic activity of hiNPSCs. To further characterize naive pluripotency, we analyzed enriched signaling pathways in naive pluripotent cells compared to primed pluripotent cells, based on Gene Ontology (GO) terms and the KEGG (Kyoto Encyclopedia of Genes and Genomes) database³¹, with FDR < 0.01. The top enriched GO terms in hiNPSCs are related to the mitochondrial electron transport chain (Supplementary Data 2) and transcriptomic analysis of genes involved in oxidative phosphorylation shows an overall upregulation in human epiblast cells and hiNPSCs (Fig. 3a). To functionally validate the importance of enriched pathways, we tested the mitochondrial activity by measuring the oxygen consumption rate and the extracellular acidification rates in hiPSCs and hiNPSCs. This showed an increased metabolic activity in naive compared to primed cells, with a combined increase of glycolysis, recently reported in hNESCs³², and oxidative phosphorylation (Fig. 3b). Interestingly, the metabolic activity was proportional to the level of naive

pluripotency predicted by transcriptomic and pathway analysis: the RSeT hiNPSCs, with a mildly increased expression of electron transport chain genes, had a modest increase in metabolic activity. Moreover, analysis of oxidative phosphorylation capacity showed that hiNPSCs derived in T2iLGö have a higher

respiratory capacity than RSeT hiNPSCs or hiPSCs, in line with the analysis of HNES1 cells¹⁹ (Supplementary Fig. 5). Clonal assays in culture conditions supplemented with 4 mM 2-deoxy-D-glucose, a competitive inhibitor of glycolysis, confirmed the ability of T2iLGö hiNPSCs to mobilize oxidative phosphorylation



to proliferate. In contrast, RSeT hiNPSCs and primed hiPSCs did not grow under these culture conditions (Fig. 3c). Therefore, metabolic activity is an important discriminant of hiNPSCs, and could be used to select for naive PSCs.

Hypomethylation and X-chromosome reactivation in hiNPSCs.

Naive pluripotency is characterized by DNA hypomethylation in human naive PSCs^{7,17} and preimplantation epiblast^{13,14}. Quantitation of 5-methyl-cytosine (5mC) by mass spectrometry showed that naive cells in T2iLGö had the lowest mC content of all tested lines, under 3% in average, while primed cells were all above 5%, in accordance with the previously published analysis of hNESc lines^{17,19} (Fig. 4a). Comparison of DNA methylation regulators expression between T2iLGö hiNPSCs and hiPSCs show that *DNMT3L* is dramatically increased (up to 0.1% of the transcriptome) in the former, while *DNMT3B* is decreased. The *TET* family has also been recently associated with human naive pluripotency, as *TET1* overexpression could transiently induce expression of naive markers³³. Quantitative DGE-seq shows a gain of *TET2* expression in T2iLGö hiNPSCs, whereas *TET1* expression level remains stable and *TET3* is poorly expressed in all samples (Fig. 4b). On another hand, RSeT hiNPSCs had intermediate levels of *DNMT3L*, *DNMT3B* and *TET2* and 5mC percentage between primed and T2iLGö hiNPSCs. Our results suggest mass spectrometry quantitation of 5mC as a convenient and accurate measurement to qualify hiNPSCs.

The presence of two active X chromosomes is considered a hallmark of human naive pluripotency^{12,18,34}. To assess the activity of the X chromosomes in primed and naive hiPSCs, we first analyzed by IF the distribution of H3K27me₃, a marker of the inactive X (Xi) chromosome. Xi-characteristic H3K27me₃ accumulation is seen in RSeT hiNPSCs and early-passage hiPSCs nuclei consistent with the presence of an Xi (Fig. 5a). In contrast, a low percentage of T2iLGö hiNPSCs and late-passage hiPSCs display H3K27me₃ foci, which could correspond either to the erosion of dosage compensation³⁵ that occurs spontaneously in primed hiPSCs³⁶ or to the reactivation of the Xi, which has been observed in naive PSCs^{15,18,30}. To discriminate between the two hypotheses, we monitored by RNA fluorescent in situ hybridization (RNA-FISH) the activity status of the X chromosomes. We first focused on the expression of the protein-coding gene *ATRX*, which has been shown to resist erosion in most studied lines^{35,37}. Biallelic expression of *ATRX* is consistently observed in T2iLGö hiNPSC lines, with 36% to 100% naive T2iLGö hiNPSCs displaying only active X (Xa) (Fig. 5b). In contrast, biallelic expression of *ATRX* was rarely observed in RSeT, and never in KSR+FGF2 culture conditions. We next probed the expression of the lncRNAs *XIST* and *XACT*, as their relative patterns of expression clearly distinguish the various X-chromosome states. While early-passage hiPSCs have a characteristic post-inactivation staining, with the Xi coated by *XIST* and the Xa coated by *XACT*, we observed loss of *XIST* and biallelic accumulation of *XACT* in late-passage hiPSCs, further confirming

the erosion of X inactivation in these cells (Fig. 5c). In striking contrast, we observed co-accumulation of *XIST* and *XACT* on one active X chromosome in a significant proportion of T2iLGö and RSeT hiNPSCs, but we only observed co-accumulation of *XIST* and *XACT* on both X chromosomes in T2iLGö hiNPSCs, similar to previous findings¹⁵. Collectively, those results show that X-chromosome reactivation has occurred only in T2iLGö hiNPSCs. To ensure that X-chromosome reactivation was happening independently of chromosomal abnormalities, we subcloned the hiNPSCs M2A18, M8A9 and M8A15. Analysis of two subclones of each line by *ATRX*, *XACT* and *XIST* RNA-FISH demonstrated that X-chromosome reactivation occurs in diploid cells (Fig. 6).

Our FISH results stress out the importance of combining the analysis of *XACT*, *XIST* and *ATRX* expression, and not to rely only on *XIST* or H3K27me₃, in order to assess X-chromosome reactivation, a critical hallmark of naive pluripotency in humans¹⁵ (Fig. 5).

Discussion

We generated hiNPSCs directly from somatic cells by using OKMS overexpression and defined culture media. Our method enables parallel generation of naive and primed hiPSCs of the same genetic background, limiting tissue culture time and extended passaging compared to previously published strategies that require primed PSCs prior to their conversion into naive PSCs⁴⁻⁹. Collectively, our results show that a human preimplantation-like state can be induced in somatic cells by directly shifting reprogramming cultures to naive conditions without the need for a primed intermediate. The resulting hiNPSCs in T2iLGö display all hallmarks of human naive pluripotency, while RSeT hiNPSCs display an intermediate naive phenotype. However, we recorded some genomic alterations in T2iLGö hiNPSCs and others showed aberrant imprinting^{17,25} of the cells. Thus, culture conditions are not yet ideal to maintain *in vitro*, the transient human naive pluripotent state.

Further analysis of the array of hiNPSCs that we generated could uncover hierarchy between molecular events necessary to achieve naive pluripotency (Fig. 7). Our data show that transcriptomics analysis is able to rank cells from morula, epiblast/T2iLGö hiNPSCs, RSeT hiNPSCs and primed hESCs/hiPSCs. This supports the concept of a “formative pluripotent state”, a state achieved during the transition from naive demethylated cells to cells primed for differentiation³⁸. Our protocol uses OKMS overexpression to achieve a state compatible with naive and primed pluripotency, in line with the first observation of a higher state of human pluripotency³⁹. One could envision to use factor-based reprogramming to capture the formative state using proper culture medium.

Besides representing a powerful model to study the differences between naive and primed pluripotency, the multiple metastable states of pluripotency could have significant biological properties. Indeed, one of the approach envisioned for regenerative medicine is to generate interspecies chimeras with human pluripotent stem

Fig. 3 T2iLGö hiNPSCs metabolic profile is closely related to preimplantation epiblast. **a** Genes coding proteins of the electron transport chain, located in the inner membrane of the mitochondria, are upregulated in human epiblast cells and T2iLGö hiNPSCs in comparison to their primed counterparts. Relative expression of genes related to oxidative phosphorylation pathway for hESC, morula and epiblast samples analyzed by single-cell RNA-seq (left), and analyzed by DGE-seq for primed or naive hiPSCs (right). Genes were classified by mitochondrion complex and hierarchically clustered. **b** T2iLGö hiNPSCs have higher metabolic activity than their isogenic counterparts in RSeT and KSR+FGF2. A Seahorse apparatus was used to measure the oxygen consumption rate and the extracellular acidification rate of hiNPSC and hiPSC lines, maintained in indicated culture conditions. This figure presents six biological replicates. Each symbol in the panel is the average of a technical triplicate. **c** T2iLGö hiNPSCs have a higher resistance to inhibition of glycolysis. Quantification of colony numbers obtained after culture with the indicated concentrations of 2-deoxy-D-glucose. Primed cells were seeded in StemMACS™ iPS Brew XF, and naive cells were seeded in the indicated medium. Error bars indicate s.d. of three technical replicates. The presented experiment is representative of four independent experiments

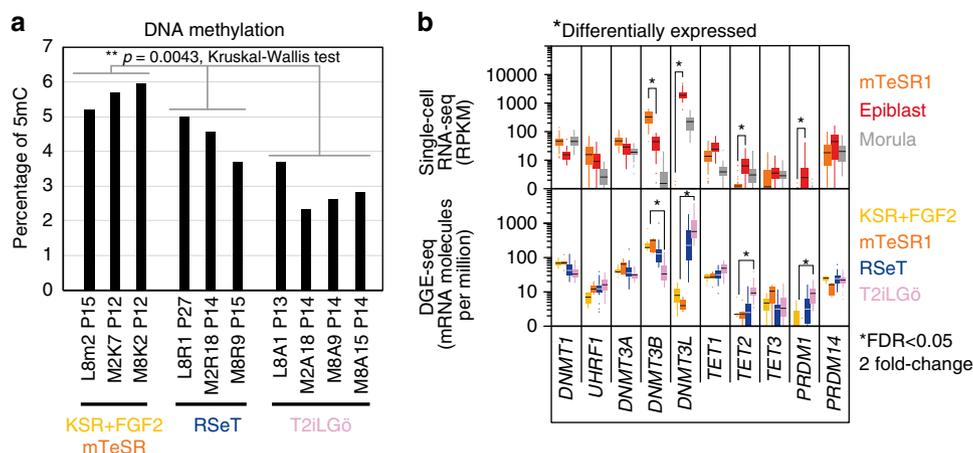


Fig. 4 T2iLGö hiNPCs are hypomethylated. **a** T2iLGö hiNPCs are hypomethylated in comparison to their RSeT and KSR+FGF2 counterparts. 5mC content is expressed as the percentage of 5mC in the total pool of cytosine for the indicated cell lines. Significance level was determined using Kruskal-Wallis test $**p < 0.01$. **b** Expression of indicated epigenetic-related genes is plotted as RPKM for single-cell RNA-seq or mRNA molecules per million of total mRNA molecules for DGE-seq. Error bars are defined as s.e.m. *Differentially expressed gene, as defined in Methods

cells. A recent report showed that specific pluripotent states might be needed depending on the recipient species, limiting the success of the chimeras⁴⁰. During the revision of our manuscript, Yang et al.⁴¹ described a specific medium allowing generation of pluripotent stem cells with extended potential to contribute to chimeras (EPS cells). However, further characterization is needed to show if those cells correspond to naive pluripotency or have specific features granting them superior chimerism capabilities. In that context, it will be interesting to see how EPS cells qualify using our proposed readouts (Fig. 7).

Altogether, direct reprogramming of somatic cells into hiNPCs will alleviate ethical issues linked with hESCs, therefore spreading the availability of this important cellular model. Indeed, naive PSCs are considered an alternative to human embryos to study regulation of human pluripotency, model preimplantation development and gonad diseases⁴². A concern for the clinical use of PSCs lies within their ability to keep a stable genome and epigenome, such as the X-chromosome dosage compensation in humans which is deregulated in primed hPSCs after long-term culture and is therefore a potential barrier for regenerative medicine³⁷. Combining knowledge obtained from both primed and naive hPSCs will contribute to a better understanding of molecular processes involved in human pluripotency like X-chromosome dynamics, facilitating the development of hPSC-based therapies.

Methods

Human preimplantation embryos. The use of human embryos donated to research was allowed by the French embryo research oversight committee: Agence de la Biomédecine, under approval number RE13-010. All human preimplantation embryos used in this study were obtained from and cultured at the Assisted Reproductive Technology unit of the University Hospital of Nantes, France, which are authorized to collect embryos for research under approval number AG11 0126AMP of the Agence de la Biomédecine. Embryos used were initially created in the context of an assisted reproductive cycle with a clear reproductive aim and then voluntarily donated for research once the patients have fulfilled their reproductive needs, or tested positive for the presence of monogenic diseases. All embryos used in this study were given to research after double consents from both parents. Donors did not receive any financial compensation. Molecular analysis of the embryos was performed in compliance with the embryo research oversight committee and The International Society for Stem Cell Research (ISSCR) guidelines⁴³.

Human preimplantation embryos culture. Day 3 cryopreserved embryos were thawed using Sydney IVF Thawing Kit (Cook Medical) and cultured in G2 Plus (Vitrolife), a specific medium for culture of embryos from day 3 to the blastocyst stage. Embryos were loaded into the Embryoscope® (Unisense Fertilitech®), a tri-gas incubator with a built-in microscope allowing time-lapse monitoring of early embryo development.

For embryos affected by a monogenic disease, insemination was achieved by intracytoplasmic sperm injection. Vitrolife® sequential media were used for embryo culture, with embryos being cultured in G1plus® medium from day 0 to day 3 and then transferred to a new pre-equilibrated slide containing G2plus® medium and cultured from day 3 onwards. Embryo biopsy of one or two blastomeres was performed at day 3 and the genetic results were obtained at day 4. Embryo culture was performed at 37 °C under a controlled atmosphere with low oxygen pressure (5% O₂, 5% CO₂). Embryos were fixed at the morula, B2 or B4 stages according to the grading system proposed by Gardner and Schoolcraft⁴⁴. Staging details of the embryos that are presented in Fig. 2b were as follows: the morula was fixed at 97 h post fertilization, and contained 24 cells; the B2 blastocyst passed through the morula stage at 59 h post thawing (8 cell stage), and was fixed at 76 h post thawing, it contained 33 cells; the B4 blastocyst passed through the morula stage at 95 h post fertilization, B2 stage at 111 h and was fixed at B4 stage at 118 h, it contained 135 cells, among which 8 PE cells (SOX17+) and 7 EPI cells (KLF17+).

Human cell lines. Three donor fibroblasts were used in this study, all of them being healthy donors: (1) B1 male fibroblasts are commercial BJ human neonatal fibroblasts extracted from normal human foreskin (Stemgent cat. no. 08-0027), (2) L8 female fibroblasts are normal human adult dermal fibroblasts extracted from a healthy woman aged 57 years, which are commercially available (Lonza cat. no. CC-2511 Lot 0000258580), (3) M1PS female fibroblasts are from three female patients from the Milieu Interieur Labex consortium, M2 and M3 are in their thirties while M8 is in the sixties. Human fibroblasts from the consortium were obtained after informed consent of patients, acknowledging the generation of hiPSC lines and use of those pluripotent lines for research. Primed hiPSCs from L7 human adult dermal fibroblasts extracted from a healthy male aged 51 years (Lonza cat. no. CC-2511 Lot 0000293971) were also used in this study. hESC lines H1 (WA01 Lot WB0111) and H9 (WA09 Lot WB0090) were obtained from the WiCell Research Institute, under authorization RE13-004 from the French embryo research oversight committee, Agence de la Biomédecine.

Tissue culture. Fibroblasts were cultured in fibroblast medium, composed of high glucose Dulbecco's modified Eagle's medium (DMEM) GlutamaxII (Life Technologies) supplemented with 10% fetal bovine serum (Hyclone), 1% sodium pyruvate (Life Technologies) and 1% non-essential amino acids (Life Technologies).

Mouse embryonic fibroblasts (MEFs) were prepared as previously described⁴⁵ and cultured in fibroblast medium supplemented with 0.5% of penicillin-streptomycin (Life Technologies). MEF isolation was performed in compliance with the French law and under supervision of the UTE animal core facility, University of Nantes. MEFs were mitotically inactivated using mitomycin C (Sigma) to be used as feeder cells.

Primed PSCs on feeder cells were cultured in DMEM/F-12 (Life Technologies) supplemented with 20% Knockout™ serum replacement (Life Technologies), 1% non-essential amino acids (Life Technologies), 1% glutamax (Life Technologies), 50 μM 2-mercaptoethanol (Life Technologies) and 10 ng/ml fibroblast growth factor 2 (Peprotech). Primed PSCs were mechanically passaged by cutting colonies with a needle. Primed PSCs in feeder-free conditions were cultured on Matrigel (BD/Corning) in mTeSR1 media; cells were non-enzymatically dissociated with StemMACS passaging solution XF (Miltenyi Biotec) for passaging.

hiNPCs were cultured on feeder cells, either in RSeT™ medium (Stem Cell Technologies) or in T2iLGö medium^{7,19} which is composed of N2B27

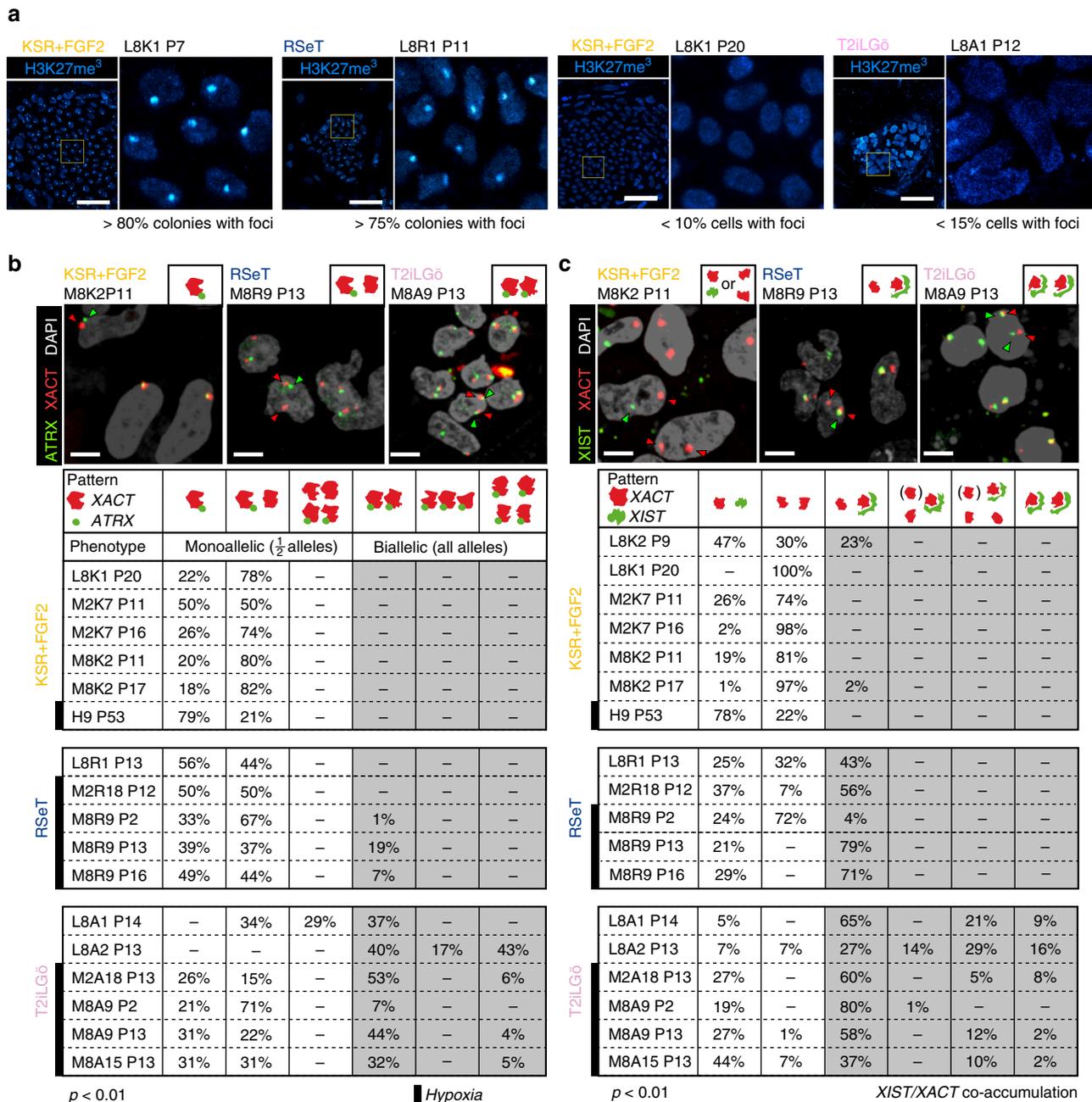


Fig. 5 T2iLGö hiNPCs have a X-chromosome status related to preimplantation epiblast. **a** T2iLGö hiNPCs or high-passage KSR-FGF2 do not display H3K27me₃ foci. Indicated cell lines were analyzed by immunofluorescence for H3K27me₃. This experiment is representative of three technical replicates performed at different passages of the cell lines. Scale bar = 50 μ m. **b**, **c** T2iLGö hiNPCs show signs of X-chromosome reactivation. mRNA FISH analysis for **b** ATRX and XACT or **c** XIST and XACT. For each cell line represented, more than 100 cells were investigated for their nuclear expression of the indicated mRNA. Quantifications for each combination are indicated below pictures. For each table, samples distribution between FISH counting were found statistically different ($p < 0.01$) by a homogeneity χ^2 test. Scale bar = 20 μ m

supplemented with 20 ng/ml LIF (Miltenyi Biotec), 1 μ M of PD0325901 (Axon Medchem), 1 μ M CHIR99021 (Axon Medchem) and 5 μ M Gö6983 (TOCRIS). N2B27 medium is composed of DMEM/F-12 (Life Technologies) supplemented with 1% N2 (Life Technologies), 1% B27 (Life Technologies), 1% non-essential amino acids (Life Technologies), 1% glutamax (Life Technologies), 0.1 mM 2-mercaptoethanol (Life Technologies), 50 μ g/ml bovine serum albumin (Sigma) and 0.5% penicillin–streptomycin (Life Technologies). hiNPCs were passaged using TrypLE (Life Technologies) for 5 min at 37 $^{\circ}$ C.

Naive and primed hPSCs were cultured at 37 $^{\circ}$ C under 20% O₂, 5% CO₂ and 10 μ M Y27632 (TOCRIS) was added in the medium upon cell seeding. M2 and M8 naive hPSCs were cultured at 37 $^{\circ}$ C under 5% O₂, 5% CO₂ and 10 μ M Y27632.

Somatic cell lines have been tested for mycoplasma presence using the MycoAlert kit (LONZA, LT07-318) before reprogramming. Only if the test was

negative, reprogramming was performed. Each iPSC line generated was tested for mycoplasma using the MycoAlert kit at various time points to ensure mycoplasma absence in both primed and naive hiPSCs.

Reprogramming of human somatic cells into iPSCs. Fibroblasts were reprogrammed using the CytoTune-iPS 2.0 Sendai reprogramming kit from Life Technologies. Two days before infection, 40,000 fibroblasts were seeded per well on a 12-well plate, coated with Matrigel. At day 0, cells were counted and infected with the three vectors: polycistronic Klf4-Oct4-Sox2, cMyc and Klf4 at a 5, 5 and 3 multiplicity of infection, respectively. At day 7 of infection, cells were dissociated using TrypLE and seeded on 3 \times 35 mm dishes coated with mouse feeder cells. Cells were switched to naive pluripotency medium (RSeT or T2iLGö) or TeSR-E7 medium at day 9. For each of our reprogramming campaigns, we obtained more

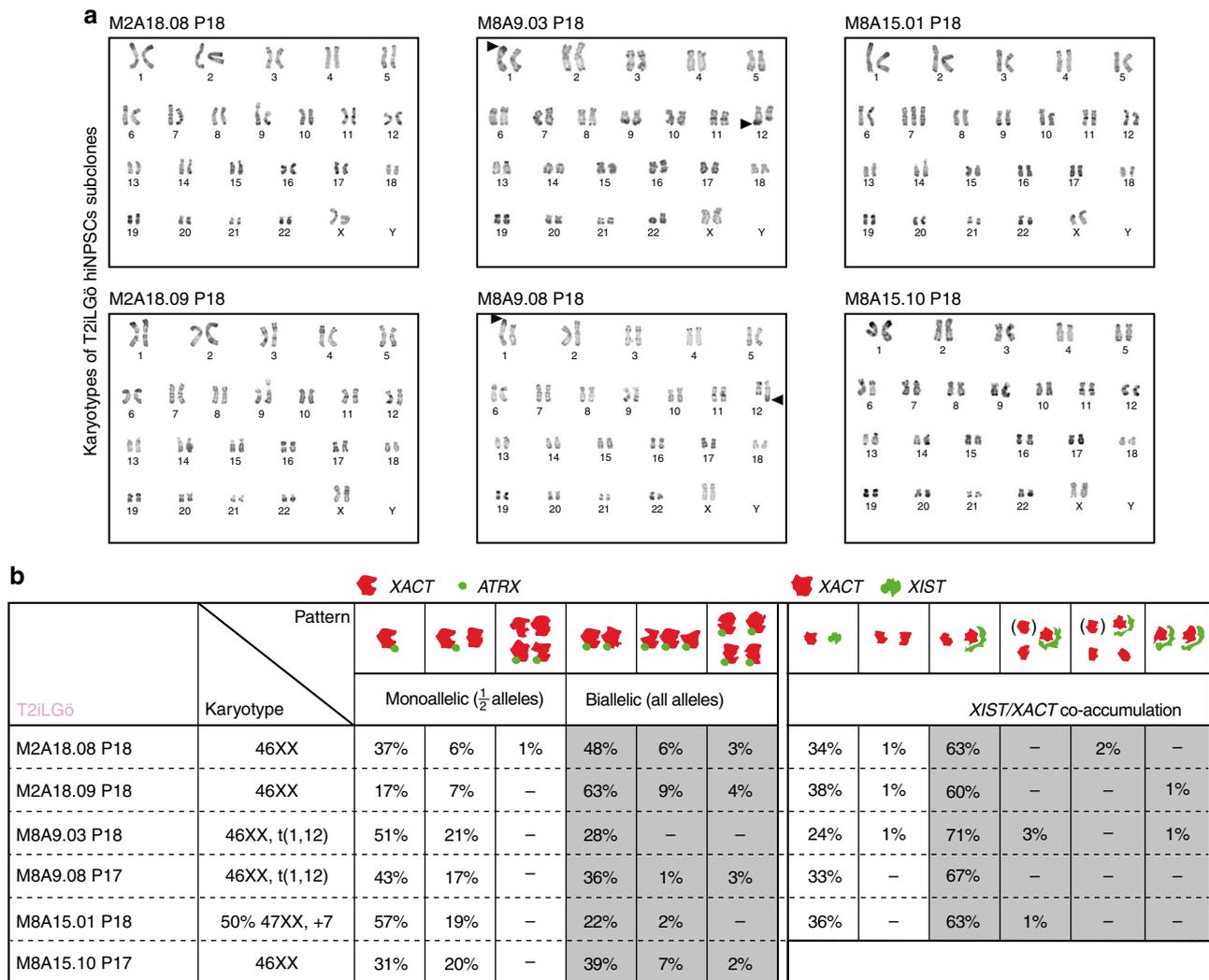


Fig. 6 Clonal analysis of X-chromosome reactivation in T2iLG6 hiNPSCs. **a** Representative karyotypes of T2iLG6 subclones. **b** Karyotype and mRNA FISH analysis for *ATR*X and *XACT* (left) or *XIST* and *XACT* (right) of indicated subclones. For each subclone represented, more than 50 cells were investigated for their nuclear expression of the indicated mRNA. Quantifications for each combination are indicated below pictures

than 100 colonies in T2iLG6, TeSR-E7 and RSeT. RSeT and T2iLG6 hiNPSCs were trypsinized for passaging, primed hiPSCs were mechanically passaged to fresh feeder-coated tissue culture dishes, in KSR+FGF2, between day 16 and day 24.

SeaHorse analysis. Oxygen consumption rate and extracellular acidification rate were measured using an XF24 Analyser (SeaHorse Bioscience). At confluence, hiNPSCs and hiPSCs were dissociated using TrypLE and cells were incubated on gelatin for 30 min at 37 °C to remove feeder cells. The SeaHorse plate was pre-treated with 2 µg/ml of Cell-Tak Cell and tissue adhesive (Corning). KSR, RSeTTM and T2iLG6 cells were seeded at 200,000, 150,000 and 100,000 cells per well respectively prior to the experiment. Cells were incubated for 1 h at 37 °C and atmospheric CO₂ in DMEM (Sigma Aldrich) supplemented with 10 mM glucose, 2 mM glutamine, 2 mM pyruvate and the pH was adjusted to 7.4 using NaOH. During the mito stress kit experiment, Oligomycin (2 µM), CCCP (0.75 µM), Antimycin-A (2 µM) and Rotenone (1 µM) were injected at indicated time points.

Colony formation assay in 2-deoxy-D-glucose. hiNPSCs or hiPSCs were seeded at 3000 cells per well in a 12-well plate, coated with feeder cells, in their respective media in addition to 10 µM Y27632. Of note, KSR cells were seeded and cultured in iPS Brew to boost clonogenicity of the cells. From day 1 after seeding onwards, 2 mM or 4 mM of 2-deoxy-D-glucose were supplemented in the culture medium. Cells were fixed between day 4 and day 6 post seeding and stained for alkaline phosphatase using the SIGMA FASTTM BCIP[®]/NBT kit (Sigma). Images were acquired using the Cellomics ArrayscanVTI (Thermo Fisher) at a 5× magnification. Colonies were counted manually. Presented results are from a representative experiment performed in triplicate.

Karyotype analysis. Karyotyping based on RTG-banding was performed at Cytogenetic Laboratory (CHU Nantes) using standard methods with minor modifications. Briefly, hiPSCs and hiNPSCs were plated on Lab-Tek chamber slide. At 70% confluence, PSCs were submitted to hypotonic shock (20% fetal bovine serum in water), fixed in methanol/glacial acetic acid 3:1, then stained in Giemsa stain. Metaphase spreads for each sample were analyzed. A total of 30 pictures per slides were automatically acquired, chromosomes counted and at least 10 karyotypes for each cell line were classified. The commercial L8 fibroblast line contains a translocation from the chromosome 10 to 7 that is found in all derived iPSC lines. The B1 and M3 fibroblasts had normal karyotypes.

Immunofluorescence. For IF analysis, cells and embryos were fixed at room temperature using 4% paraformaldehyde for 15 min and 30 min (on a rotating shaker for embryos), respectively. Samples were then permeabilized and blocked in IF buffer (IF buffer: phosphate-buffered saline (PBS)–0.2% Triton, 10% fetal bovine serum) for 60 min at room temperature. Samples were incubated with primary antibodies overnight at 4 °C. Incubation with secondary antibodies was performed for 2 h at room temperature along with 4',6-diamidino-2-phenylindole (DAPI) counterstaining. Primary and secondary antibodies with dilutions used in this study are listed in Supplementary Table 2.

Imaging flow cytometry. Cells were stained with the Zombie NIRTM viability kit (BioLegend cat. no. 423105) for 20 min on ice. Potential nonspecific binding sites were blocked by incubation with human serum (obtained from healthy donor at the French blood establishment EFS) diluted to 1:20 in PBS for 30 min. Before performing intracellular staining, cells were fixed and permeabilized for 30 min on ice, using the Fixation/Permeabilization concentrate (eBioscience cat. no. 00-5123)

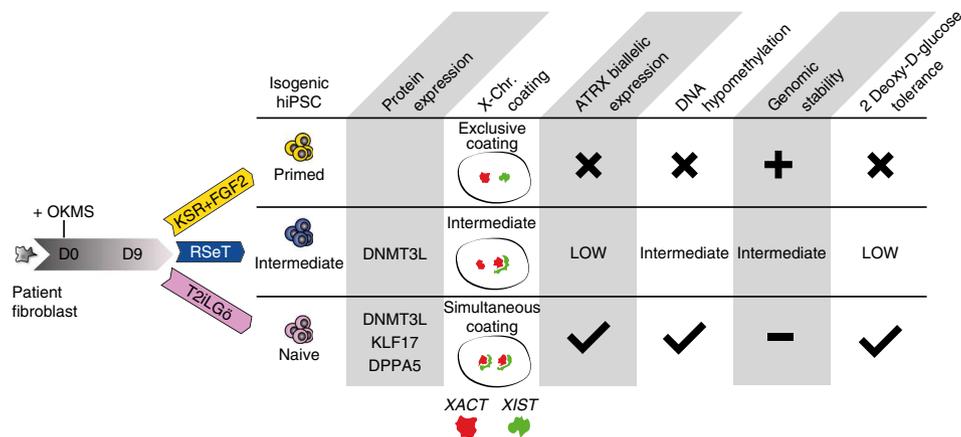


Fig. 7 hiNPCs in T2iLG6 achieve the most naive pluripotency hallmarks. The presented reprogramming method enables to simultaneously generate isogenic hiPSCs in KSR+FGF2, RSeT and T2iLG6 media. The naive pluripotency level of the generated cell lines can be assessed with specific markers, X-chromosome activity status, DNA methylation level and capacity to tolerate glycolysis inhibition

diluted in Fixation/Permeabilization Diluent (eBioscience cat. no. 00-5223) and in the Permeabilization buffer (eBioscience cat. no. 00-8333). Cells were incubated with the primary antibody (anti-KLF17, 1:100) for 45 min at room temperature. Further incubation with the secondary antibody (goat anti-rabbit, 1:500) was performed for 30 min on ice. Before imaging, DAPI was added to stain nuclei of the cells. References of antibodies with dilutions used in this study are listed in Supplementary Table 2.

Analyses were performed using an ImageStreamX Mark II Imaging Flow Cytometer (Amnis Corporation, Seattle, WA) equipped with the INSPIRE software.

A 40× magnification was used for all samples. Data analysis was performed using the IDEAS software (Amnis Corporation). The Zombie NIR[®] was excited with a 642 nm laser (power 50 mW) and the fluorescence signal was collected on channel 12 (745–800 nm). The DAPI was excited with a 405 nm laser (power 60 mW) and the fluorescence signal was collected on channel 7 (430–505 nm). KLF17 coupled to an Alexa 488 was excited with a 488 nm laser (power 80 mW) and the fluorescence signal was collected on channel 2 (480–560 nm). Intensity-adjusted brightfield images were collected on channel 1 (430–480 nm). The gating strategy for analysis involved the selection of focused, single and living cells on viability marker, then on DAPI and KLF17 fluorescence.

DNA methylation. For mass spectrometry analysis of DNA methylation, 1 µg of genomic DNA was analyzed using liquid chromatography triple-quadrupole mass spectrometry (KU Leuven Metabolomics Core). The concentration (µM) of Cytosine (unmodified), 5mC and 5-hydroxymethyl-cytosine (5hmC) were obtained using standard curves of known C, 5mC and 5hmC amounts. The percentage of 5mC or 5hmC in DNA was obtained by calculating the ratio of 5mC or 5hmC to the total pool of C.

RNA-FISH. RNA-FISH was performed as previously described³⁵. Briefly, cells were fixed between 36 h and 50 h post seeding in 3% paraformaldehyde for 10 min at room temperature. Cells were permeabilized in CSK buffer supplemented with 1 mM EGTA, 0.5% Triton and RNaseOUT inhibitor (20 U/ml) for 5 min on ice. After 3 washes in 70% EtOH, cells were dehydrated in 90% and 100% EtOH and incubated overnight with probes at 37 °C. After three 50% formaldehyde/2× SSC washes and three 2× SSC washes at 42 °C for 4 min, coverslips were mounted in Vectashield plus DAPI. SpectrumGreen or SpectrumRed-labeled probes (Vysis) were generated by nick translation for human *XIST*, *XACT* (RP11-35D3, BACPAC) and *ATRX* (RP11-42M11, BACPAC Resource). Images were acquired on an inverted Nikon A1 confocal microscope, according to the Shannon–Nyquist sampling rate. mRNA expression of *XIST*, *XACT* and *ATRX* are manually counted in more than 100 cells per cell line.

Western blot. Cells were lysed in 100 µl of TNTE buffer supplemented with protease inhibitor cocktail (Sigma) and phosphatase inhibitor cocktail (Sigma). TNTE buffer were composed of 50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA and 0.5% Triton X-100. Then, 20 µg of proteins samples were denatured using NuPAGE sample reducing agent and LDS sample buffer (Invitrogen) for 5 min at 98 °C. Next, 6 µl of spectra[™] multicolor broad-range protein ladder (Invitrogen) or 20 µg of denatured protein samples were loaded on a 4–15% mini-PROTEAN[®] TGX stain-free[™] precast gels (BioRad), transferred on trans-blot[®] turbo[™] RTA midi nitrocellulose transfer kit membranes (BioRad). The membranes were blocked for 1 h in tris-buffered saline with Tween-20/5% milk, incubated overnight with primary antibodies and incubated 1 h with secondary antibodies. Signal was revealed with Super signal west femto maximum sensitivity substrate (Thermo scientific) for DPPA5 or clarity[™] ECL western blotting

substrate (BioRad) for glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and imaged on a Chemidoc[™] MP system (BioRad). Primary and secondary antibodies are listed in Supplementary Table 2. The stain-free blot image and the uncropped blot images can be found in Supplementary Fig. 6.

RNA extraction and quantitative real-time PCR. Total RNA was extracted using RNeasy[®] columns and DNase-treated using RNase-free DNase (Qiagen). For quantitative PCR, first-strand complementary DNAs (cDNAs) were generated using 500 ng of RNA, M-MLV reverse transcriptase (Invitrogen), 25 µg/ml polydT and 9.6 µg/ml random primers (Invitrogen).

To quantitate transcripts, absolute quantitative PCR was performed on a Viia 7 (Applied Biosystems) using power SYBR green PCR master mix (Applied Biosystems), for genes listed in the primers table (Supplementary Table 3). For each sample, the ratio of specific mRNA level relative to GAPDH levels was calculated. Experimental results are shown as levels of mRNA relative to the highest value.

All quantitative real-time PCR primers have a hybridization temperature of 60 °C and their sequences are listed in Supplementary Table 3. All amplicons span two adjacent exons. SeV primers are from Life Technologies (cytotune 2.0 kit).

Expression profiling by single-cell RNA-seq and DGE-seq. For single-cell RNA-seq, H1 and H9 cells were sorted on a FACS Aria in 5 µl lysis buffer (1:500 Phusion buffer, NEB; 1:20 RNASE out, Life Technologies), and frozen at –80 °C. The SmartSeq2 libraries were prepared according to the SmartSeq2 protocol^{46,47} with some modifications⁴⁸. Briefly, total RNA was purified using RNA-SPRI beads. Poly(A)+mRNA was converted to cDNA which was then amplified. cDNA was subject to transposon-based fragmentation that used dual indexing to barcode each fragment of each converted transcript with a combination of barcodes specific to each sample. In the case of single-cell sequencing, each cell was given its own combination of barcodes. Barcoded cDNA fragments were then pooled prior to sequencing. Sequencing was carried out as paired end 2 × 25 bp with an additional 8 cycles for each index.

The FASTQ files were mapped with Tophat⁴⁹ on GRCH37.75.gtf genome version with Bowtie2⁵⁰ (human_g1k_v37). Of note, FASTQ from¹² were generated by single-end RNA-seq and in our data set in paired-end RNA-seq. HTSeq⁵¹ was used to generate raw counts tables from BAM files. For each sample, Q30 percentage was calculated with FASTQC, and samples with a score above 75% were kept. Additional filters were employed: samples with more than 5000 genes detected were kept, and a final gene filtering step was performed to keep genes with a sum of at least 10 counts across the 1976 samples. Samples with total quantity of count 2 s.d. away from mean total quantity of count were excluded.

Counts were normalized with scran⁵² and log₂ transformed for variance analysis (PCA and clustering on correlations matrix)

To obtain RPKM, BAM were computed to count using featureCount from Rsubread with GRCH37. Then, counts were normalized with calcNormFactors from edgeR⁵³ with default parameters, and RPKM were finally obtained using rpkm function from edgeR on normalized counts. RPKM were used for Figs. 2a and 4b.

For 3' DGE, RNA-sequencing protocol was performed according to ref. 28. Briefly, the libraries were prepared from 10 ng of total RNA. The mRNA poly(A) tails were tagged with universal adapters, well-specific barcodes and unique molecular identifiers (UMIs) during template-switching reverse transcriptase. Barcoded cDNAs from multiple samples were then pooled, amplified and tagged using a transposon-fragmentation approach which enriches for 3' ends of cDNA. A library of 350–800 bp was run on an Illumina HiSeq 2500 using a HiSeq Rapid SBS Kit v2-50 cycles (ref FC-402-4022) and a HiSeq Rapid PE Cluster Kit v2 (ref PE-402-4002).

Read pairs used for analysis matched the following criteria: all 16 bases of the first read had quality scores of at least 10 and the first 6 bases correspond exactly to a designed well-specific barcode. The second reads were aligned to RefSeq human mRNA sequences (hg19) using bwa version 0.7.4 with non-default parameter “-l 24”. Reads mapping to several positions into the genome were filtered out from the analysis. DGE profiles were generated by counting for each sample the number of unique UMIs associated with each RefSeq genes. DGE-sequenced samples were acquired from three sequencing runs. Batch effects were corrected with the limma library function “removeBatchEffect”. All sequenced samples were retained for further analysis.

DESeq2 was used to normalize expression with the DESeq function⁵⁴. Normalized counts were transformed with vst (variance stabilized transformation) function from DESeq library. This log-like transformation was used for variance analysis.

Batch effects were corrected with the limma library function “removeBatchEffect”. All sequenced samples were retained for further analysis. This represents 78 samples cultured from passage 1 to 40 (Supplementary Table 1) in following media: 26 from T2iLGo, 26 RSet, 9 KSR+FGF2, 5 mTESR, 4 fibroblasts and 8 E7.

Samples assignment of single-cell RNA-seq data. Single-cell samples used in epiblast–hESC comparisons came from two data sets: the first was a subset of 52 hESCs from our own RNA-sequencing, the second was 104 cells from Petropoulos et al.¹². E4 samples were labeled as morula; epiblast cells were labeled from E5 and E6 blastocysts, after a filtering of cells expressing epiblast markers (Supplementary Table 4). Blastocysts from Yan et al.¹¹ were clustered by preimplantation lineage markers (Supplementary Table 4), and a cluster of 5 cells were selected and annotated as epiblast. Two outliers were removed from cells annotated as morula.

For the analysis of KLF17, GATA2 and SOX17 expression profile over development time, samples were stratified per embryonic days. Trophectoderm, and inner cell mass cells were segregated by unsupervised clustering using known lineage markers (Supplementary Table 4). A second clustering was applied on inner cell mass cells to segregate epiblast and primitive endoderm.

Differential expression profiling. For the DGE-seq data set, differential expressed *p*-values were processed with DESeq2 and FDRs were estimated with Benjamini–Hochberg procedure. For single-cell data set, *p*-values and FDR were computed with ROTS⁵⁵. Genes with FDR < 0.05 and a fold change < 0.5 or > 2 were qualified as differentially expressed for both data sets. Two-sided Fisher’s exact test was computed to test dependency between single-cell differentially expressed genes and DGE-seq differentially expressed genes with the contingency table found in Supplementary Table 5.

Processing of principal component analysis. PCAs were computed with R princomp functions from centered data and plotted with R library ggplot2. PCA from Fig. 1d were computed from four data sets: Yan et al.¹¹, Petropoulos et al.¹² and this paper (single-cell and Bulk DGE-seq data). Each data set was transformed into transcripts per million, quantile normalized and *z*-scored by row separately. The data sets were reunified on the 15,315 genes in common for PCA computing.

To generate Supplementary Fig. 2c, a PCA was made for each of the three sets of pathways from Supplementary Fig. 2b (Supplementary Data 2 Tables 3–8). Each single-cell or DGE-seq sample was projected on the first component of the 3 PCA, and the first component coordinates were used to generate the three-dimension graph.

Processing of heatmaps. Heatmaps were drawn with the library complexHeatmap with *z*-score of expression. Cluster trees were computed with pvclust⁵⁶ with correlation method as distance calculation and Ward criteria as construction method.

For expression profile of KEGG pathways, genes were ordered per fold changes.

Functional enrichment. topGO⁵⁷ was used to identify enriched GO terms (Supplementary Data 2 Tables 9–10). Enrichment was performed by comparing GO terms present in differentially expressed genes vs. the whole transcriptome data set. Three annotation databases of GO terms were used (org.Hs.eg.db): Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). According to the reference manual, *p*-values were computed with the “classic” and “elim” method algorithm parameter and “Fisher” as statistic parameter.

A gene set analysis method GAGE⁵⁸ was used with KEGG database to identify differentially regulated pathways. Pathways with FDR < 0.01 were retained for further analysis. Gage was used on unpaired mode with parameter “same dir” in false mode.

Quantification and statistical analysis. All data presented are representative of at least three independent experiments that yielded similar results. Statistical analyses were performed using the software Prism (Graphpad) or R.

DESeq2 was used for analysis considering RNA-seq data were following a negative binomial distribution. Other statistical tests were performed considering their specific assumption and hypothesis, notably for Pearson correlation’s test of

Supplementary Fig. 3c and homogeneity χ^2 tests of Fig. 5b, c. All graphical representations were chosen to accurately display variation within each group.

For each experiment, sampling was done to have comfortable group size that provide statistically robust results. For each figure and statistical analysis from RNA-seq data, size of each group is listed in the Supplementary Table 6.

Data and software availability. The authors declare that all data supporting the findings of this study are available within the article and its supplementary information files or from the corresponding author upon reasonable request.

The raw read sequence data and sample annotations generated for this paper are available at European Nucleotide Archive (ENA) with accession number PRJEB18663.

Received: 16 August 2017 Accepted: 27 October 2017

Published online: 24 January 2018

References

- Nichols, J. & Smith, A. Naive and primed pluripotent states. *Cell Stem Cell* **4**, 487–492 (2009).
- Weinberger, L., Ayyash, M., Novershtern, N. & Hanna, J. H. Dynamic stem cell states: naive to primed pluripotency in rodents and humans. *Nat. Rev. Mol. Cell Biol.* **17**, 155–169 (2016).
- Chen, H. et al. Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. *Nat. Commun.* **6**, 7095 (2015).
- Chan, Y. S. et al. Induction of a human pluripotent state with distinct regulatory circuitry that resembles preimplantation epiblast. *Cell Stem Cell* **13**, 663–675 (2013).
- Gafni, O. et al. Derivation of novel human ground state naive pluripotent stem cells. *Nature* **504**, 282–286 (2013).
- Valamehr, B. et al. Platform for induction and maintenance of transgene-free hiPSCs resembling ground state pluripotent stem cells. *Stem Cell Rep.* **2**, 366–381 (2014).
- Takashima, Y. et al. Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* **158**, 1254–1269 (2014).
- Theunissen, T. W. et al. Systematic Identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* **15**, 471–487 (2014).
- Ware, C. B. et al. Derivation of naive human embryonic stem cells. *Proc. Natl. Acad. Sci. USA* **111**, 4484–4489 (2014).
- Blakeley, P. et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* **142**, 3151–3165 (2015).
- Yan, L. et al. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.* **20**, 1131–1139 (2013).
- Petropoulos, S. et al. Single-Cell RNA-Seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* **165**, 1012–1026 (2016).
- Guo, H. et al. The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
- Smith, Z. D. et al. DNA methylation dynamics of the human preimplantation embryo. *Nature* **511**, 611–615 (2014).
- Vallot, C. et al. XACT noncoding RNA competes with XIST in the control of X chromosome activity during human early development. *Cell Stem Cell* **20**, 102–111 (2017).
- De Los Angeles, A. et al. Hallmarks of pluripotency. *Nature* **525**, 469–478 (2015).
- Pastor, W. A. et al. Naive human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. *Cell Stem Cell* **18**, 323–329 (2016).
- Sahakyan, A. et al. Human naive pluripotent stem cells model X chromosome dampening and X inactivation. *Cell Stem Cell* **20**, 87–101 (2016).
- Guo, G. et al. Naive pluripotent stem cells derived directly from isolated cells of the human inner cell mass. *Stem Cell Rep.* **6**, 437–446 (2016).
- Collier, A. J. et al. Comprehensive cell surface protein profiling identifies specific markers of human naive and primed pluripotent states. *Cell Stem Cell* **20**, 874–890 e877 (2017).
- Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317 (2007).
- Maherali, N. et al. Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1**, 55–70 (2007).
- Wernig, M. et al. In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318–324 (2007).
- Ying, Q. L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
- Guo, G. et al. Epigenetic resetting of human pluripotency. *Development* **144**, 2748–2763 (2017).

26. Choi, J. et al. DUSP9 modulates DNA hypomethylation in female mouse pluripotent stem cells. *Cell Stem Cell* **20**, 706–719 e707 (2017).
27. Cacchiarelli, D. et al. Integrative analyses of human reprogramming reveal dynamic nature of induced pluripotency. *Cell* **162**, 412–424 (2015).
28. Soumillon, M., Cacchiarelli, D. & Semrau, S. Characterization of directed differentiation by high-throughput single-cell RNA-Seq. Preprint at *bioRxiv* <https://www.biorxiv.org/content/early/2014/03/05/003236> (2014)
29. Niakan, K. K., Han, J., Pedersen, R. A., Simon, C. & Pera, R. A. Human pre-implantation embryo development. *Development* **139**, 829–841 (2012).
30. Theunissen, T. W. et al. Molecular criteria for defining the naive human pluripotent state. *Cell Stem Cell* **19**, 502–515 (2016).
31. Ogata, H. et al. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).
32. Gu, W. et al. Glycolytic metabolism plays a functional role in regulating human pluripotent stem cell state. *Cell Stem Cell* **19**, 476–490 (2016).
33. Durruthy-Durruthy, J. et al. Spatiotemporal reconstruction of the human blastocyst by single-cell gene-expression analysis informs induction of naive pluripotency. *Dev. Cell* **38**, 100–115 (2016).
34. Okamoto, I. et al. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* **472**, 370–374 (2011).
35. Vallot, C. et al. Erosion of X chromosome inactivation in human pluripotent cells initiates with XACT coating and depends on a specific heterochromatin landscape. *Cell Stem Cell* **16**, 533–546 (2015).
36. Mekhoubad, S. et al. Erosion of dosage compensation impacts human iPSC disease modeling. *Cell Stem Cell* **10**, 595–609 (2012).
37. Patel, S. et al. Human embryonic stem cells do not change their X inactivation status during differentiation. *Cell Rep.* **18**, 54–67 (2017).
38. Smith, A. Formative pluripotency: the executive phase in a developmental continuum. *Development* **144**, 365–373 (2017).
39. Buecker, C. et al. A murine ESC-like state facilitates transgenesis and homologous recombination in human pluripotent stem cells. *Cell Stem Cell* **6**, 535–546 (2010).
40. Wu, J. et al. Interspecies chimerism with mammalian pluripotent stem cells. *Cell* **168**, 473–486 e415 (2017).
41. Yang, Y. et al. Derivation of pluripotent stem cells with in vivo embryonic and extraembryonic potency. *Cell* **169**, 243–257 e225 (2017).
42. von Meyenn, F. et al. Comparative principles of DNA methylation reprogramming during human and mouse in vitro primordial germ cell specification. *Dev. Cell* **39**, 104–115 (2016).
43. Daley, G. Q. et al. Setting global standards for stem cell research and clinical translation: the 2016 ISSCR guidelines. *Stem Cell Rep.* **6**, 787–797 (2016).
44. Alpha Scientists in Reproductive Medicine and ESHRE Special Interest Group of Embryology. The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Hum. Reprod.* **26**, 1270–1283 (2011).
45. Samavarchi-Tehrani, P. et al. Functional genomics reveals a BMP-driven mesenchymal-to-epithelial transition in the initiation of somatic cell reprogramming. *Cell Stem Cell* **7**, 64–77 (2010).
46. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
47. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
48. Trombetta, J. J. et al. Preparation of single-cell RNA-Seq libraries for next generation sequencing. *Curr. Protoc. Mol. Biol.* **107**, 4 22 21–17 (2014).
49. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
51. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
52. Lun, A. T., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* **5**, 2122 (2016).
53. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
54. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
55. Suomi, T., Seyednasrollah, F., Jaakkola, M. K., Faux, T. & Elo, L. L. ROTS: An R package for reproducibility-optimized statistical testing. *PLoS Comput. Biol.* **13**, (e1005562) (2017).
56. Suzuki, R. & Shimodaira, H. Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).
57. Alexa, A., Rahnenfuhrer, J. & Lengauer, T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**, 1600–1607 (2006).
58. Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10**, 161 (2009).

Acknowledgements

We thank S. Bézie (UMR1064) and the following core facilities for their support: GenoBIRD, Cytocell, Micropicell, the Center of Excellence Nikon Nantes and iPSC core facility. We thank the "service de génétique médicale" of CHU de Nantes for the help with the karyotypes. This work was supported by "Paris Scientifique region Pays de la Loire: HUMPLURI" and IHU CESTI. S.K. is recipient of fellowships from Progreffe and Fondation pour la Recherche Médicale (FDT20160435459). D.M. is supported by FINOX Biotech FORWARD initiative. M.C. is supported by the "Who Am I?" Laboratory of Excellence #ANR-11-LABX-0071 funded by the French Government through its "Investments for the Future" program operated by the ANR under grant #ANR-11-IDEX-0005-01. L.F. is supported by the "Paris Scientifique region Pays de la Loire: TIPS" and LabEx IGO project (no. ANR-11-LABX-0016-01) funded by the «Investissements d'Avenir» French Government program, managed by the French National Research Agency (ANR). Milieu Interieur is supported by the French government's Invest in the Future Program (ANR, reference 10-LABX-69-01). Research in the Pasque lab is supported by The Research Foundation – Flanders (FWO) (Odysseus Return Grant G0F7716N) and KU Leuven Research Fund (BOFZAP starting grant StG/15/021BF, C1 grant C14/16/077 and project financing PF/10/019). We thank Pr A. Smith and Dr G. Guo (Wellcome Trust-MRC Stem Cell Institute, University of Cambridge, UK) for providing HNES1 mRNA and sharing their protocols. L.Q.-M. and M.L.A. are co-coordinators of the Milieu Intérieur consortium. We thank Dr C. Hirsh (University of Toronto, Canada) for critical review of the manuscript.

Author contributions

S.K. and L.D. designed the study, with input from T.M., T.F., V.P., C.L.C., C.P., M.S., R.R., and J.B. S.K., D.M., C.R. and L.D. wrote the manuscript with input from all authors. A.R. and J.L. manipulated human embryos. D.M., Y.L. and E.C. performed bioinformatics analysis. S.K., D.M., C.C., A.R., A.G., C.V., A.D., M.S., W.D., S.N., M.C., J.S. and L.D. performed experiments. T.F. and P.B. supervised human embryo donation. All authors approved the final version of the manuscript.

Additional information

Supplementary Information accompanies this paper at <https://doi.org/10.1038/s41467-017-02107-w>.

Competing interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npj.nature.com/reprintsandpermissions/>

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

Stéphanie Kilens^{1,2,3}, Dimitri Meistermann^{1,2,3,4}, Diego Moreno^{1,2,3}, Caroline Chariou⁵, Anne Gaignerie⁵, Arnaud Reignier^{1,2,3,6}, Yohann Lelièvre⁴, Miguel Casanova⁷, Céline Vallot⁷, Steven Nedellec⁸, Léa Flippe^{1,2,3},

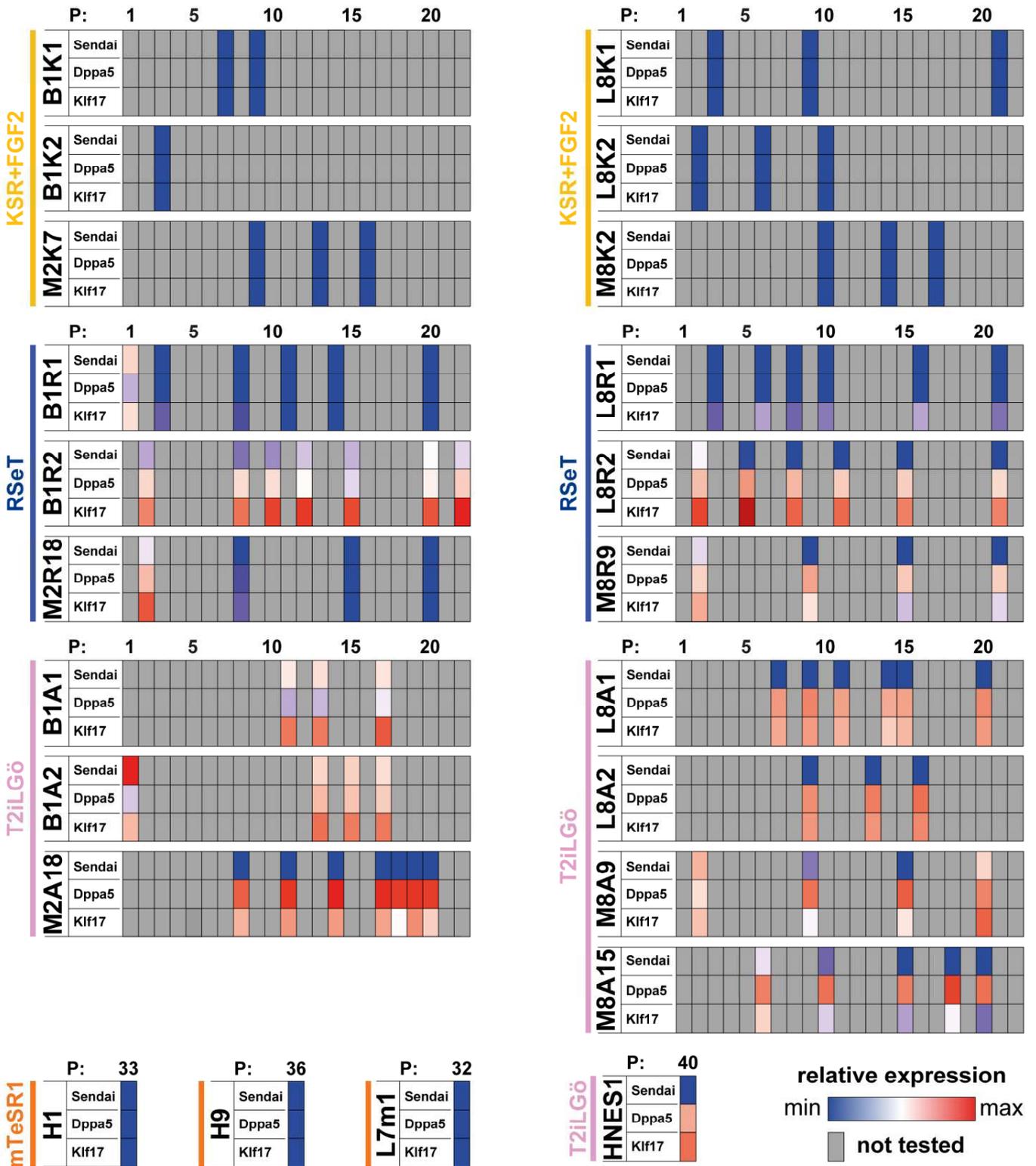
Julie Firmin^{1,2,3,6}, Juan Song⁹, Eric Charpentier¹⁰, Jenna Lammers^{1,2,3,6}, Audrey Donnart¹⁰, Nadège Marec¹¹, Wallid Deb¹², Audrey Bihouée¹⁰, The Milieu Intérieur Consortium, Cédric Le Caignec^{12,13}, Claire Pecqueur¹⁴, Richard Redon^{10,15}, Paul Barrière^{1,2,3,6}, Jérémie Bourdon⁴, Vincent Pasque⁹, Magali Soumillon^{16,17}, Tarjei S. Mikkelsen^{16,18} , Claire Rougeulle⁷, Thomas Fréour^{1,2,3,6} & Laurent David^{1,2,3,5}

¹Centre de Recherche en Transplantation et Immunologie UMR1064, INSERM, Université de Nantes, Nantes, France. ²Institut de Transplantation Urologie Néphrologie (ITUN), CHU Nantes, Nantes, France. ³LabEx IGO “Immunotherapy, Graft, Oncology”, Nantes, France. ⁴Laboratoire des Sciences du Numérique de Nantes, LS2N, UMR CNRS 6004, Université de Nantes, Nantes, France. ⁵INSERM UMS 016, SFR Francois Bonamy, iPSC Core Facility, Nantes, France; CNRS, UMS 3556, Nantes, France; Université de Nantes, Nantes, France; CHU Nantes, Nantes, France. ⁶CHU Nantes, Service de Biologie de la Reproduction, Nantes, France. ⁷Sorbonne Paris Cité, Epigenetics and Cell Fate, UMR 7216 CNRS, Université Paris Diderot, Paris, France. ⁸INSERM UMS 016, SFR Francois Bonamy, MicroPicell Core Facility, Nantes, France; CNRS, UMS 3556, Nantes, France; Université de Nantes, Nantes, France; CHU de Nantes, Nantes, France. ⁹KU Leuven-University of Leuven, Department of Development and Regeneration, Stem Cell Biology and Embryology Unit, Leuven Stem Cell Institute, Herestraat 49, B-3000 Leuven, Belgium. ¹⁰INSERM UMR1087, CNRS UMR6291 Université de Nantes l’institut du thorax, Nantes, France. ¹¹INSERM, UMS 016, SFR Francois Bonamy Cytocell Core Facility, Nantes, France; CNRS, UMS 3556, Nantes, France; Université de Nantes, Nantes, France; CHU Nantes, Nantes, France. ¹²CHU Nantes, Service de génétique médicale, Nantes, France. ¹³INSERM, UMR1238, Bone Sarcoma and Remodeling of Calcified Tissue, Nantes, France. ¹⁴CRCINA, INSERM, Université de Nantes, Nantes, France. ¹⁵CHU Nantes, l’institut du thorax, Nantes, France. ¹⁶Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA, 02138, USA; Broad Institute, Cambridge, MA 02142, USA.; Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138, USA. ¹⁷Present address: Berkeley Lights Inc., 5858 Horton Street, Emeryville, CA 94608, USA. ¹⁸Present address: 10x Genomics, 7068 Koll Center Pkwy #401, Pleasanton, CA 94566, USA. Stéphanie Kilens and Dimitri Meistermann contributed equally to this work. A full list of consortium members appears at the end of the paper.

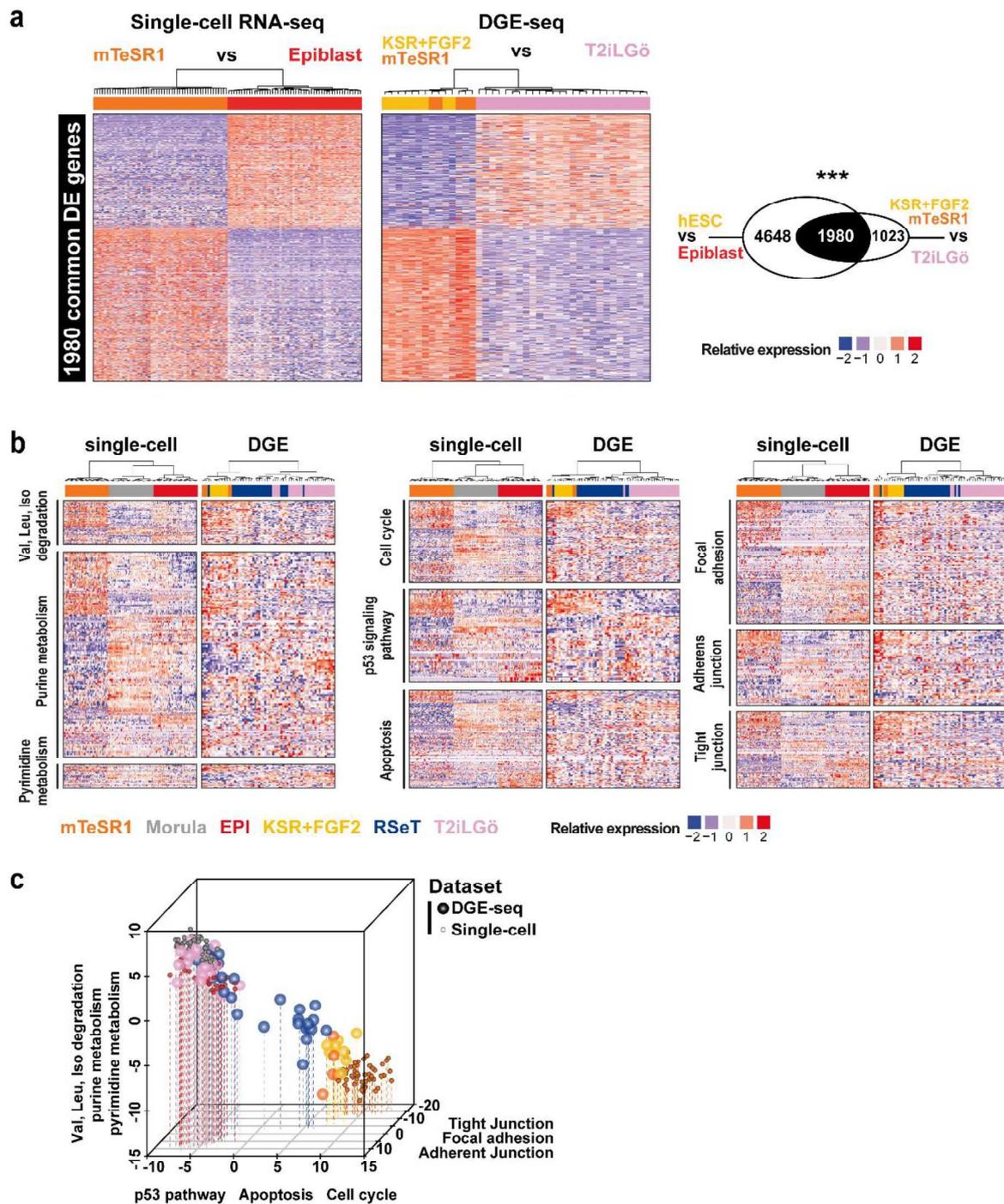
The Milieu Intérieur Consortium

Laurent Abel¹⁹, Andres Alcover²⁰, Kalla Astrom²¹, Philippe Bousso²², Pierre Bruhns²³, Ana Cumano²⁴, Darragh Duffy²⁵, Caroline Demangel²⁶, Ludovic Deriano²⁷, James Di Santo²⁸, Françoise Dromer²⁹, Gérard Eberl³⁰, Jost Enninga³¹, Jacques Fellay³², Antonio Freitas³³, Odile Gelpi³⁴, Ivo Gomperts-Boneca³⁵, Serge Hercberg³⁶, Olivier Lantz³⁷, Claude Leclerc³⁸, Hugo Mouquet³⁹, Etienne Patin⁴⁰, Sandra Pellegrini⁴¹, Stanislas Pol⁴², Lars Rogge⁴³, Anavaj Sakuntabhai⁴⁴, Olivier Schwartz⁴⁵, Benno Schwikowski⁴⁰, Spencer Shorte⁴⁶, Vassili Soumelis⁴⁷, Frédéric Tangy⁴⁸, Eric Tartour⁴⁹, Antoine Toubert⁵⁰, Marie-Noëlle Ungeheuer⁵¹, Lluís Quintana-Murci⁵² & Matthew L. Albert²⁵

¹⁹Laboratory of Human Genetics of Infectious Diseases, Necker Branch, INSERM U1163, Paris, France. ²⁰Institut Pasteur, Department of Immunology, Lymphocyte Cell Biology Unit, Paris, France. ²¹Department of Clinical Pathology and Cytology, Karolinska University Hospital, Stockholm, Sweden. ²²Institut Pasteur, Dynamics of Immune Responses Unit, 75015 Paris, France. ²³Unit of Antibodies in Therapy and Pathology, Department of Immunology, Institut Pasteur, Paris, France. ²⁴Unit for Lymphopoiesis, Immunology Department, Pasteur Institute, Paris, France. ²⁵Laboratory of Dendritic Cell Immunobiology, Department of Immunology, Institut Pasteur, 75015, Paris, France; INSERM U1223, 75015 Paris, France; Center for Translational Research, Institut Pasteur, 75015 Paris, France. ²⁶Immunobiology of Infection Unit, Institut Pasteur, 75015 Paris, France. ²⁷Genome Integrity, Immunity and Cancer Unit, Department of Immunology, Paris, France; Department of Genomes and Genetics, Institut Pasteur, 75015 Paris, France. ²⁸Innate Immunity Unit, Institut Pasteur, INSERM, U1223 Paris, France. ²⁹Institut Pasteur, Department of Mycology, Molecular Mycology-CNRS URA3012, Paris, France. ³⁰Unité Microenvironnement and Immunity, Institut Pasteur, 75724 Paris, France. ³¹Department of Cell Biology and Infection, Institut Pasteur, Paris, France. ³²Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland. ³³Unité de Biologie des Populations Lymphocytaires, Department of Immunology Institut Pasteur, and Centre National pour la Recherche Scientifique, URA1961, 75724 Paris, France. ³⁴Center for Translational Research, Institut Pasteur, Paris, France. ³⁵Institut Pasteur, Unité Biologie et génétique de la paroi bactérienne, Dept. Microbiologie, Paris, France. ³⁶Université Paris 13, Sorbonne Paris Cité, Equipe de Recherche en Épidémiologie Nutritionnelle, Centre de Recherche en Épidémiologies et Biostatistiques, Inserm (U1153), Inra (U1125), Cnam, F-93017 Bobigny, France. ³⁷Laboratoire d’Immunologie clinique, CIC-4218 et Unité INSERM 932, Institut Curie, Paris, France. ³⁸Institut Pasteur, Unité de Régulation Immunitaire et Vaccinologie, Paris, France. ³⁹Laboratory of Humoral Response to Pathogens, Department of Immunology, Institut Pasteur, INSERM U1222, 75015 Paris, France. ⁴⁰Center of Bioinformatics Biostatistics, and Integrative Biology, Institut Pasteur, 75015 Paris, France. ⁴¹Institut Pasteur, Unit of Cytokine Signaling, Paris, France. ⁴²Université Paris Descartes et Département d’hépatologie, Groupe Hospitalier Cochin Hôtel-Dieu, Paris, France. ⁴³Immunoregulation Unit, Institut Pasteur, 75724 Paris, France. ⁴⁴Functional Genetics of Infectious Diseases Unit, Department of Genomes and Genetics, Institut Pasteur, 75015 Paris, France. ⁴⁵Virus & Immunity Unit, Department of Virology, Institut Pasteur, Paris, France. ⁴⁶Institut Pasteur, Imagopole-CITech, 75015 Paris, France. ⁴⁷PSL Research University, INSERM U932, Institut Curie, Paris, France. ⁴⁸Unité de Génomique Virale et Vaccination, Institut Pasteur, CNRS UMR3569, Paris, France. ⁴⁹Department of Immunology, Hôpital Européen Georges Pompidou, Paris, France. ⁵⁰INSERM UMR1160, Université Paris Diderot, AP-HP, Hôpital St Louis, Paris, France. ⁵¹Center for Translational Research, ICAREB Platform, Center for Translational Research, Institut Pasteur, Paris, France. ⁵²Laboratory of Human Evolutionary Genetics, Department of Genomes and Genetics, Institut Pasteur, Paris, 75015, France; CNRS URA3012, Paris 75015, France

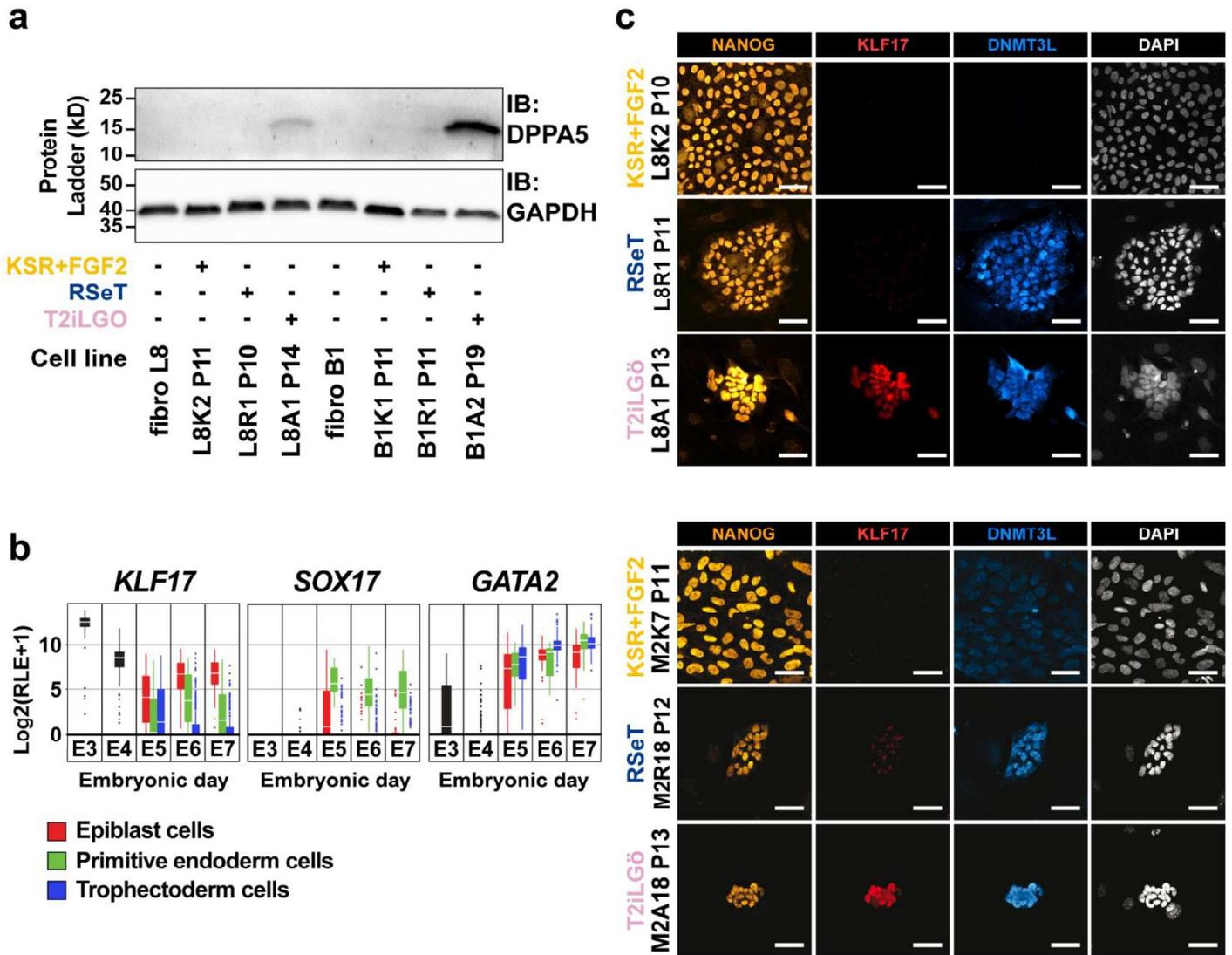


Supplementary Figure 1. **Evaluation of Sendai virus, *DPPA5* and *KLF17* expression across all cell lines**
 Expression levels of Sendai virus, *DPPA5* and *KLF17* measured by qPCR in indicated cell lines at indicated passages and represented in a blue to red relative scale.

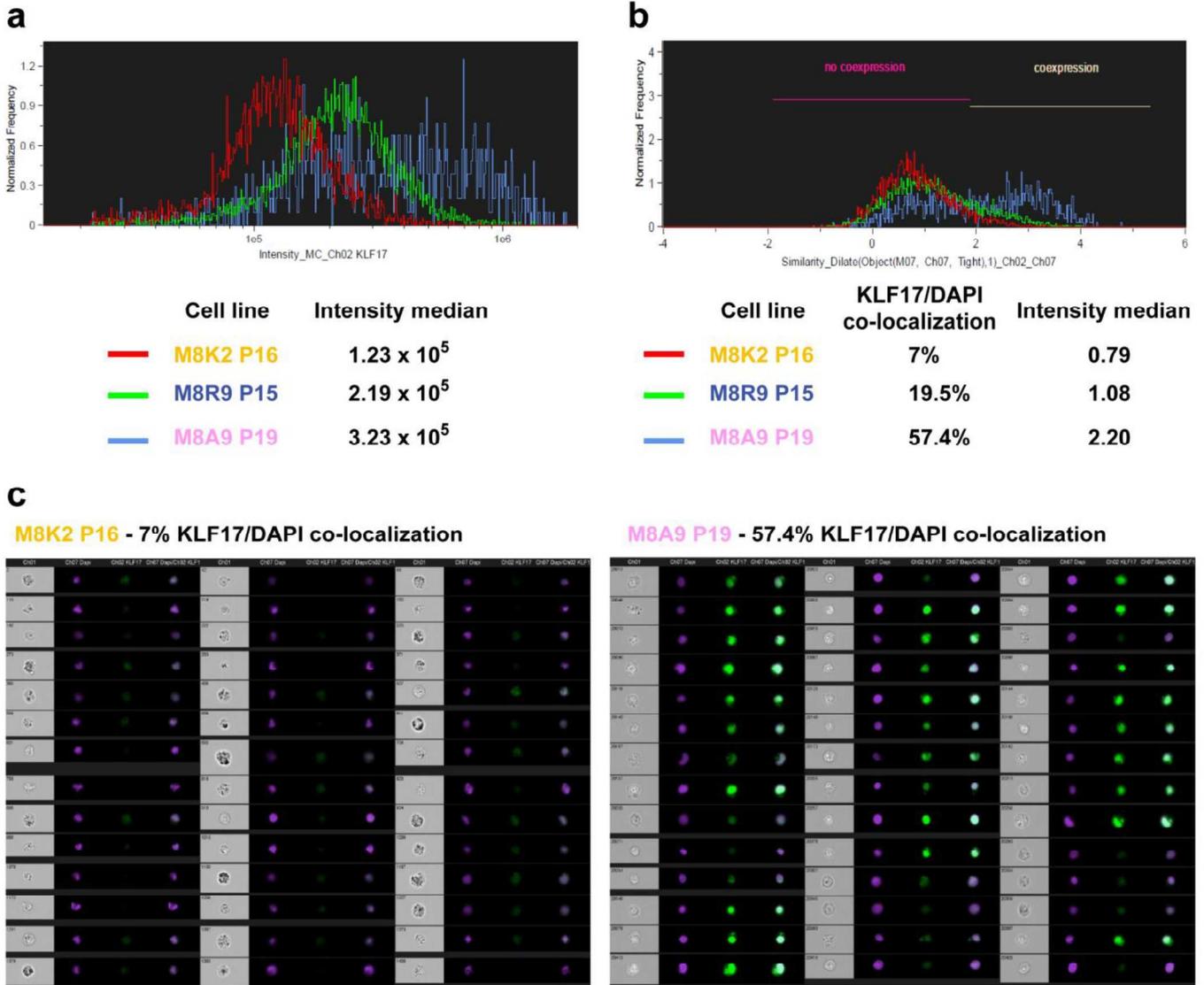


Supplementary Figure 2. Differential expression analysis and functional enrichment of human epiblast cells and hPSCs

(a) Statistically significant differentially expressed genes, as defined in the Differential Expression profiling section of the Methods, between indicated primed and naive cells, from single-cell RNA-seq on the left and DGE-seq on the right, are represented as a heatmap after hierarchical clustering. Venn diagrams show significant overlaps (Two-sided Fisher exact test, $p < 0.01$) in differentially-expressed genes between single-cell and DGE-seq analyses. (b) Expression profile of genes associated with significant dysregulated KEGG pathways (GAGE method, $FDR < 0.01$) related to metabolism (Val, Iso and Leu degradation, purine and pyrimidine metabolism), cell cycle (p53 pathway, cell cycle and apoptosis) and cell-cell junctions (adherent and tight junctions, focal adhesion). For each of the three sets of pathways, expression is represented for hESC and epiblast samples analysed by single cell RNA-seq (left), and for primed or naive hPSCs analysed by bulk DGE-seq (right). Genes are classified by pathway then by fold-change. (c) Projection of single-cell and DGE-seq samples on the first component of 3 principal component analysis. Each PCA was made on genes from a set of putative dysregulated pathway between primed and naive pluripotency. Each component was found statistically linked with other components by a Pearson correlation test ($p < 0.01$).



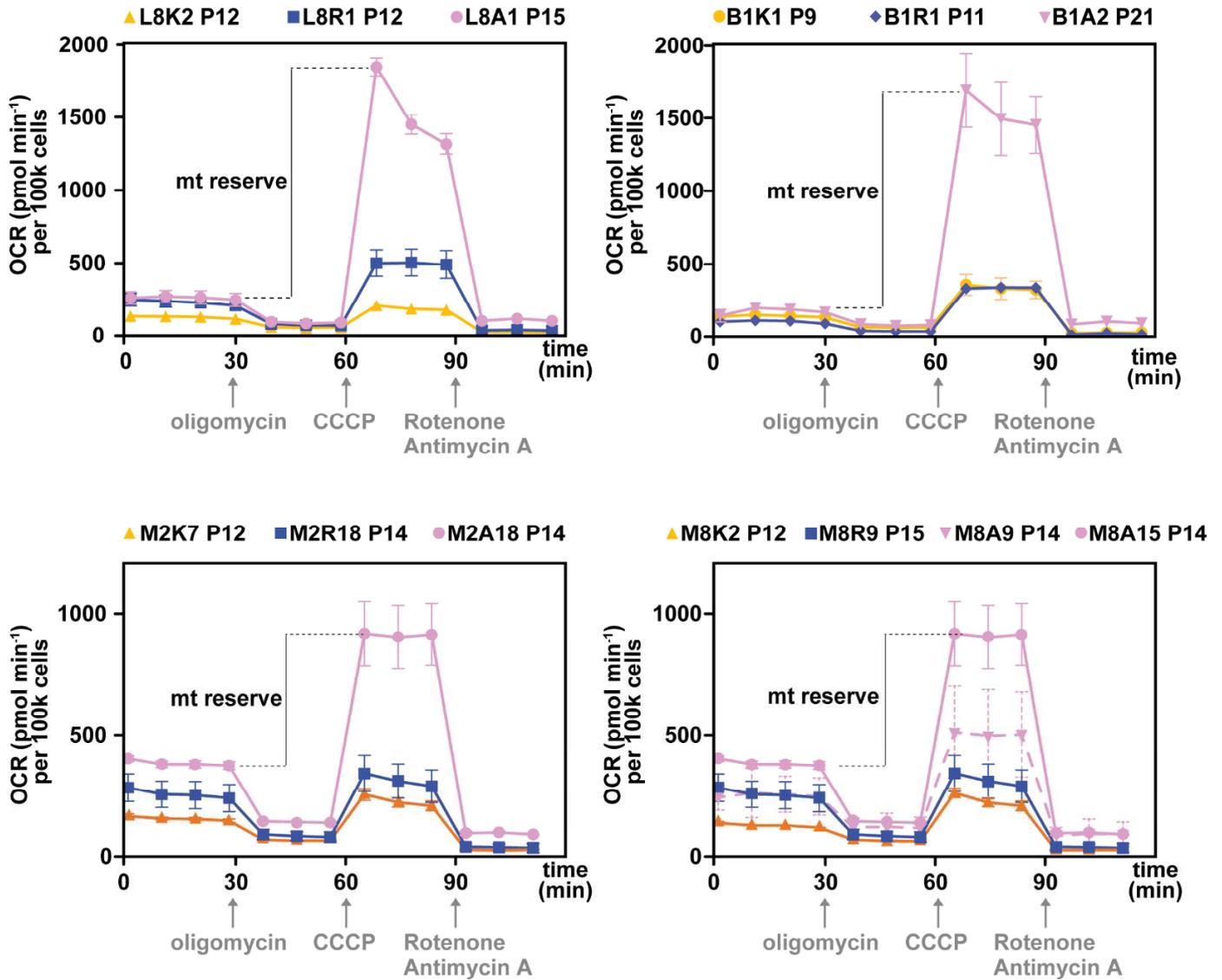
Supplementary Figure 3. **KLF17 and DPPA5 protein expression is restricted to T2iLGö hiNPSCs**
(a) DPPA5 protein is only expressed in T2iLGö hiNPSCs in comparison to their RSeT and KSR+FGF2 counterparts. Indicated cell lines were analyzed by western blot for DPPA5 (expected molecular weight: 14kDa) and GAPDH (expected molecular weight: 36kDa). This western blot represents 2 of 3 biological replicates. **(b)** Expression profile of *KLF17*, *SOX17* and *GATA2* in single cells of human blastocysts by embryonic day, and classified by lineage from day 5 onwards using previously known markers (**Supplementary table 4**). RLE = relative log expression. Error bars are defined as s.e.m. **(c)** Indicated hiPSCs were analyzed by immunofluorescence for NANOG (yellow), KLF17 (red) and DNMT3L (cyan). This figure is representative of 5 biological replicates. Scale bar = 50 μ m.



Supplementary Fig. 4. KLF17 flow imaging confirms intense nuclear expression specifically in T2iLGö hiNPSCs

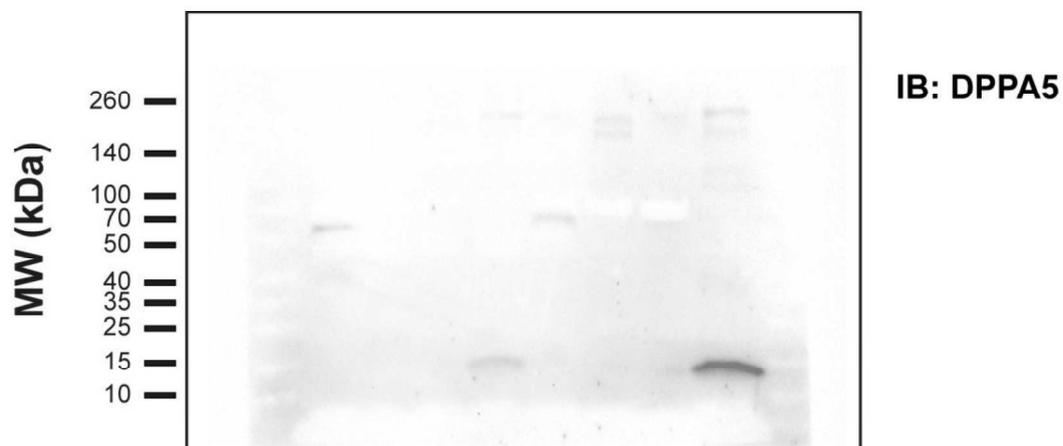
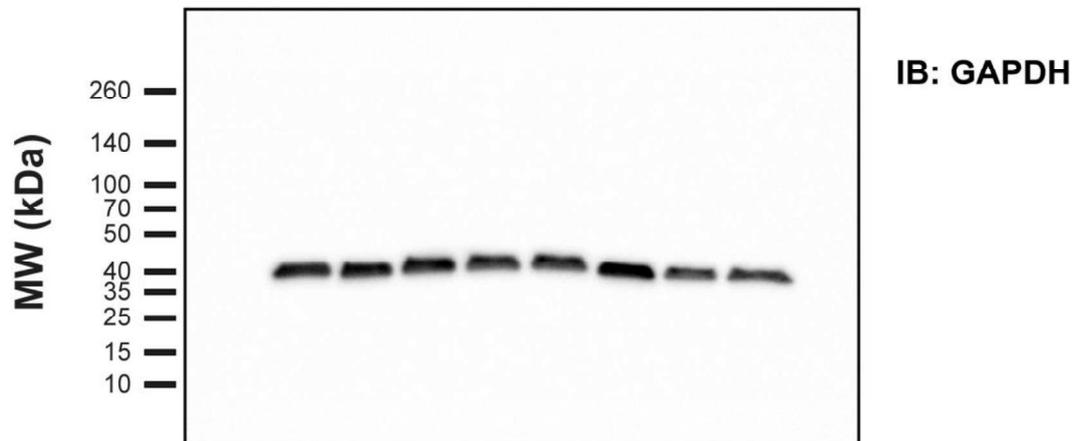
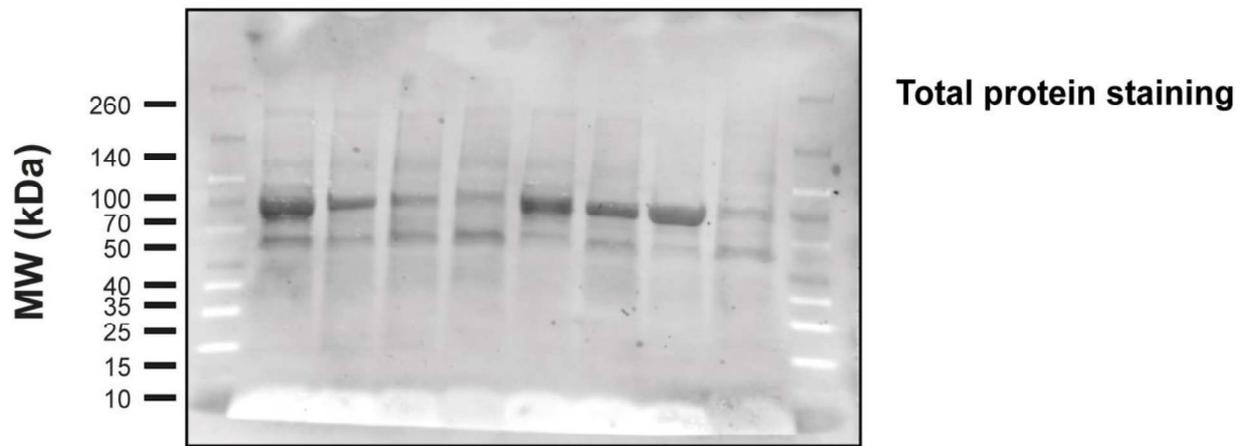
Indicated cells were stained for KLF17 and DAPI, and analysed on an image stream X flow imager. Signal was specifically measured from the nuclear region. (a) Median value of the signal intensity. (b) Percentage of KLF17-DAPI co-localisation and median value of the corrected intensity. (c) Image samples from positive hiPSC in KSR+FGF2 (left) or hiNPSC in T2iLGö (right) showing a striking difference in staining intensity.

Medium: **KSR+FGF2** **RSeT** **T2iLGö**



Supplementary Figure 5. **T2iLGö hiNPSCs possess a higher respiratory capacity than RSeT and primed hiPSCs**

Oxygen consumption rate profiles measured by SeaHorse of indicated cell lines. Oligomycin, CCCP and rotenone/antimycin A were injected at indicated time points to evaluate mitochondrial capacity. This figure represents a biological duplicate. Each point of this figure is a technical triplicate and is representative of 5 independent experiments, with s.d. as error bars.



KSR+FGF2	-	+	-	-	-	+	-	-
RSeT	-	-	+	-	-	-	+	-
T2iLGO	-	-	-	+	-	-	-	+
Cell line	fibro L8	L8K2 P11	L8R1 P10	L8A1 P14	fibro B1	B1K1 P11	B1R1 P11	B1A2 P19

Supplementary Figure 6. **Original Western blot membranes**

From top to bottom images are total protein staining, immunoblot for GAPDH or for DPPA5. For all western blots, Spectra™ Multicolor Broad Range Protein Ladder (Invitrogen) was used.

Supplementary table 1: Overview of cell lines generated in this study starting from 5 fibroblast cell lines

	Cell line	Gender	DGE-seq passage in Fig 1B	Sendai	Karyotypic abnormalities acquired	
mTeSR1	H1	♂	P33*	N/A	-	
	H9	♀	P36*	N/A	-	
	L7m1	♂	P32	N/A	none at P30	
	M2m7	♀	P13	neg at P10	none at P12	
	M8m2	♀	P14	neg at P11	none at P12	
KSR+FGF2	L8K1	♀	P3, P7, P9, P16	neg after P3	none at P20	
	L8K2	♀	P6	pos at P2, neg after P6	none at P9	
	B1K1	♂	P9	neg after P7	none at P7	
	B1K2	♂	P3	neg after P3	-	
	M3K1	♀	P4	neg after P3	-	
RSeT	L8R1	♀	P3, P8, P16, P21	neg after P3	none at P10 and P12	
	L8R2	♀	P8, P11	pos at P2, neg after P8	none at P14 and P16	
	B1R1	♂	P1, P3, P8, P20	pos at P1, neg after P3	none at P12	
	B1R2	♂	P2, P8	low until P20, neg at P22	P18: 40% 47,XY,+7 or +5 50% other trisomies	
	B1R3	♂	P2, P7	pos at P2, neg after P7	-	
	M3R1	♀	P1, P3, P8	pos at P3, neg after P8	-	
	M3R2	♀	P2, P7	pos at P7	-	
	M3R3	♀	P7	pos at P7	-	
	M2R18	♀	P2, P8	neg after P8	none at P12	
	M8R9	♀	P2, P9	neg after P9	none at P13	hypoxia
T2iLGo	L8A1	♀	P7, P9, P11, P15, P20	pos until P9, neg after P11	P14: 50% 92,XXXX	
	L8A2	♀	P9	neg after P9	P13: 12% 92,XXXX; 88% 69,XXX	
	B1A1	♂	-	pos at P17	P16: 100% 92,XXYY	
	B1A2	♂	P11, P13, P17	pos at P21	none at P15	
	M3A1	♀	P20, P25	pos at P25	-	
	M2A18	♀	P8, P11, P14	neg after P8	P14: 67% 92, XXXX	
	M8A9	♀	P2, P9, P15	neg after P15	P13: 25% 46XX, t(1,12) 17% 92, XXXX	hypoxia
	M8A15	♀	P6, P10, P15	neg after P15	P15: 64% 47, XX, +7; 14% 92, XXXX	
	HNES1	♂	P40	N/A	none at P42	

*also in single-cell RNA-seq

N/A stands for not applicable.

A dash means we do not have the data for the particular passage or cell line.

Supplementary table 2: **Antibodies used in this study**

ANTIBODY	SUPPLIER	IDENTIFIER	DILUTION USED
Anti-H3K27me3 mouse monoclonal	Abcam	Cat# ab6002, RRID:AB_305237	1:200
Anti-SOX17 goat polyclonal	R and D Systems	Cat# AF1924, RRID:AB_355060	1:200
Anti-NANOG goat polyclonal	R and D Systems	Cat# AF1997, RRID:AB_355097	1:200
Anti-KLF17 rabbit polyclonal	Sigma-Aldrich	Cat# HPA024629, RRID:AB_1848933	1:500 1:100 for ISX
Anti-GATA2 mouse monoclonal	Sigma-Aldrich	Cat# WH0002624M1, RRID:AB_1841726	1:50
Anti-DNMT3L mouse monoclonal	Abcam	Cat# ab93613, RRID:AB_10562109	1:100
Anti-DPPA5 goat polyclonal	R and D Systems	Cat# AF3125, RRID:AB_2094168	1:3000
Anti-GAPDH mouse monoclonal	Santa Cruz Biotechnology	Cat# sc-32233, RRID:AB_627679	1:1000
Zombie NIR™ Fixable Viability Kit	BioLegend	Cat# 423105	1:500
Anti-rabbit Alexa 488	Thermo Fisher Scientific	Cat# A-21206, RRID:AB_2535792	1:1000
Anti-mouse Alexa 568	Thermo Fisher Scientific	Cat# A10037, RRID:AB_2534013	1:1000
Anti-goat Alexa 647	Thermo Fisher Scientific	Cat# A21447, RRID:AB_10584487	1:1000
Anti-rabbit Alexa 488	Thermo Fisher Scientific	Cat# A-11034 RRID:AB_10562715	1:500 for ISX
Anti-goat HRP	Santa Cruz Biotechnology	Cat# sc-2922, RRID:AB_656965	1:5000
Anti-mouse HRP	Santa Cruz Biotechnology	Cat# sc-2055, RRID:AB_631738	1:5000

Antibodies either identified by the catalog number of the supplier or by their Research Resource Identifiers (RRIDs)

ISX stands for Image Stream X

Supplementary table 3: **RT-qPCR Primers used in this study**

Gene Name	Primer sequence 5'-3'	Amplicon size (bp)	Melting Temp. (°C)	Position
GAPDH	AATCCCATCACCATCTTCCA TGGACTCCACGACGTA	82	80.5	494-576
KLF17	TCAGGAAGGGACTGGTAGAA GTACCCGCATATGTCGTCTAAG	206	83	862-1067
DPPA5	TCCCGAAGACCTGAAAGATCCAGA AATAGGAGCCGTAAACCACGACCT	177	83.5	99-275
SeV	GGATCACTAGGTGATATCGAGC ACCAGACAAGAGTTTAAGAGATATGTATC	181	79	N.A.
NANOG	ATAGCAATGGTGTGACGCAGAAGG CTGGTTGCTCCACATTGGAAGGTT	116	82	701-816

Supplementary Table 4: **Lineage specific makers used to distinguish trophectoderm, epiblast and primitive endoderm cells**

Gene Symbol	Lineage
CDX2	Trophectoderm
CLDN10	Trophectoderm
GATA2	Trophectoderm
GATA3	Trophectoderm
TGFBR3	Trophectoderm
KRT18	Trophectoderm
KRT8	Trophectoderm
EFNA1	Trophectoderm
ARGFX	Epiblast
BMP2	Epiblast
DPPA2	Epiblast
DPPA5	Epiblast
ETV4	Epiblast
FGF4	Epiblast
FOXD3	Epiblast
GDF3	Epiblast
IL6R	Epiblast
KLF17	Epiblast
LEFTY	Epiblast
NANOG	Epiblast
NODAL	Epiblast
OTX2	Epiblast
POU5F1	Epiblast
PRDM14	Epiblast
SOX2	Epiblast
TDGF1	Epiblast
VENTX	Epiblast
ZIC3	Epiblast
GATA6	Primitive endoderm
FOXA2	Primitive endoderm
PDGFRA	Primitive endoderm
BMP6	Primitive endoderm
GATA4	Primitive endoderm
COL4A1	Primitive endoderm

Supplementary table 5: **Contingency table of differentially expressed genes**

Differentially Expressed :	In Single-cell RNA-seq	Not in single-cell RNA-seq
In DGE-seq	1980	1023
Not in DGE-seq	4648	10286

Supplementary table 6: **Sampling size for indicated cell types**

Dataset	Group	Sampling size
Single-Cell	hESC	52
	Epiblast	52
	Morula	52
DGE-seq	T2iLGö (+HNES)	26
	KSR+FGF2	9
	mTeSR1	5
	RSet	26
	E7	8
	Fibroblast	4

2.3 Discussion

L'analyse du DGE-Seq seul a été effectuée selon une méthodologie d'analyse classique de bulk RNA-Seq. L'analyse de ces données a ainsi marqué le début de l'écriture de mon pipeline d'analyse de bulk RNA-Seq (<https://gitlab.univ-nantes.fr/E114424Z/BulkRNAseq>). Le séquençage avait été effectué en 3 plaques (3 *runs*) mais l'existence de conditions redondante et de répliques techniques entre les plaques a considérablement facilité la mise en place d'une correction batch par simple régression linéaire. Après cette étape, une analyse par composante principale (Kilens, Meistermann et al. Figure 1c) permet de distinguer un axe majeur entre des cellules non reprogrammées (fibroblastes, E7) et des cellules pluripotentes. Le deuxième axe de variation est marqué par un continuum entre les hNESC (Guo et al., 2016) et les hESC classiques (Thomson et al., 1998), et donc, entre la pluripotence naïve et amorcée. Cette ACP montre que les cellules placées en milieu t2iLGö lors de la reprogrammation ont un profil transcriptomique de pluripotence naïve et que celles placées en milieu RSet ne sont pas stables et possèdent des caractéristiques plus ou moins naïves et amorcées selon la lignée et le nombre de passage (la maturité) de ces lignées. Dans la même période que la publication de ce travail, il a été montré que des milieux de cultures autre que t2iLGö permettent de capturer une pluripotence naïve stable tel que 5iLAF (Liu et al., 2017).

L'aspect le plus complexe du travail d'analyse fut l'intégration des données de scRNAseq aux données de DGE-Seq. La comparaison directe des données par ACP (Kilens, Meistermann et al. Figure 1d) a été rendu possible après de nombreuses étapes de transformation de données, dont certaines importantes comme un centrage/réduction et une normalisation par quartile. La présence de cellules souches pluripotentes amorcées à la fois dans le jeu de données single et bulk a toutefois permis de désigner ces transformations et d'en assurer la validité. Pour l'instant il n'existe pas de standard pour comparer bulk et scRNASeq, ce qui représente un véritable handicap dans la constitution de résultats robustes et reproductibles impliquant de tels types de comparaison. Le résultat montre que les cellules naïves *in vitro* sont les plus proches de l'épiblaste préimplantatoire et donc un bon modèle de celui-ci. Les cellules de Morula sont à l'extérieur du continuum pluripotent, bien que les cellules de Morula expriment certains marqueurs de

pluripotence naïve comme *ARGFX* et *KLF17* (Kilens, Meistermann et al. Figure 2a). Cette constatation est un indice de la non-existence de cellules pluripotentes naïves dans la Morula. Cet aspect sera approfondi dans le second manuscrit présenté dans cette thèse. Une autre façon d'intégrer les données de single-cell et bulk est indirecte. Nous avons comparé les résultats des analyses de chaque jeu de données. Le cas le plus simple est la comparaison des listes de gènes différentiellement exprimés hNESC-hNIPSC versus hESC/hIPSC (DGE-Seq) et épiblaste préimplantatoire versus hIPSC/hESC (single-cell). Les listes de gènes différentiellement exprimés sont significativement associées et sont donc un proxy montrant la parenté transcriptomique des hNIPSC et de l'épiblaste (Kilens, Meistermann et al. Supplementary Figure 2a).

2.4 Conclusion

Nous avons montré qu'il était possible de capturer l'état naïf par reprogrammation OSKM et ainsi produire des lignées de hNIPSC. Ce travail a aussi permis de mieux caractériser les différences entre pluripotence naïve et amorcée. L'analyse du DGE-Seq par expression différentielle a mis en évidence l'importance de marqueurs propres à chaque type de pluripotence, tel que des facteurs de transcriptions clés (Kilens, Meistermann et al. Figure 2a). Parmi les gènes différentiellement exprimés, nous retrouvons les gènes *TET2* et *DNMT3L* contrôlant la méthylation de l'ADN (Kilens, Meistermann et al. Figure 4b). Il apparaît que le génome des cellules naïves est en effet hypométhylé par rapport aux cellules amorcées (Kilens, Meistermann et al. Figure 4a)

L'enrichissement fonctionnel a fait ressortir la phosphorylation oxydative comme significativement dérégulée entre pluripotence naïve et amorcée (Figure 3a), tout comme d'autres fonctions métaboliques (Kilens, Meistermann et al. Supplementary Figure 2b-c). L'hypothèse d'une respiration cellulaire différente entre pluripotence naïves et amorcée soulevée par l'analyse du DGE-Seq a été confirmée par mesure de l'activité des mitochondries par *SeaHorse* (Kilens, Meistermann et al. Figure 3b). Ce résultat montre l'importance d'utiliser des données OMIC dans un premier temps pour générer des hypothèses qui détermineront les expériences à effectuer pour valider ces mêmes hypothèses.

2.5 Perspectives

L'obtention des hNIPSC permet d'avoir un modèle de l'épiblaste préimplantatoire humain *in vitro* en variant facilement les fonds génétiques et ainsi étudier de façon robuste des mécanismes tels que l'inactivation du X. Ces cellules ouvrent aussi la voie à la constitution de modèle *in vitro* de blastocystes chez l'Homme. Ces modèles sont constitués d'un assemblage de cellules pluripotentes naïves au centre, entourées de cellules souches trophoblastiques et appelés blastoïdes (Rivron et al., 2018b). Des blastoïdes ont été obtenus chez la souris. Chez l'Homme, l'obstacle majeur à l'obtention de blastoïdes en masse est la génération de cellules souches trophoblastiques, induites ou non, proches du trophoctoderme préimplantatoire. Cette quête est devenue un but majeur de notre domaine de recherche en général et de notre équipe en particulier, et de nettes avancées ont déjà été effectuées. Une des plus importante est l'obtention de cellules souches trophoblastiques humaine (hTSC) (Okoe et al., 2018). L'étude de ces deux modèles *in vitro* permettra de cerner les limites de la plasticité entre épiblaste naïf et trophoctoderme. Une publication qui concerne cette plasticité est en cours de révision dans l'équipe. Nous montrons dans ce travail que les cellules pluripotentes naïves peuvent être converties en cellules souches trophoblastique identiques à celles obtenus par l'équipe d'Okoe. L'inverse ne semble pas possible. Ainsi, de la même façon qu'il existe un état de pluripotence naïve préimplantatoire et amorcée postimplantatoire, il pourrait exister des cellules trophoblastiques naïves et amorcées. Les hTSC obtenues par Okoe seraient donc des cellules trophoblastiques postimplantatoire ayant perdu leur plénipotence.

Les apports de cet article concernant la compréhension de la pluripotence naïve ouvrent la voie à la constitution de modèles plus complexes. Ces modèles pourraient décrire les causes et les conséquences des différences métaboliques entre la pluripotence naïve et amorcée. Une approche courante pour modéliser le métabolisme est la *Flux Balance Analysis* (Fell and Small, 1986). La création d'un tel modèle métabolique serait un atout dans notre compréhension profonde de la pluripotence. Une telle compréhension pourrait avoir des répercussions directes dans tous les domaines utilisant des cellules souches pluripotentes. En premier lieu le développement humain, mais aussi la médecine régénérative.

3 Manuscrit #2 : Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophectoderm specification

3.1 Contexte

3.1.1 Le développement préimplantatoire humain par le prisme du scRNA-Seq

La première étude scRNA-Seq d'embryons humains préimplantatoires a été publiée en 2013 (Yan et al., 2013). Ce jeu de données contient 124 cellules provenant d'hESC (cf. 2.1, en page 70) d'ovocytes et d'embryon préimplantatoires (90 cellules). L'étude a permis de confirmer le fait que les hESC ne sont pas dans le même état de pluripotence que l'épiblaste préimplantatoire. Ce travail nous a aussi permis d'avoir un premier aperçu des variations transcriptomiques au sein du développement préimplantatoire, notamment comme on pouvait s'y attendre, il existe d'important changement lors de la ZGA. Les 3 blastocystes séquencés sont au stade de l'éclosion (**Figure 1a**). Les 30 cellules provenant de ces blastocystes ont été classées en EPI, PrE et TE suivant une liste de marqueurs établie dans des études utilisant le profilage d'expression par puce à ADN (Assou et al., 2012; Galán et al., 2010). Ce classement a été comparé à un clustering ascendant hiérarchique (CAH) en utilisant tous les gènes comme contrôle. Les auteurs disposent de l'information de provenance du trophectoderme (côté mural et polaire), et notent que la CAH discrimine les deux pôles. Ce travail a également permis de constituer des listes de centaines de marqueurs associées aux lignées EPI, PrE et TE.

En 2015 une publication (Blakeley et al., 2015) réanalyse les premiers jeux de données de scRNA-Seq du développement préimplantatoire humain (Yan et al., 2013) et murin (Deng et al., 2014). Le jeu de données est complété avec 30 nouvelles cellules de blastocyste séquencées en scRNA-Seq. Les auteurs notent tout d'abord que l'ACP de Yan et al. n'est pas reproductible, et effectuent ensuite une CAH pour

partitionner les cellules en TE / EPI / PrE. Des différences significatives entre le développement humain et murin sont mises en évidence :

- l'absence ou la faible expression de *ID2*, *CDX2*, *EOMES*, *ELF5* chez l'Homme, alors que ses orthologues chez la souris ont été montrés comme marqueurs du TE ;
- la non-spécificité au TE de *TFAP2C* chez l'Homme, qui est exprimé dans tout le blastocyste. Alors que l'induction de *Tcfap2c* permet de dériver des cellules souches trophoblastiques chez la souris (Kuckenberget al., 2010) et est indispensable à la mise en place et au maintien du TE (Auman et al., 2002).

Cette étude a aussi permis de mettre en évidence de nouveaux marqueurs des lignées EPI et TE. En dehors de l'analyse des données de scRNA-Seq, la voie de signalisation TGF- β a été bloquée dans des cellules de l'EPI. Le résultat a été la diminution de l'expression de NANOG et OCT4, des facteurs clef de la pluripotence. La voie TGF- β semble donc être nécessaire au maintien de la pluripotence chez l'Homme. L'enrichissement fonctionnel montre des différences dans les gènes impliqués dans la voie TGF- β entre l'EPI et le TE. De façon intéressante la stimulation par BMP4 (activateur de la voie BMP, sous voie de TGF- β) d'IPS humaines induit une différenciation trophoblastique (Li and Parast, 2014).

Le plus volumineux jeu de données de scRNA-Seq préimplantatoire humain contient 1529 cellules (Petropoulos et al., 2016) et couvre le stade 8 cellules jusqu'au blastocyste prêt à l'implantation. Une t-SNE des données montre un continuum entre les cellules de blastocystes. Cela montre qu'il est difficile d'établir des clusters sur des bases objectives dans le contexte du développement préimplantatoire humain. J'ai dû moi-même prendre en compte cet aspect des données dans mon analyse. Une approche de pseudo-temps a été employée pour différencier l'EPI, le TE et le PrE, à partir d'une réduction de dimension *diffusion map* (Haghverdi et al., 2015). Les auteurs concluent que les lignées TE / EPI / PrE se spécifient simultanément à jour 5, et notent qu'il existe deux populations de TE qu'ils attribuent au TE mural et polaire. Leur analyse montre que la morula semble être plus proche de l'ICM que du TE. Un axe majeur de l'article est l'étude de la compensation de dosage du X (Lyon, 1961). En effet, l'expression des gènes du

chromosome X doit être équivalente entre les cellules mâles et femelles dans l'embryon préimplantatoire sous peine de mort de l'embryon (Goto and Takagi, 1998). Or, il y a deux fois plus de matériel génétique du X chez les cellules femelles (XX) par rapport aux cellules mâles (XY). Chez la souris, un des deux X est inactivé de façon transitoire pour garantir cette compensation, puis est réactivé avant l'implantation dans l'épiblaste (Heard et al., 2004). Un des résultats de cette étude est que l'expression reste biallélique dans toutes les cellules mais diminue sur chaque allèle en comparaison des cellules mâles. Cet article marque un pas en avant dans deux axes : la quantité de cellules et la qualité des outils statistiques employés. Cependant, il me semble que la normalisation RPKM utilisée pour les analyses peut induire des biais problématiques dans l'analyse.

Le premier KO par CRISPR-Cas9 sur embryon humain cible OCT4, gène clef de la pluripotence (Fogarty et al., 2017). Les embryons KO ont beaucoup de mal à survivre jusqu'au stade blastocyste. L'intérêt de l'étude est surtout de montrer la faisabilité d'un KO d'embryon humain, à des fins de recherche. 80 cellules ont été séquencées en scRNA-Seq pour cet article, provenant d'embryons KO ou contrôles.

L'équipe de Paul Bertone publie en 2018 deux études qui réanalysent les données de Petropoulos et al. La première (Stirparo et al., 2018) utilise une liste restreinte de marqueur pour discriminer les lignées TE / EPI / PrE, face à la difficulté de partitionner les données par des analyses non supervisées. Les auteurs font aussi le choix de se débarrasser du TE pour une partie des analyses, afin de réduire le problème de la surreprésentation de la lignée TE dans les données de blastocyste. Des ensembles de marqueurs sont déterminés, notamment par WGCNA. Une des conclusions de l'article est que la spécification EPI / PrE est plus tardive que la spécification ICM / TE, contrairement à la conclusion de Petropoulos et al. Les auteurs font l'hypothèse que les embryons à jour 5 peuvent en fait être à des stades de développement différents. En effet tous les embryons humains ne se développent pas à la même vitesse, surtout à l'échelle du développement préimplantatoire où l'embryon change d'heure en heure. La deuxième étude (Boroviak et al., 2018) s'intéresse aux différences entre l'Homme, la souris, et un primate de la famille des ouistitis, le marmoset.

Globalement les données de scRNA-Seq de développement préimplantatoire humain ont rarement été analysées avec des outils spécifiques au single-cell. Par exemple, seule une étude a utilisé une normalisation spécifique au single-cell (Fogarty et al., 2017). La raison la plus évidente est que dans la majorité des cas ces outils n'étaient pas encore disponibles lors de l'écriture de ces articles. En conséquence, les différentes analyses sont peu reproductibles car tributaires des variations indésirables dans les données. Un autre problème est que les analyses effectuées dans ces articles sont souvent restreintes et ne permettent pas d'avoir une vision d'ensemble intuitive du développement préimplantatoire humain.

3.1.2 Motivation de l'étude

La principale motivation pour ce travail est de comprendre la hiérarchie des événements du développement préimplantatoire humain, en utilisant au maximum le potentiel du scRNA-Seq. Dans ce but, nous avons effectué une analyse de quatre jeux de données de scRNA-Seq. Trois d'entre eux proviennent de la littérature (Fogarty et al., 2017; Petropoulos et al., 2016; Yan et al., 2013). Le quatrième jeu de données a été construit pour cet article et provient d'embryone donnée à la recherche par l'intermédiaire du centre d'assistance médical à la procréation de Nantes. L'intérêt de ce jeu de données réside dans l'annotation des cellules. Pour chacune d'entre elles nous avons la vidéo du développement de l'embryon d'où elle provient, grâce à un embryoscope. Ces vidéos ont permis à des cliniciens d'annoter précisément les stades de développement de chaque embryon. Enfin, si la cellule provient du blastocyste, nous avons le pôle de provenance de celle-ci (**Meistermann, Loubersac et al. Figure 3a**). Cette information permet de connaître de façon certaine le type cellulaire des cellules du pôle mural, en effet celui-ci ne contient que du trophoctoderme.

Le cœur de notre analyse du scRNA-Seq se base sur l'utilisation complémentaire de modélisation par pseudo-temps avec Monocle2 (Qiu et al., 2017), de reconstruction des modules de gènes coexprimés avec WGCNA (Langfelder and Horvath, 2008), et d'analyse des dynamiques de transcription avec Velocity (La Manno et al., 2018). Nous avons réalisé des immunofluorescences pour valider les aspects spatio-temporels découverts lors de notre analyse des données transcriptomiques.

Un élément qui a pris de plus en plus de place dans notre article au fur et à mesure de sa conception est la comparaison de nos résultats aux données murines. Cette comparaison nous permet en effet de montrer que notre workflow d'analyse donne des résultats cohérents avec les connaissances sur le développement préimplantatoire murin. Cette comparaison agit donc comme contrôle des nouveautés trouvées par l'analyse des données humaines. Le deuxième intérêt est de pouvoir comparer facilement les trajectoires de destin cellulaire des deux espèces. Le plus gros jeu de données de scRNA-Seq de développement préimplantatoire murin compte 262 cellules (Posfai et al., 2017). Il existe donc une véritable différence de perception du développement préimplantatoire entre l'Homme et la souris en termes de données disponibles. Cette différence de vision, mécanistique chez la souris et transcriptomique chez l'Homme, pourrait être à l'origine de la difficulté à comparer les événements préimplantatoires de l'Homme et de la souris. Cette problématique nous a poussé à ré-analyser le jeu de données de Posfai avec le même pipeline d'analyse que celui utilisé chez l'Homme.

3.2 Manuscrit

Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophectoderm specification

Dimitri Meistermann^{1,2,4,*}, Sophie Loubersac^{1,3,*}, Arnaud Reignier^{1,3}, Alexandre Bruneau^{1,2}, Julie Firmin^{1,3}, Valentin Francois - - Champion^{1,2}, Stéphanie Kilens^{1,2}, Yohann Lelièvre⁴, Jenna Lammers³, Magalie Feyeux^{1,5}, Phillipe Hulin⁵, Steven Nedellec⁵, Betty Bretin^{1,2}, Gaël Castel^{1,2}, Simon Covin^{1,2}, Audrey Bihouée⁶, Magali Soumillon^{7,8}, Tarjei Mikkelsen^{7,9}, Paul Barrière^{1,3}, Jérémie Bourdon^{4,6}, Thomas Fréour^{1,3,†}, Laurent David^{1,2,5,†}.

1. CRTI, UNIV Nantes, INSERM, Nantes, France;
2. ITUN, CHU Nantes, Nantes, France;
3. Service de Biologie de la Reproduction, CHU Nantes, Nantes, France;
4. LS2N, UNIV Nantes, CNRS, Nantes, France;
5. SFR-SANTE, UNIV Nantes, INSERM, CNRS, CHU Nantes, Nantes, France;
6. Institut du thorax, UNIV Nantes, INSERM, CNRS, Nantes, France;
7. Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA; Broad Institute, Cambridge, MA 02142, USA; Harvard Stem Cell Institute, Harvard University, Cambridge, MA 02138, USA;
8. Present address: Torpedo Diagnostics, Inc. LabCentral, 700 Main Street, Cambridge, MA 02139, USA;
9. Present address: 10x Genomics, 7068 Koll Center Pkwy #401, Pleasanton, CA 94566, USA.

***Authors contributed equally**

†co-corresponding authors

Correspondence to:

L. David (laurent.david@univ-nantes.fr), T. Fréour (thomas.freour@chu-nantes.fr).

Lead contact:

L. David

Highlights

- Mouse and human preimplantation development reconstructed by a pseudotime model from scRNAseq
- Mouse and human preimplantation models are browsable through a quick and efficient user interface
- Precise gene expression hierarchy provides molecular markers of development stage
- Polar TE develops faster than mural TE

Keywords

Epiblast ; Trophectoderm ; Pluripotency ; cell fate ; Lineage specification ; Blastocyst ; Preimplantation development ; scRNA-seq ; Pseudotime ; Human and mouse development

Summary

Recent technological advances such as single-cell (sc) RNAseq and CRISPR-CAS9-mediated knock-out have allowed an unprecedented access to processes orchestrating human preimplantation development. However, the sequence of events which occur during human preimplantation development are still unknown. In particular, the timing of the first human lineage specification, the process by which the morula cells acquire a specific fate, remains elusive. Here, we present a mouse and human preimplantation development model based on transcriptomic pseudotime modelling of scRNAseq biologically validated by spatial information and precise time-lapse staging. We show that in human, trophectoderm (TE) and inner cell mass (ICM) specification is not complete until the B3 blastocyst stage. This reveals a delay with the mouse first specification. We identified novel markers enabling precise staging of human preimplantation embryos: IFI16, associated to the developed epiblast (EPI) and NR2F2, associated with TE development. Strikingly, NR2F2 positive TE cells arise from the polar side, shortly after specification, supporting a model of faster polar TE development leading to polar implantation in the endometrium. Altogether, our study unravels preimplantation lineage specification in human embryos and provides a browsable resource for mapping spatio-temporal events orchestrating human preimplantation development.

Introduction

Shortly after fertilization, human embryos establish three lineages bearing specific features, essential for the development of the fetus: 1/ the trophectoderm (TE), at the periphery of the blastocyst, that regulates implantation and generates the placenta, 2/ the epiblast (EPI) is pluripotent and will give rise to the fetus, 3/ the primitive endoderm (PrE) will contribute to the yolk sac and other extraembryonic annexes (Rossant and Tam, 2017). Lineage specification has mostly been studied in mouse, using a combination of genetic manipulation, lineage tracing and immunofluorescence, leading to a two-steps model of mammalian lineage specification.

In mouse, the first specification, segregating TE from ICM, occurs at the morula stage and is driven by a YAP/TEAD4/CDX2 axis (Chazaud and Yamanaka, 2016; Rossant and Tam, 2017). The second specification, separating ICM cells into EPI and PrE, occurs at the blastocyst stage, which is driven by a NANOG/FGF4/GATA6 axis (Chazaud and Yamanaka, 2016; Rossant and Tam, 2017). However, differences between mouse and human development are accumulating: in human, CDX2 expression appears after specification of the TE (Niakan and Eggan, 2013), GATA6 is restricted to PrE cells upon implantation (Deglincerti et al., 2016) and FGF signaling did not seem to regulate EPI/PrE specification (Kuijk et al., 2012; Roode et al., 2012).

Human development has been difficult to study due to limited access to biological samples. The advent of scRNAseq methods has dramatically changed our ability to understand human preimplantation development (Blakeley et al., 2015; Petropoulos et al., 2016; Yan et al., 2013). However, scRNAseq remains challenging and requires biological references to validate models. In absence of a continuous transcriptomic model of mouse preimplantation development, human scRNAseq datasets are compared with functional data obtained in mouse preimplantation development studies, with discrepancies in conclusions influenced by the technic used (**Figure S1A**).

The principal question of interest that has been studied in human preimplantation development was to understand EPI and PrE segregation. Few TE markers have been discovered: CDX2 in expanded blastocysts; GATA3, LAMA1, LAMA3 and KRT7 in hatched/post-implantation blastocysts; GATA2 in

blastocysts after B2/B3 stage. Moreover, the molecular events that enable competent human blastocysts to interact with receptive luminal endometrium have remained largely elusive. An important observation has been that human blastocysts attach to the endometrial surface epithelium from the polar side (Aberkane et al., 2018; Grewal et al., 2008; Lindenberg, 1991). This suggests specific molecular events occurring on the polar TE side of the blastocyst, however neither the events nor the underlying mechanisms have been proposed.

Using scRNAseq technology coupled to fine annotation of embryos, we now provide a human comprehensive resource, characterizing the transcriptional dynamics of lineage specification. Validation of our transcriptomic model unveiled human EPI specific events. We also propose some mechanistic insight in TE maturation towards implantation.

Results

Generation of a continuous transcriptomic models of mouse and human preimplantation development

Much knowledge has been gained on mouse preimplantation development through genetic studies, immunofluorescence analysis and treatment or inhibition of signaling pathways. However, scRNAseq datasets are generally not providing samples from all lineages at multiple developmental stages. One dataset offers diversity of cells and stage, from early 16-cells morula to 64 cell stage (Posfai et al., 2017) (**Figure 1A**). In order to have a view of how the transcriptomic profile of cells during mouse preimplantation development evolves, we reanalyzed the dataset from Posfai and colleagues by computing a pseudotime model with reversed-graph embedded projection to generate cell fate trajectories: the likeliest path that temporally orders cells based on their transcriptome (Qiu et al., 2017) (**Figure 1B** and **Figure S1C**). The mouse pseudotime model shows separation in two branches: one identified as ICM expressing *Sox2*; the other one identified as TE expressing *Cdx2* (**Figure 1C**) (Frum et al., 2018).

We applied this analysis strategy to human embryo scRNAseq datasets (**Figure 1A**). Importantly, we included a novel dataset that we generated with precise time-lapse annotation of embryos before

scRNAseq analysis. We used embryos donated to research, and not poor quality embryos excluded from IVF procedure or that have undergone preimplantation genetic diagnostic. Our precise annotation is particularly important since the main difficulty impairing the analysis of public scRNA-seq datasets of human embryos is the variability of *in vitro* culture leading to potential miss-annotation. For *in vitro* fertilization (IVF) cycles, morphology-based staging are preferred to time, especially as multiple morphological events occur during the fifth day post-fertilization (dpf) (Alpha Scientists in Reproductive and Embryology, 2011) (**Figure S1D**). Time-lapse analysis coupled with the pseudotime allowed us to link morphology-based stage and transcriptomic events. Individual morula cells were recovered after incubation of morula with decompaction media; blastocysts were laser-dissected to separate the mural TE from the polar TE/ICM (**Figure S2A-C**). Our set of 24 embryos and 150 cells were pooled with available datasets, totaling 128 embryos and 1751 cells. Normalization and batch correction were applied (Lun et al., 2016) and a pseudotime model was generated, with all 4 datasets interspersed (**Figure S2D-F**). The resulting human pseudotime model is characterized by 5 branches, each branch containing samples from at least 2 datasets, and 2 branching points; pseudotime value indicates the amount of transcriptomic changes from the root cell (**Figure 1D**). Projection of the expression levels of transcription factors that we previously analyzed by immunofluorescence allowed to identify the EPI, PrE and TE branches of the blastocyst, respectively positive for KLF17 (EPI), SOX17 (PrE) and GATA2 (TE) (**Figure 1E**) (Kilens et al., 2018).

We compared lineage classification by Petropoulos et al., Stirparo et al. and our annotation based on the pseudotime. The best overlap between analysis strategies is the determination of TE cells (**Figure S3A-D**). The main discrepancy is identification of an ICM: Stirparo et al. arbitrarily decided of an ICM (5 dpf, not TE cells), while we did only observe 41 cells that are fitting between the main branch and the EPI/PrE branch. Finally, the PrE identification is rather variable between studies, suggesting that PrE fate might not be well established in human preimplantation embryos. Finally, we projected cells from OCT4 KO human embryos (Fogarty et al., 2017). Only 2/17 OCT4 KO embryos passed the ICM/TE specification point, however, only cells that express OCT4 have reach developed EPI, in CAS9

control embryos or in mosaic KO embryos. Interestingly, morphokinetic information suggests that the OCT4 KO embryos have been analyzed at B3 stage, whereas the control embryos have been analyzed at B4 stage. This highlights how our model can be used to refine human preimplantation studies by serving as reference of developmental stage molecular signature progression towards specification (**Figure S3E-F**).

Overall, our global approach yielded a hierarchic model of transcriptomic signature towards specification. All annotations and expression profiles can be found in a user-friendly online browser that we generated to ease access to human preimplantation datasets (**Figure S3G-H**).

Dissection positioning and developmental stage annotations validate pseudotime models

We projected developmental stage annotations on the mouse pseudotime model to confront it to biological knowledge. The unspecified branch is composed mostly of cells from early 16-cells and late 16-cells morulae. All cells from early 32 and late 32 cells morula are in one of the two specified branch; the cells from 64 cells stage blastocysts were all at the tips of specified branches of the pseudotime model (**Figure 2A**). In the original manuscripts, the authors reported Cdx2 differential protein expression levels at late 16 cells morula (Posfai et al., 2017). This raised the hypothesis that there might be a delay between asymmetric expression of Cdx2, and the onset of the TE transcriptomic profile. To validate this hypothesis, we applied a computational method allowing interrogation of RNA velocity, a measurement of non-spliced mRNA (La Manno et al., 2018). This strategy reveals mRNA being actively transcribed in cells and not yet translated, thereby inferring cellular fate. RNA velocity analysis of unspecified cells showed that despite being transcriptomically similar, cells had velocity trending toward ICM or TE, supporting that Cdx2 protein levels are beginning to be heterogeneous at late 16 cells but transcriptomic differences are visible at 32-cells morula stage (**Figure 2B**).

For human pseudotime model, we projected the stage (morula or blastocyst) and position of dissection prior to scRNAseq (mural TE or polar TE/ICM). This revealed the positioning of morula cells in the middle of the first branch of the pseudotime model (**Figure 2C**). The branching occurs in blastocysts, corroborating the conclusion

from Petropoulos et al (Petropoulos et al., 2016). RNA velocity analysis of unspecified and specified cells from the pseudotime model showed that despite having similar transcriptomes, unspecified cells are already starting to transcribe RNA related to the EPI or TE branch (**Figure 2D**). This phenomenon is well illustrated by *KLF17* and *GATA2* expression profiles: cells at early blastocyst stage co-express *KLF17* and *GATA2* at the mRNA and protein level (Kilens et al., 2018), but are enriched in either *KLF17* or *GATA2* RNA velocity (**Figure S4A**).

Biological information validated mouse and human pseudotime models. RNA velocity approach confirmed that despite their transcriptomic similarity, unspecified cells had already initiated transcription of genes characteristic of one fate or the other.

Human TE and EPI lineages are transcriptomically distinct at B2/B3 stage

To link developmental stage and molecular events, we performed an embryo-based analysis, linking cells with a similar embryo origin. We projected all cells from each embryo on the pseudotime model. For example, projection of blastocyst ID=13 (B4 stage upon scRNAseq analysis) showed that at this stage, some cells are populating the EPI branch while other cells are populating the beginning of the TE branch, confirming that specification has occurred at this stage (**Figure 2E-F**). Our annotation based on time-lapse and *zona pellucida* thickness on all our embryos shows that in human, TE and ICM cells are specified between the B2 and B3 stages, after the beginning of blastocyst cavitation (**Figure S4B** and **Figure 2G**). The combination of our scRNAseq analysis of time-lapse annotated human embryos with previous datasets resolved annotation overlap between 4, 5 and 6 dpf, which limited the translation of those datasets into sequence of events pacing human preimplantation development. Indeed, during the 5th dpf, human embryos progress from early blastocysts (B1 and B2) to blastocysts (B3) and expanded blastocysts (B4).

Altogether, transcriptomic modeling of mouse and human lineage specification shows a remarkable difference in the timing between species and identified that the first specified lineages transcriptomically distinguishable appear at B2/B3 stage in human, whereas they appear at 16/32-cell stage in mouse morula.

Human morula and early blastocyst cells are heterogeneous but yet unspecified

As specified cells appears later in human compared to mouse, we checked whether initiation of specification occurs in a similar manner in human compared to other mammalian species: by the sequential establishment of adherent junctions, marked by *CDH1*, followed by the establishment of tight-junctions, marked by *ZO1* (Barcroft et al., 1998; Barcroft et al., 2003; White et al., 2018). Immunostaining of *ZO1* and *CDH1* during human preimplantation development confirmed basolateral expression of *CDH1* from the morula stage, with subsequent apical recruitment of *ZO1* (**Figure 3A**). This supports that lineage specification in human is initiated by morphogenesis event similar to other mammalian species (Reijo Pera and Prezzoto, 2016).

Following up on our analysis at the embryo level, we then clustered embryos based on the position on the pseudotime model of all the cells that belong to each embryo. Each embryo can only belong to one cluster. We then measured the pseudotime spreading of each cluster, and reported annotations associated with embryos from each cluster. For example, cluster 1 spans 5 pseudotime distance, and contains only 3 dpf embryos (**Figure 3B**). The clusters K3 and K4 are standing out: those clusters have the highest pseudotime distance variation, 33 and 32, respectively. Interestingly, cluster K3 contains mostly 4 dpf embryos. This support heterogeneity within human morula, reminiscent of *Cdx2* asymmetric expression in 16-cells mouse morula. We have yet to define which transcription factor would be initiating TE specification in human.

Coupling gene expression correlation with pseudotime highlights sequence of events during lineage specification

Our pseudotime model gave us the opportunity to explore sequence of expression within the embryo. To further investigate whether multiple states were present within each branch, we performed a gene-based analysis. 8 major modules of co-expressed genes were identified using weighted correlation network analysis (WGCNA)(Langfelder and Horvath, 2008), subdividing the pseudotime into 9 states defined by absence or presence of gene module (**Figure 4A-B**). Modules are named after one of its representative gene. Each module contains specific expression dynamics, linked to

specific transcription factors and enriched functions (**Figure S5**).

We investigated proteins that could be markers of outer cells in human morula and found that AQP3 expression profile was high in morula, then restricted to TE (**Figure 4C**). An important event in the mammalian preimplantation development is the expression of aquaporins, at the morula/early blastocyst stage (Barcroft et al., 2003). Remarkably, AQP3 is the only water channel expressed during human preimplantation development and one of the few genes that are initially expressed in all cells then restricted to TE. Immunostaining for AQP3 revealed cytoplasmic membrane-localization at the morula stage, before being restricted apically at B2 stage (**Figure 4D**). AQP3 is therefore an early event marking lineage specification of the outer cells of the embryo.

Transcription factors are of particular interest since they can drive changes in transcriptomic signatures. We analyzed specifically the transcription factors expressed in each gene modules (**Figure 4E**). Some genes modules reflected overall changes in the embryos: ZSCAN4 and DUXA modules are associated with zygotic genome activation between 8-cell and morula stages; the DNMT3L module is emerging in all cells at the blastocyst stage and contains epigenetic regulators in line with methylation dynamics in human preimplantation development and metabolic pathways. The POU5F1B module contains pluripotency genes. Its expression starts in all cells of the morula, and then is restricted to the developed EPI. This make POU5F1B module the first expressed module that has a cell lineage specific expression later on. De Paepe *et al* showed that TE cells from early blastocysts (B3) are plastic and can give rise to ICM cells but this plasticity is lost in expanded blastocysts (B4) (De Paepe et al., 2013), which correlates with loss of expression of POU5F1B module in the TE. Other modules are related to specified lineages: IFI16 module for developed/stable-EPI, GATA4 module for PrE, GATA2 module for early-TE/TE, and NR2F2 module for developed-TE.

IFI16 is a marker of epiblast proper

Using our transcriptomic model, we sought markers of developed-EPI and developed-TE in B5 blastocyst, as this stage is easily characterized by hatching. Among transcription factors, we identified IFI16 as a candidate for developed-EPI. IFI16 is

induced by interferon gamma in other cellular contexts. We analyzed protein expression of IFI16 in a B5 embryos which showed IFI16 positive staining restricted to ICM cells (**Figure 4F**). Analysis of time-lapse assessed developmental stages showed no IFI16 expression before B2 and a robust expression after B4 (**Figure 4G**). IFI16 is therefore specific to EPI, marking the onset of EPI proper. Identifying regulators of IFI16 could solve pathways involved in EPI specification.

Maturation of TE occurs at the interface between EPI and polar TE in human

Another transcription factor of interest was NR2F2, as it was expressed in the TE branch after expansion. In B5 blastocysts, immunofluorescence showed nuclear localization of NR2F2 in TE, but restricted to the polar TE cells juxtaposed to the ICM marked with NANOG (EPI) (**Figure 5A**). To investigate if NR2F2 is expressed readily in specified TE, we investigated B4 blastocyst, a stage immediately following lineage specification. Immunostaining revealed GATA3 and GATA2 expression throughout TE while NR2F2 was only expressed in the polar TE cells (**Figure 5B**). Therefore, from 6 dpf, TE cells can be classified as NR2F2-positive or NR2F2-negative cells. Systematic quantification of NR2F2-positive cells showed a robust expression from B4 stage on, underlining the use of NR2F2 as a molecular marker of developed-TE (**Figure 5C**). Of note, at B5 stage the majority GATA4 positive cells are still IFI16 positive, suggesting that PrE might not be fully specified at that point (**Figure 5D** and **Figure S6**).

As NR2F2-positive cells localized at the end of the TE branch on the pseudotime model, the most developed part of TE signature, we propose that TE development is stimulated by contact with the ICM. The observation that polar TE is developing faster than mural TE is consistent with the observation that the majority of human blastocysts attached to endometrial cells by the polar side with subsequent spreading of TE development to mural cells (Aberkane et al., 2018; Bentin-Ley et al., 2000; Deglincerti et al., 2016; Grewal et al., 2008; Lindenberg, 1991; Shahbazi et al., 2016). Analysis of most enriched signaling pathways pointed out prime candidates potentially driving molecular dialog between EPI and TE: TGF β , IGF1, BMP2, IL6 and FGF4 (**Figure S7** and **Figure 5E**). FGF4 signaling has been previously ruled out to have an impact during human preimplantation development, but our

analysis calls for a finer analysis of the role of FGF signaling during human preimplantation development (Kuijk et al., 2012; Roode et al., 2012). BMP signaling has recently been shown to inhibit cavitation (De Paepe et al., 2019), but its function after the first lineage specification remains to be studied. The role of IL6 has been recently studied in pig: IL6 KO or inhibition of JAK signaling perturbed preimplantation development, supporting the hypothesis that in human and pigs, in absence of LIF, JAK/STAT signaling is triggered by IL6. Polar TE-EPI interactions in human is reminiscent of the ICM-polar TE molecular dialog observed in mouse blastoids (Rivron et al., 2018; White et al., 2018). However, human specific events might be involved as some genes critical for mouse TE maturation are not expressed in human, such as KLF6.

Discussion

Here, we generated mouse and human continuous preimplantation transcriptomic models by implementing a pseudotime-based approach on scRNAseq. The mouse models links transcriptomic profiles with functional investigations. The human model shows the precise timing of transcriptomically specified TE: at the B2/B3 stage. An embryo-based analysis strategy revealed that transcriptomic heterogeneity could be found in human morula, despite the segregation in two lineages. Coupling the pseudotime model with gene-based analysis revealed dynamics of gene modules, defining specific states within each branch of the pseudotime model. This led us to functionally validate by immunofluorescence the onset of IFI16 in specified EPI. Moreover, we discovered that development of TE was initiated on the polar side which yielded a novel model of TE/EPI and TE/endometrium molecular dialogs. Altogether, our study refines our understanding of the spatio-temporal events occurring during human lineage specification (**Figure 6** and **Table S1**).

The delay in the onset of distinct transcriptomic profiles of TE/ICM in mouse in human is rather surprising since compaction seems to involve rapid epithelialization in both cases (White et al., 2018). This observation points out to additional parameters that could be linked to lineage specification. One of them is cell division. Human morulae compact at 10 cells, cavitate at 20 cells and B2 stage blastocysts are generally composed of 40 cells. Therefore,

transcriptomic timing could be dependent on cell cycle and require 2 cell divisions, like in mouse: from 8-cells at compaction to 32 cells for transcriptomic distinct signatures. Finally, it is interesting to note that 90% of unspecified cells will become TE but that the unspecified cells are more transcriptomically similar to EPI cells, in human. Pseudotime models of mouse and human preimplantation development will be an important basis to understand molecular regulation of lineage specification and to assess the similarities and differences between species.

Our analysis suggests that within the ICM, the EPI fate is the first to be achieved. Indeed, individual analysis of the 128 human embryos showed that EPI cells are found in 19/35 embryo at 5 dpf, 16/36 embryos at 6 dpf and 4/17 embryos at 7 dpf while PrE cells are found in 8/35 embryo at 5 dpf, 11/36 embryos at 6 dpf and 10/17 embryos at 7 dpf. At the molecular level, a cluster of pluripotency-associated genes is expressed in all cells before specification then restricted to the EPI. Therefore, fewer genes are necessary to establish EPI fate than PrE fate. Our model highlights how the core pluripotency circuitry evolves during EPI establishment. Among the transcription factors, we identified novel and less-studied transcription factors, such as IFI16, that marks developed-EPI. IFI16 expression suggests that IFN γ signaling might be important for EPI specification; an hypothesis is that IFI16 inhibits the burst of transposable elements expression at the morula stage (Theunissen et al., 2016).

Transcriptomically, TE is specified at B2/B3 with cells progressively losing expression of pluripotency-associated genes, correlating with loss of plasticity of TE cells (De Paepe et al., 2013). We bring new evidences showing that an EPI/TE molecular dialog could be involved in TE development. However, human implantation will occur on the polar side while mouse embryo attachment is mediated by mural TE cells. Understanding this aspect of human preimplantation development is of utmost importance for improvement of in vitro fertilization. Indeed, studies have demonstrated the importance of TE morphological quality for prediction of successful IVF (Ahlgren et al., 2011; Chen et al., 2014; Hill et al., 2013; Thompson et al., 2013). Moreover, we will now be able to focus on specific markers associated with NR2F2 that could inform clinicians about the implantation potential of the blastocyst before transfer. It is therefore an important milestone for IVF since research on

human embryo implantation has largely been focused on factors involved in murine reproduction (Aplin and Ruane, 2017).

The toughest lineage to resolve is PrE, regardless of approaches undertaken (Blakeley et al., 2015; Petropoulos et al., 2016; Stirparo et al., 2018). We consider two hypotheses: either early ICM cells are transcriptomically closer to EPI than to PrE, resulting in ICM cells to fall in the EPI branch of our pseudotime, or PrE differentiates from EPI in human. The later hypothesis is supported by IFI16/GATA4 co-staining in B5 blastocysts: we detected IFI16 positive cells, IFI16/GATA4 positive cells, and few GATA4 cells. This observation also raised the hypothesis that PrE cells are still undergoing specification at B5 stage, and therefore will be specified post-implantation; this is also supported by the fact that there is only a limited set of genes characterizing PrE preimplantation (GATA4.mod, 26 genes). These hypotheses necessitate more studies, especially in human post-implantation embryos to monitor development of PrE.

Our model also provides key knowledge for the modeling of preimplantation development with cellular models. Embryonic signature comparison with pluripotent stem cell residing in various state of pluripotency showed established hierarchy of pluripotent stem cells to mimic EPI (Kilens et al., 2018; Liu et al., 2017; Stirparo et al., 2018). However, no developmental stage has been proposed for human trophoblast stem cells (Okae et al., 2018). The expression of NR2F2 and correlation with NR2F2 positive TE cells support that the TSC are equivalent to polar preimplantation TE or post-implantation cytotrophoblasts (**Castel et al, co-submitted manuscript**). Moreover, we uncovered plasticity between pluripotent stem cells and trophoblast stem cells which is reminiscent of late specification in human (**Castel et al, co-submitted manuscript**). More studies are necessary, including post-implantation human embryo analysis, to further understand the TE and EPI fate boundaries.

In conclusion, our study highlights signaling pathways and events involved during human preimplantation development linked to IVF. Our model therefore paves the way to more efficient media formulation and readouts to assess development of human embryos. Improvement of IVF procedures is necessary since the procedural average efficiency is below 27% (Embryology,

2015). Our study contributes to an improved understanding of human preimplantation, a gateway to improve IVF success rates.

Methods

Detailed methods are provided in the online version of this paper and include the following:

- Human preimplantation embryos
- Human preimplantation embryos culture
- Human embryo time-lapse imaging
- Immunofluorescence of human embryos
- Imaging
- Single-cell RNA sequencing
- Raw count table treatments
- Computation of pseudotime model
- WGCNA
- Enrichment analysis
- Loess regressed expression by pseudotime
- Subdivision of pseudotime branches
- Usage of expression per scRNAseq counts table type
- Data visualization
- Mouse single-cell RNA-Seq analysis
- RNA velocity
- Data and software availability

Supplemental information

Supplemental information contains 7 supplementary figures and 5 supplementary tables.

Acknowledgements

We thank our colleagues F. Lanner and K. Niakan for sharing data and providing feedback. We thank J. Jullien, V. Pasque and J. Chappel for critical review of the manuscript. FINOX-Gedeon Richter and MSD contributed to the project. We thank core facilities: BIRD, PFIPSC and MicroPicell. This work was supported by “Paris Scientifique region Pays de la Loire: HUMPLURI” and IHU CESTI.

Author contributions

TF and LD designed the study. DM, SL and LD wrote the manuscript with input from all authors. SL, AR, JF, JL and ABr performed embryo manipulation. PH, SN, SL, AR and LD performed IF analysis. MS and TM performed single-cell RNAseq. DM, VFC, YL, performed bioinformatics analysis under supervision

of ABi and JB and input from BB and GC. TF and PB supervised human embryo donation. All authors approved the final version of the manuscript.

Declaration of interests

Competing interest declaration: DM is supported by FINOX Biotech FORWARD initiative 2016.

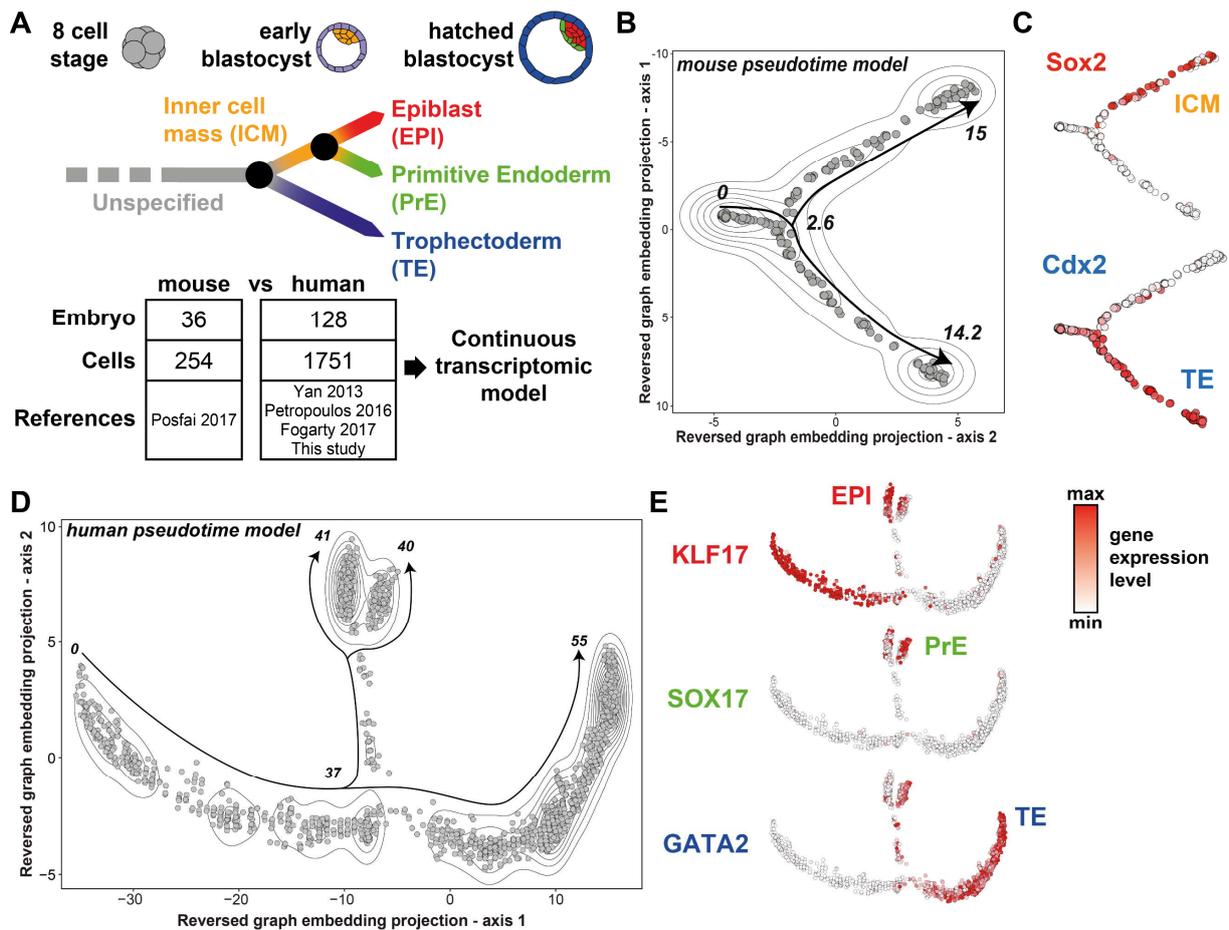


Figure 1. Transcriptomic model of human and mouse preimplantation embryo development

(A) Schematic of chronology and hierarchy of cell lineages during preimplantation development of the placental mammals. Our objective was to combine indicated scRNAseq datasets to generate a continuous transcriptomic model.

(B) Projection from the mouse scRNAseq samples (Posfai et al, 2017) from the reversed graph embedding method (Monocle2). Density curves indicate sample density; pseudotime values represent transcriptomic distance from the root cell.

(C) Projection of lineage marker expression levels on the mouse model: Sox2 (Inner cell mass), Cdx2 (trophectoderm).

(D) Projection of the human dataset from the reversed graph embedding method.

(E) Projection of lineage marker expression levels on the human model: KLF17 (epiblast), SOX17 (primitive endoderm), and GATA2 (trophectoderm).

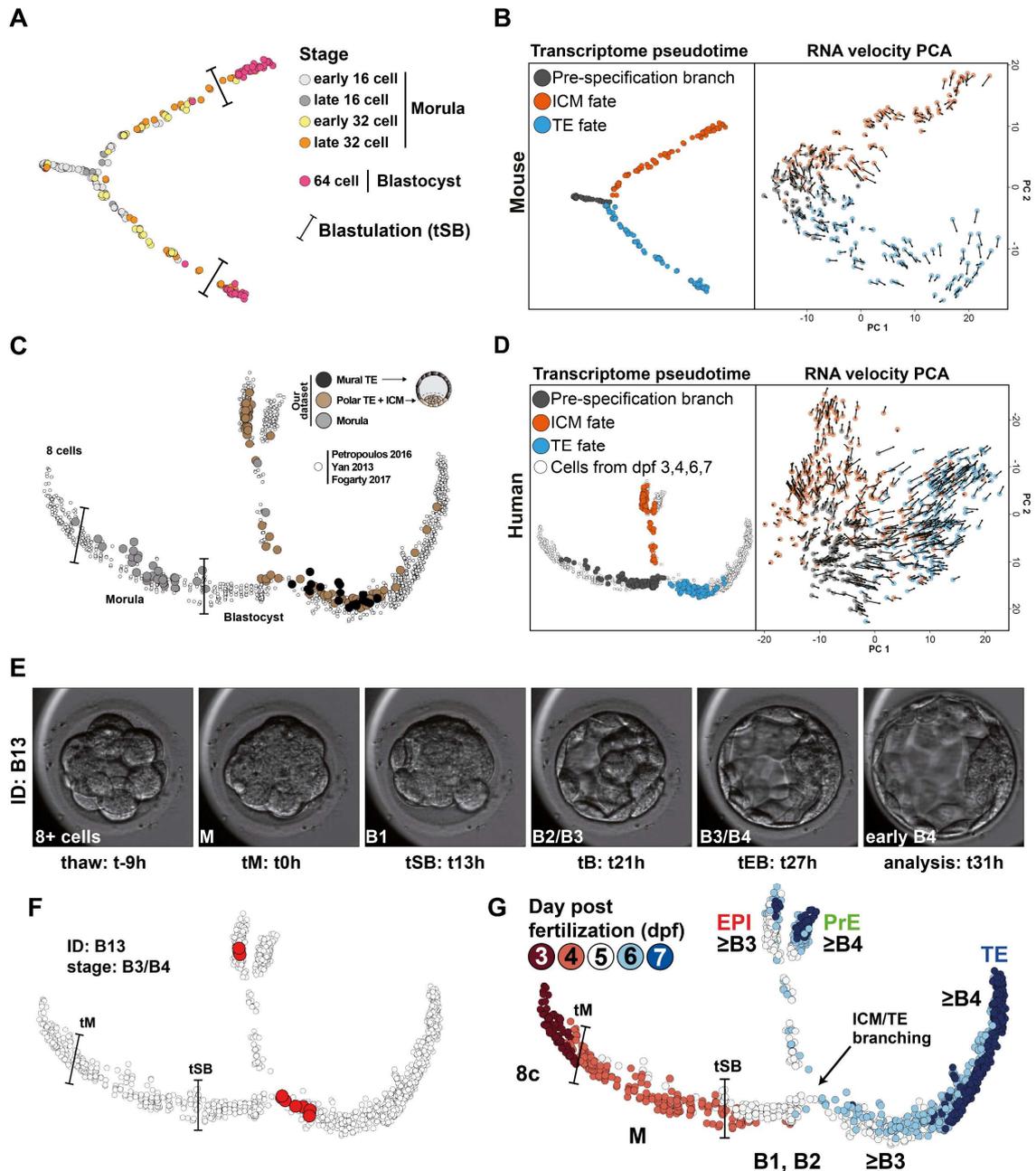


Figure 2. scRNAseq reveals the hierarchy of events during human and mouse preimplantation development

(A). Projection of the developmental stages on the mouse pseudotime model.

(B) Transcriptome pseudotime and RNA velocity comparison of cells before and after specification in mouse.

(C) Projection of cell annotation sequenced for this paper on the pseudotime model. morula cells (grey) and blastocyst dissection origin of cells are indicated (mural trophectoderm: black, polar trophectoderm or inner cell mass: brown).

(D) Transcriptome pseudotime and RNA velocity comparison of cells before and after specification in human.

(E-F) Frames from time-lapse microscopy for embryo "B13". For each embryo sequenced in this study, their morphokinetics were acquired by time-lapse microscopy (E). Developmental events include morula compaction (tM), blastulation (tSB) leading to B1 stage, full blastocyst (tB) at B3 stage and expanded blastocyst once the zona pelucida thickness is halved (tEB) at B4 stage. tM is used as t0 to compare thawed embryos. A projection of cells from embryo B13 on the pseudotime model shows the correspondence between the pseudotime and the stage of the embryo (F).

(G) Projection of developmental day (E3 to E7) for all samples combined for this study, and the result of our refined staging.

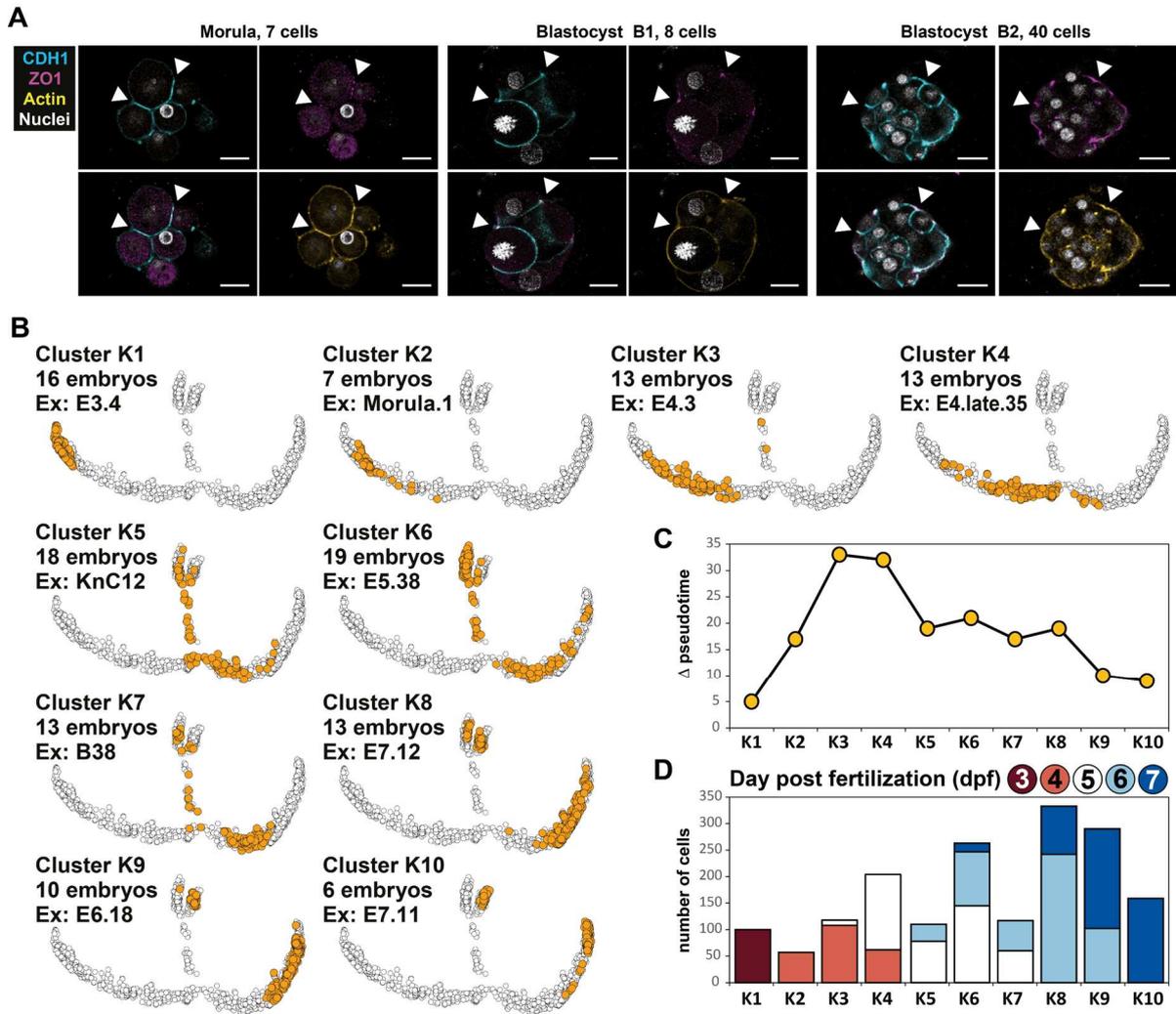


Figure 3. Development and diversity of the preimplantation embryos

(A) Immunofluorescence of CDH1 (cyan), ZO1 (purple), Actin (yellow) and nuclear counterstaining (white) at indicated stages. Arrowhead points to the apical part of the cell, with ZO1 staining next to CDH1 staining.

(B) Projection of each embryo cluster according to the position of their cells on the pseudotime model. For each embryo cluster, the corresponding number of embryos is indicated, as well as a representative embryo is indicated. Clusters are ordered by their mean pseudotime.

(C) Range of pseudotime values per embryo cluster.

(D) Distribution of number of cells per embryo cluster subdivided by day post fertilization.

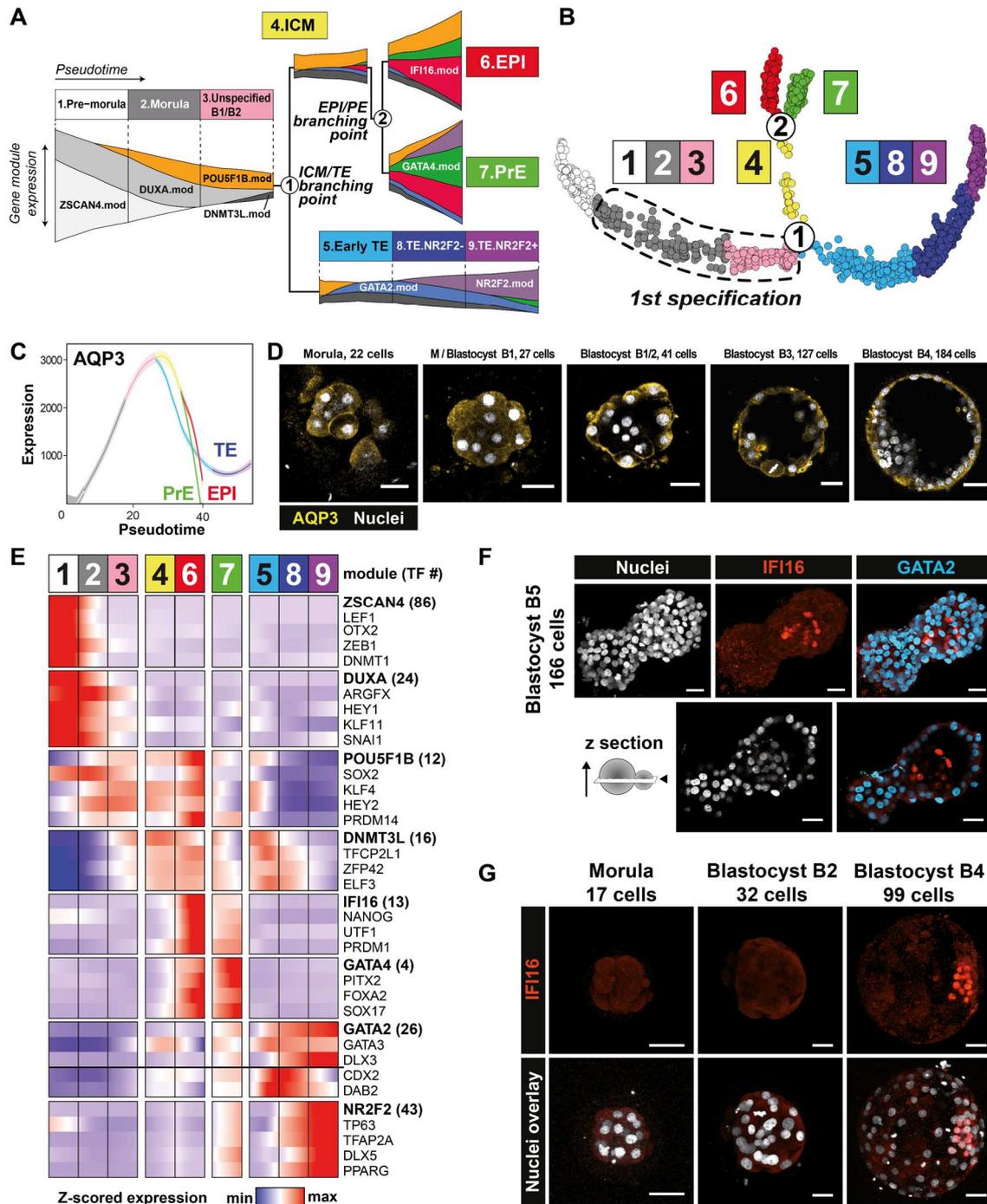


Figure 4. Defining spatio-temporal events pacing human preimplantation development

(A) Streamgraph of gene module expression along pseudotime states. Each module was named by one of its representative genes. Expression level of each module is represented by the thickness of its ribbon on the streamgraph and corresponds to the WGCNA eigengene metric.

(B) Subdivision of branches of the pseudotime model according to module expression: Pre-specification branch is subdivided into 3 states (1.Pre-morula, 2.Morula, 3.Unspecified B1/B2) and the TE branch is subdivided into 3 states (5.Early-TE, 8.TE.NR2F2-, 9.TE.NR2F2+). This yields a total of 9 states, numbered by their order of apparition in the pseudotime.

(C) Expression profile of AQP3 during preimplantation development. Colors correspond to the legend in B.

(D) Immunofluorescence of AQP3 (yellow) and DAPI (white) as nuclear counterstaining, at indicated stages.

(E) Expression levels of transcription factors specific of gene modules. The heatmap is subdivided by states and ordered by pseudotime for each state. Genes were selected by their membership degree for the gene module.

(F) Immunofluorescence of IFI16 (red) and GATA2 (cyan) at B5 blastocyst stage. z-section indicates a z-cutting plane.

(G) Immunofluorescence of IFI16 at indicated stages. Nuclear counterstaining is in white.

Scale bars = 47 μ m.

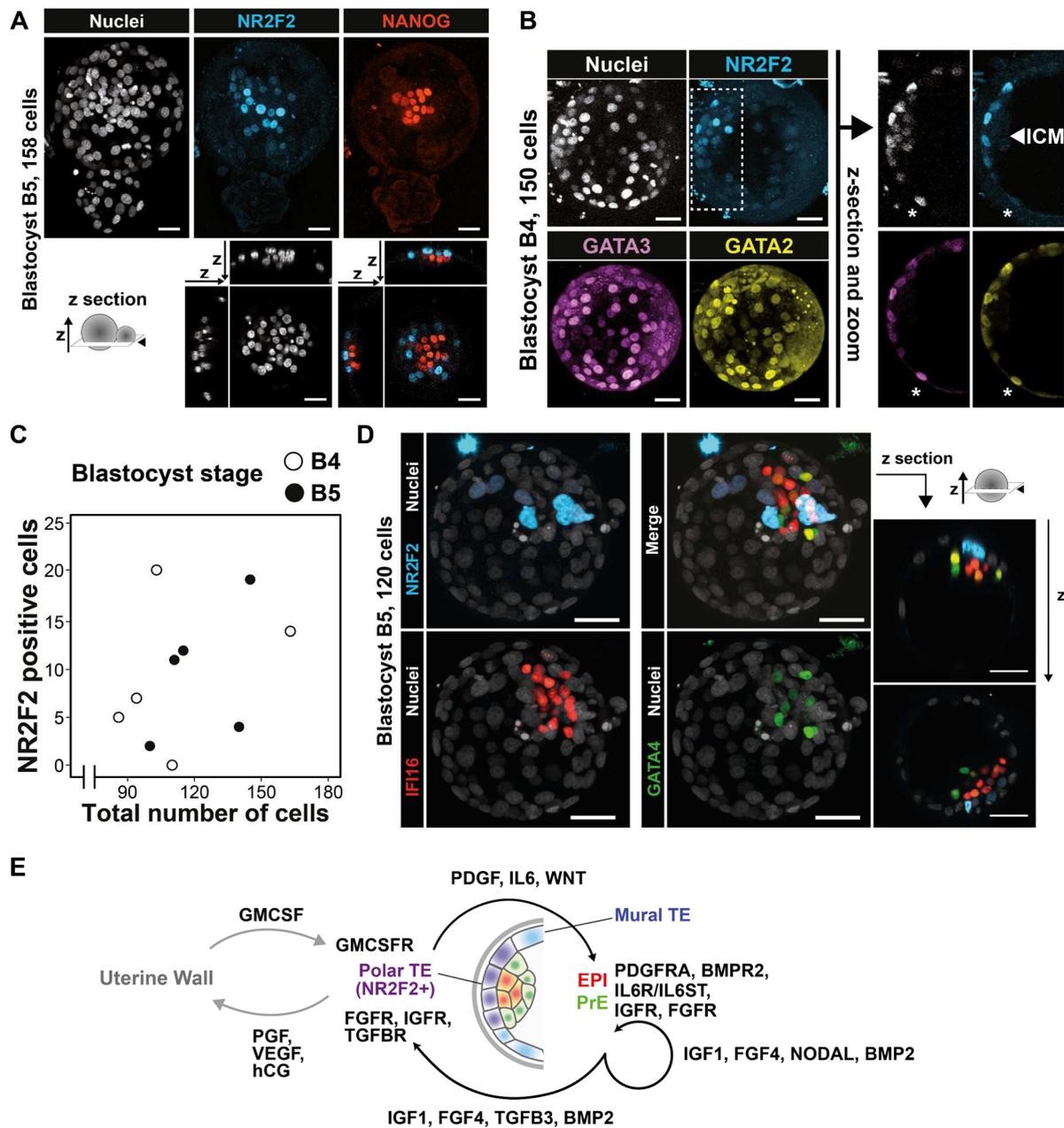


Figure 5. NR2F2 reveals the differential maturation between the polar and mural TE

(A-B) Immunofluorescence of NR2F2 (cyan) and NANOG (red) at B5 stage (A) or with GATA2 (yellow) and GATA3 (purple) at B4 stage (B). Asterisk marks a cell negative for NR2F2 and positive for GATA3 / GATA2.

(C) Quantification of NR2F2 positive cells in human embryos relative to their total number of TE cells or their developmental stage.

(D) Immunofluorescence of IFI16 (red), co-stained with GATA4 (green) and NR2F2 (cyan). Nuclear counterstaining (dapi) is in white.

(E) Schematic representing cytokine-receptors loops identified.

Scale bar = 47 μ m.

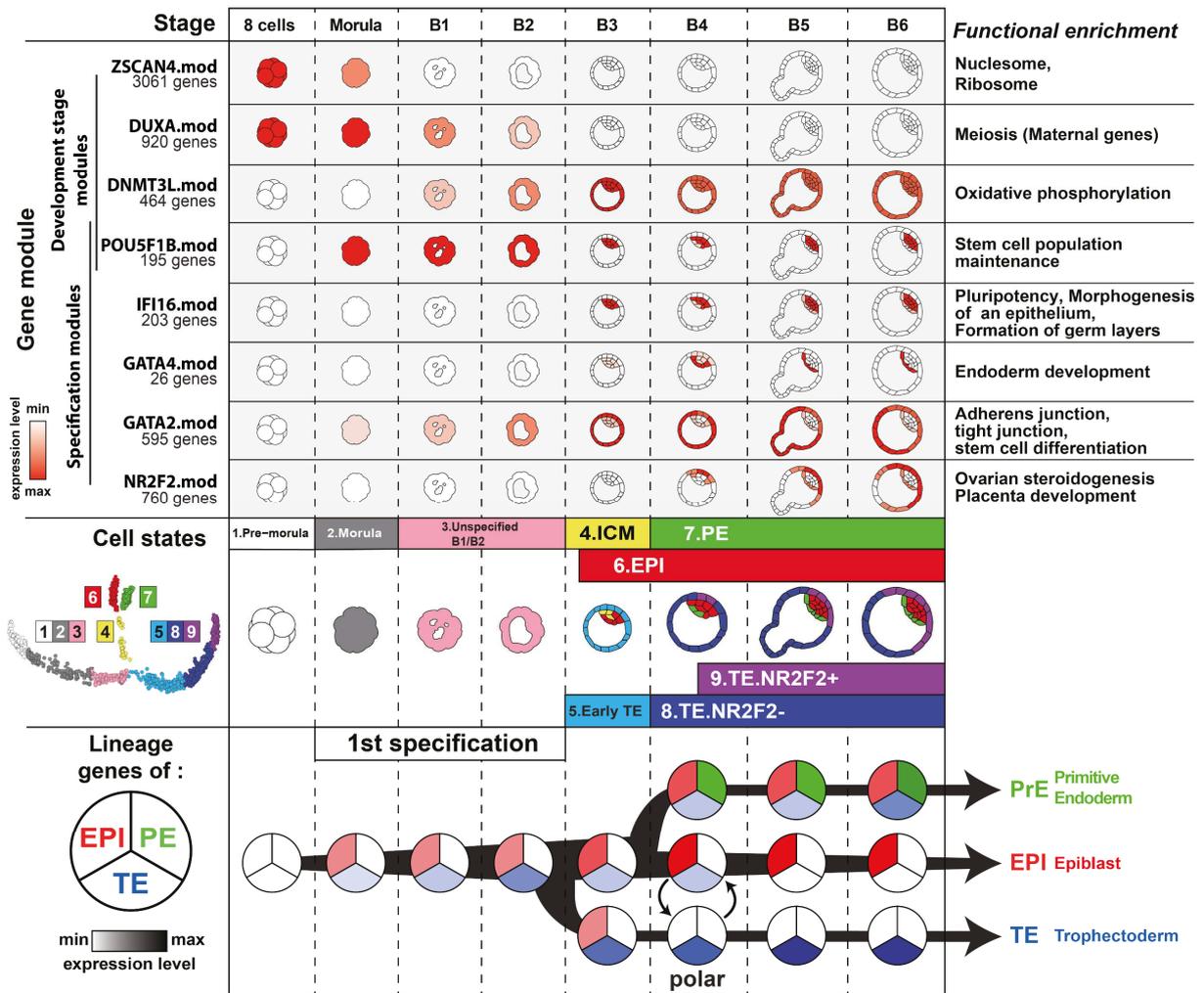


Figure 6. Human preimplantation development model

Each annotation is represented per morphology-based stages of human preimplantation development. Upper panel: Schematic representation of gene module expression. Functional enriched functions are presented on the right. Complete enrichment data are presented in Figure S5B. Middle panel: Schematic of pseudotime cell states spatiotemporal localization. A projection of cell states on the pseudotime model is shown on the left. Lower panel: expression of lineage specific stage per stage and per cell fate.

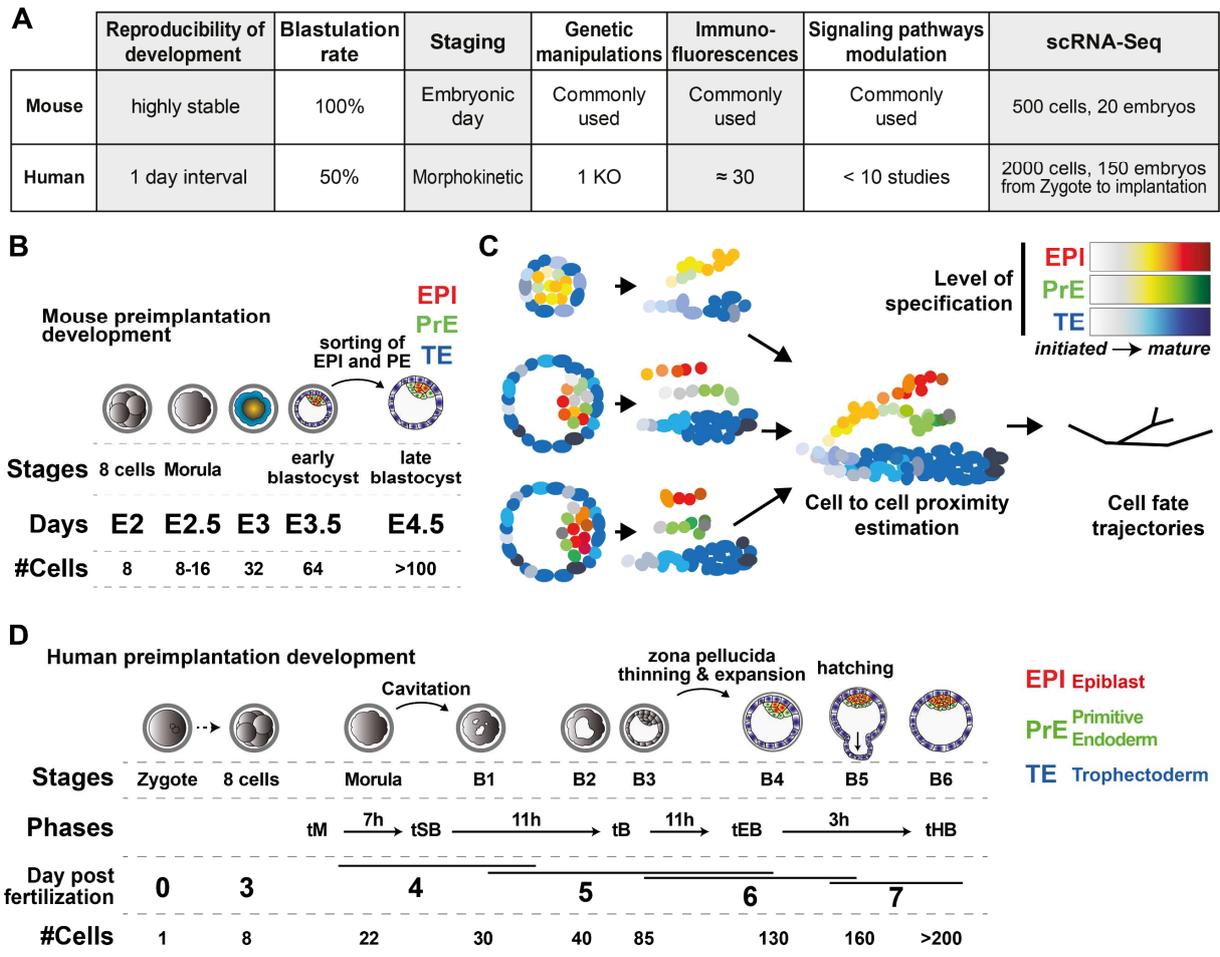


Figure S1 – related to figure 1. Strategy and context of the study

(A) Summary of differences between the data availability of mouse and human in the context of preimplantation development.

(B) Schematic of mouse preimplantation development.

(C) Schematic presentation of pseudotime analysis principle. All single-cell transcriptomic profiles are combined to create the likeliest probabilistic path from one cell to another.

(D) Schematics of human preimplantation development. In IVF clinics, “Stages” and “Phases” are interchangeable. Average time scale for fresh embryos are shown between phases: morula compaction (tM), starting of blastulation (tSB), blastocyst stage (tB), initiation of expansion of the blastocyst (tEB) and hatched blastocyst (tHB). Number of cells indicated is an average from what we observed and from published IF.

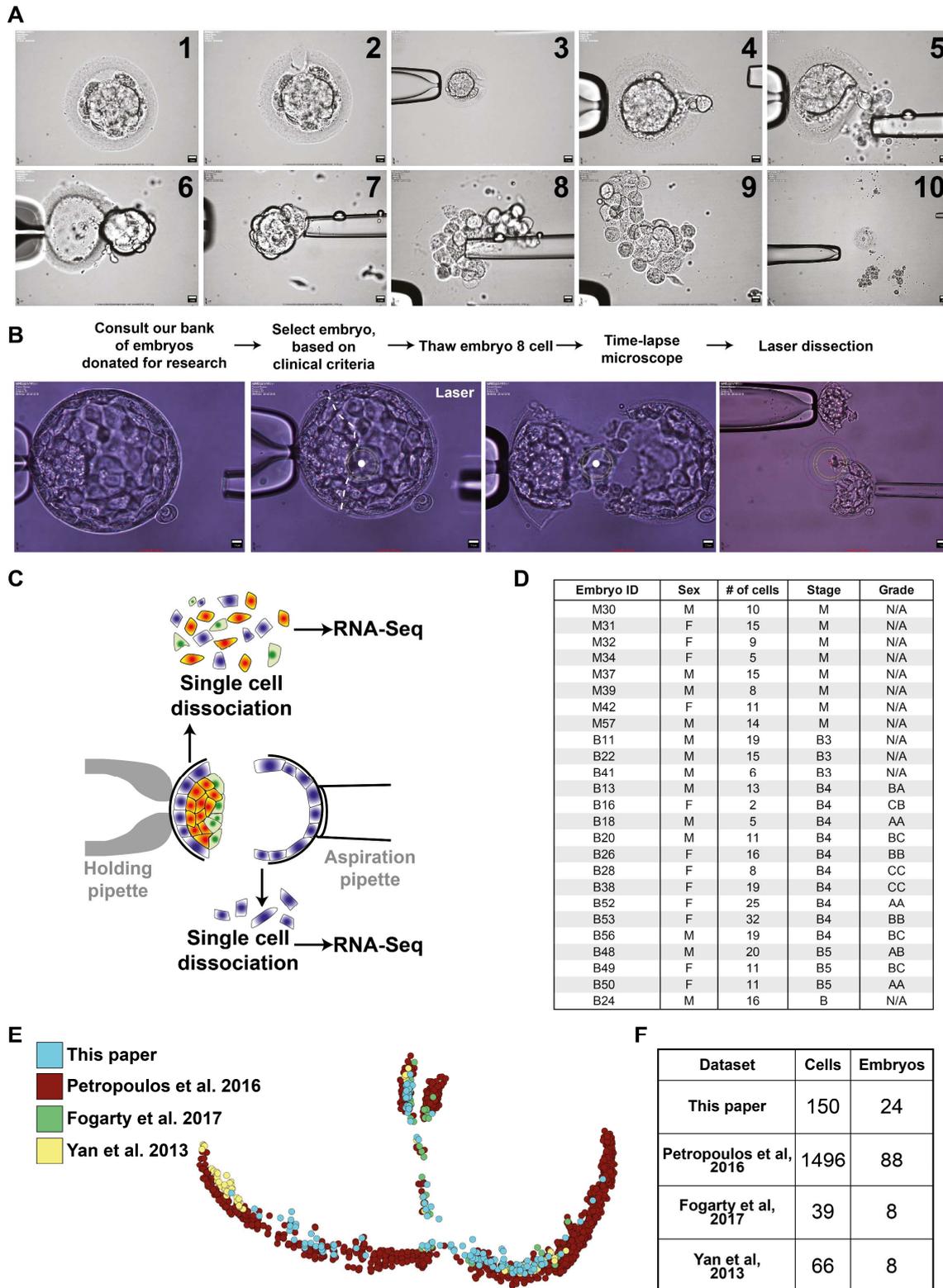


Figure S2 – related to figure 1. Sample preparation and dataset overview

(A-C), Detailed manipulation of embryos before scRNAseq. Morula are first incubated in decompaction media, then the zona pellucida (ZP) is pierced by a laser (dot). Using micropipettes, fragments are removed, and the morula is extracted from the ZP. Cells are then manipulated until dissociated (A). For blastocysts, embryos are thawed and monitored until desired developmental stage (B). Embryos are then laser-dissected, the polar side (containing TE, PrE and EPI cells) and the mural side (containing TE cells) (C).

(D) Detailed sampling of our dataset. For each embryo, the embryo ID, sex, number of cells sequenced, stage and grade are indicated.

(E-F) Projection of samples per dataset on the pseudotime model (E) and sampling of the four datasets (F).

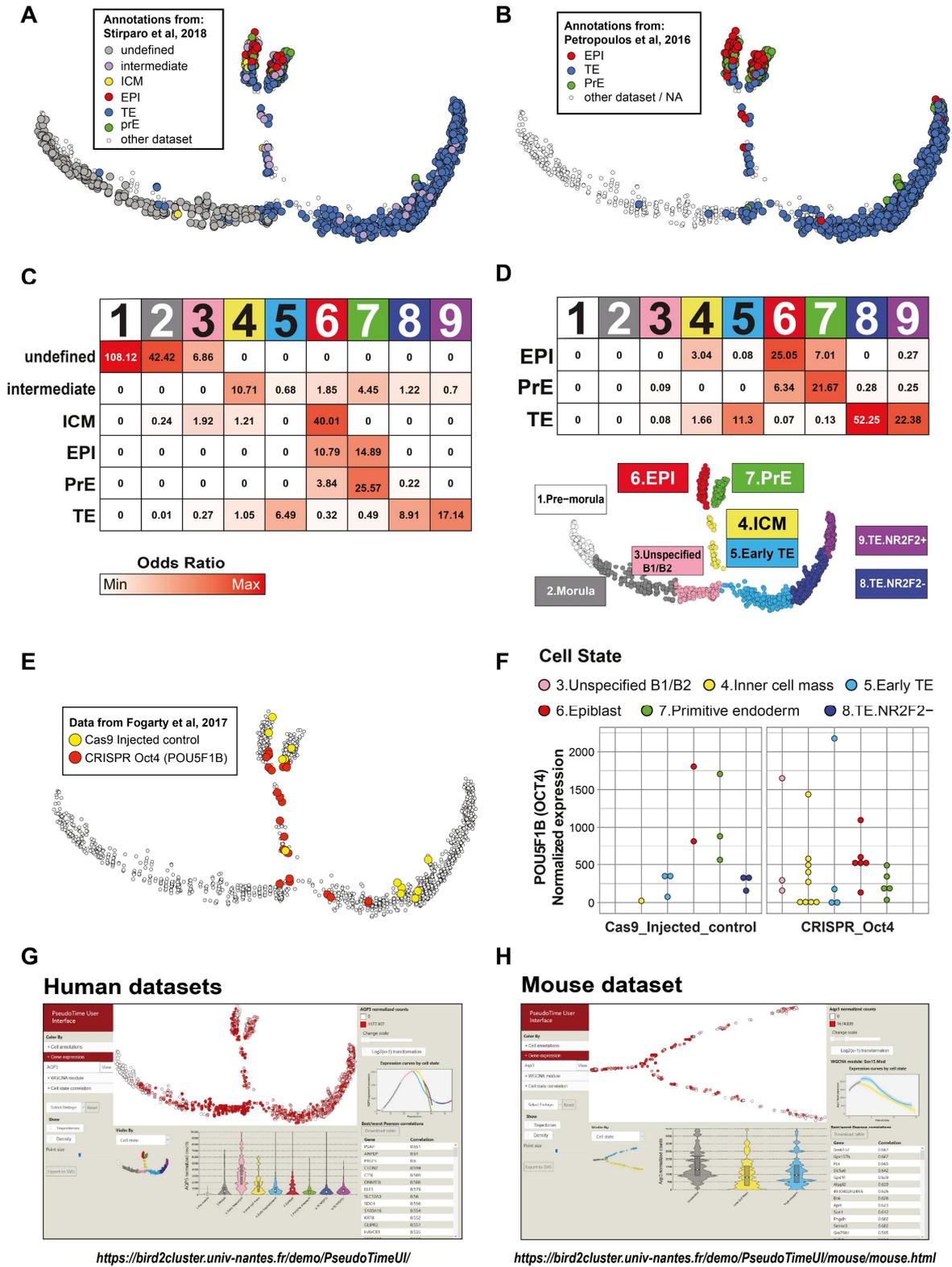


Figure S3 – related to figure 1. Projection of additional annotations on the pseudotime and overview of the pseudotime user interfaces (PTUI)

(A-B), Projection of lineage annotation by Stirparo et al. 2018 (A) or by Petropoulos et al. 2016 (B).
 (C-D) Odds ratios between our lineage annotation and those from Stirparo et al. (C) or Petropoulos et al. (D).
 (E) Projection of human single-cells KO for POU5F1 (orange) or controls (yellow). Samples from Fogarty et al. 2018.
 (F) Expression of POU5F1B in cells from KO or control (cas9 injected only) embryos, per cell state.
 (G-H) Screenshot of our human (E) and mouse (F) web application “PseudoTime User Interface”. URLs are indicated below.

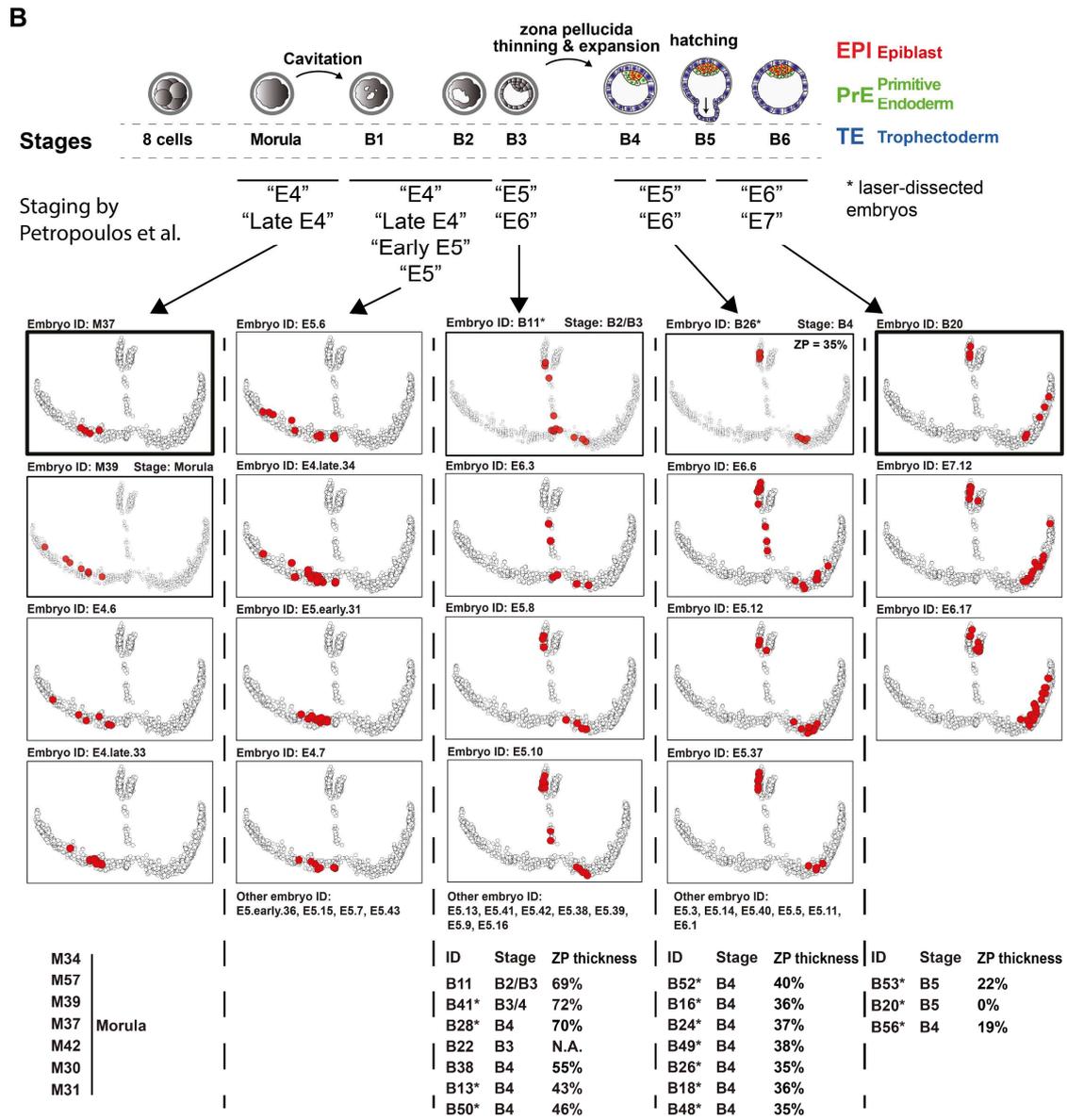
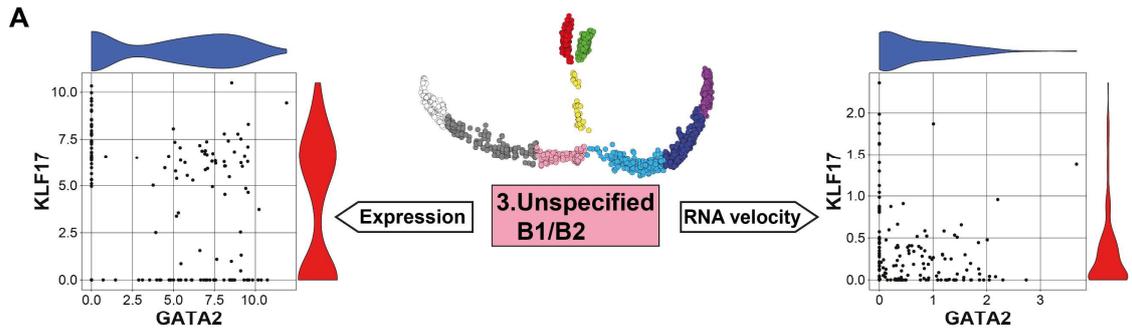


Figure S4 – related to figure 2. Additional results from the study of the pseudotime model

(A) RNA expression and RNA velocity for KLF17 and GATA2 for cells from the pre-specification state.

(B) Refined annotation from all cells and embryos included in our study. Embryos are ranked based on the average positioning of all cells from each embryo included in the study. This highlights that embryos annotated “E5” can be found at the exit of morula, just before specification, just after specification or later on, interspersed with “E6”. This could be due to difficulty to annotate embryos based on a single observation, on the fact that multiple morphologies are present at day 5, or simply because the subset of cells sequenced from an embryo is not representative of the molecular state of the embryo. Our model allows us to bin embryos based on their molecular profile, allowing a finer analysis of events pacing human preimplantation development, especially at E5.

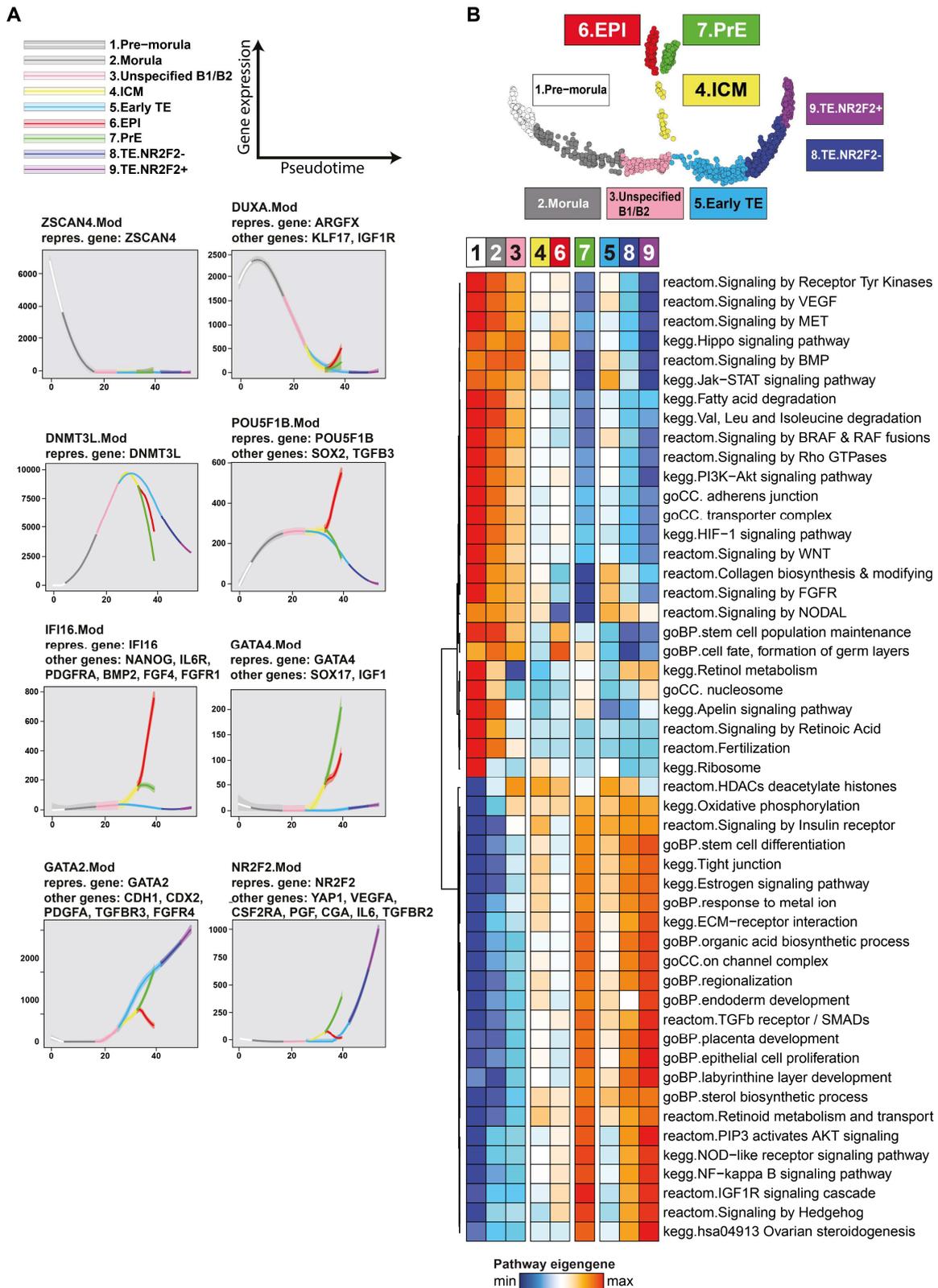


Figure S5 – related to figure 4. Details on WGCNA co-expression gene modules

(A) Expression profile of a representative gene of each gene module, based on their similarity to the global behavior of the module.

(B) Heatmap representing the average of 51 pathway eigengene across the pseudotime states. Pathway eigengene is defined as the first principal component of a principal component analysis with the genes of the pathway as initial dimensions. Enrichment was done in gene modules according to 5 databases: Gene ontology (GO Molecular Function, GO Cellular component, GO Biological process), KEGG and Reactom. All terms were significantly enriched in at least one gene module with an adjusted pvalue < 0.05.

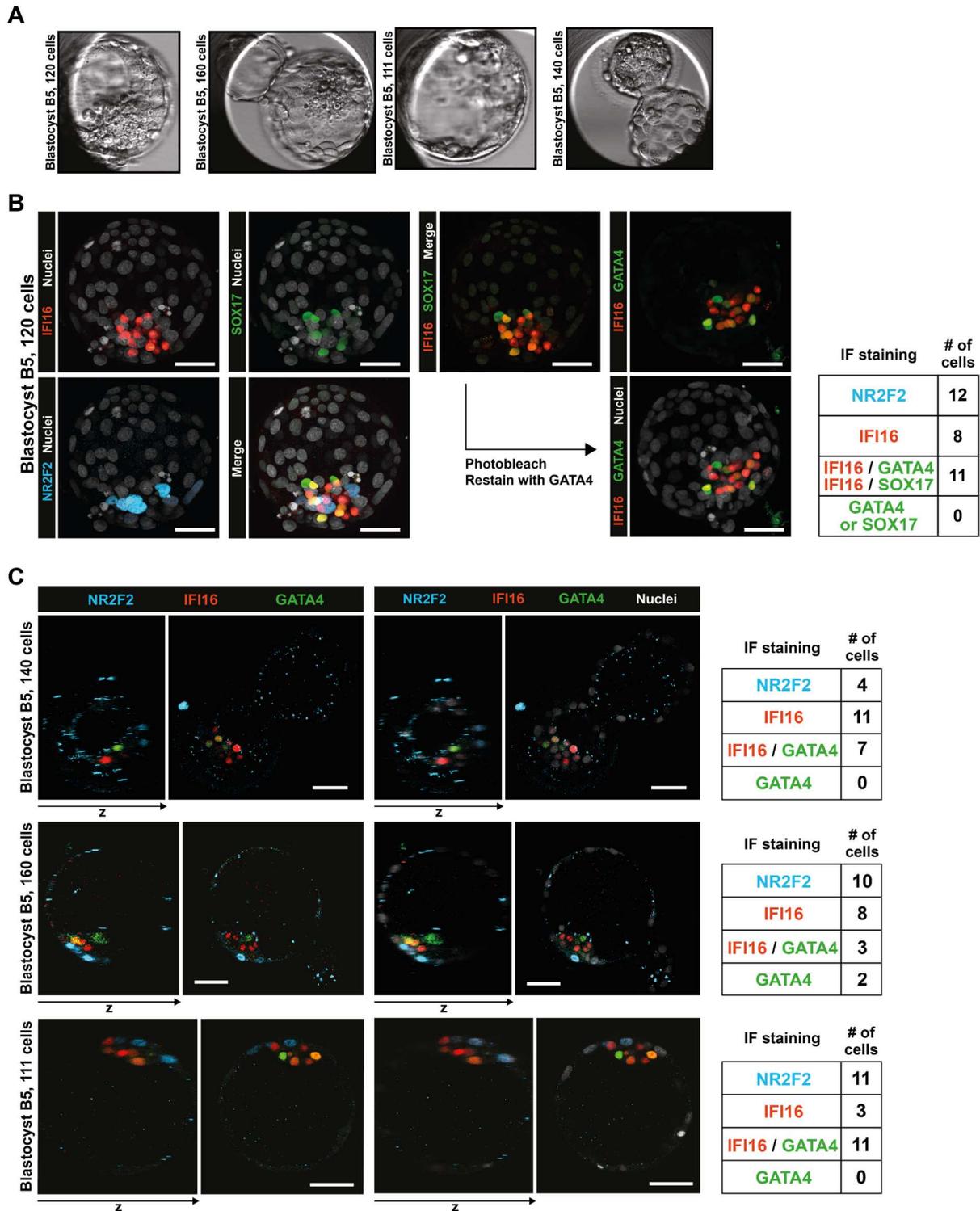


Figure S6 – related to figure 5. Additional immunofluorescence visualization

(A) Extended focus visualization of immunofluorescence of ACTIN (yellow), CDH1 (cyan) and ZO1 (purple) with nuclear counterstaining (white) for a B4 blastocyst.

(B) Split-channels immunofluorescence for AQP3 (yellow) with nuclear counterstaining (white) at indicated stages.

(C-D) Immunofluorescence of IF116 (red), co-stained with GATA4 (green) and NR2F2 (cyan). Number of positive cells for each staining are indicated. The embryo in c was first stained for SOX17, then stained for GATA4. Scale bar = 47µm.

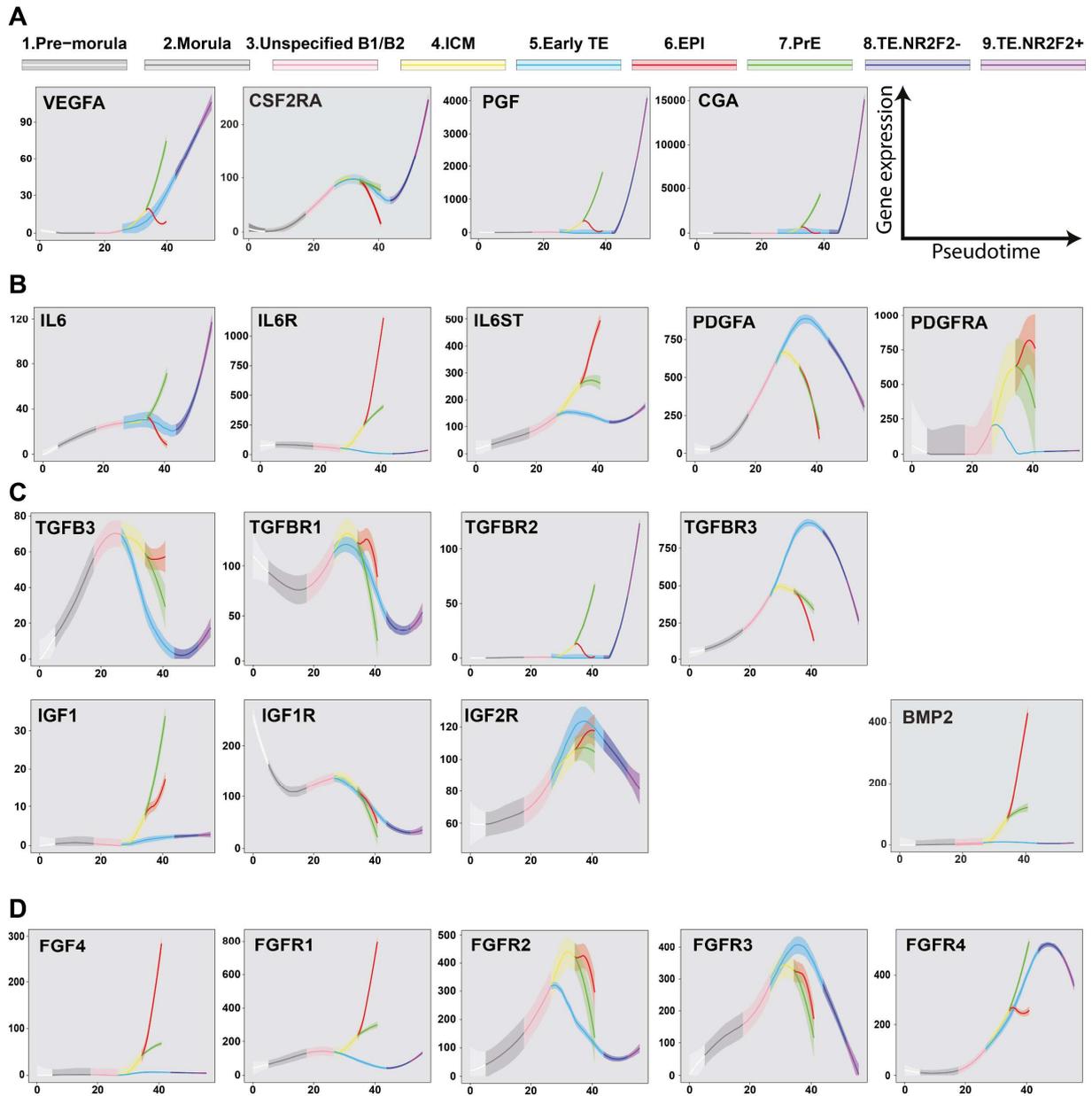


Figure S7 – related to figure 6. Expression profile of selected cytokines and receptors

(A-D), Expression profiles signaling pathways components during preimplantation development. Cytokines secreted by the TE targeting the uterine wall (A), cytokines expressed by the TE and targeting the ICM (B), cytokines expressed by the ICM and targeting the TE (C) and cytokines expressed by the ICM targeting the whole embryo (D).

Supplementary Table Legends

Supplementary legends are available at the preprint page:

<https://www.biorxiv.org/content/10.1101/604751v1.supplementary-material>

Table S1. Order of apparition of cell states and WGCNA modules

For each day post fertilization, data are represented as follows: embryo with at least one positive cell / total number of embryos. For WGCNA modules, a cell was considered positive when the module eigengene was superior to 0.

Table S2. Antibodies

Detail of primary and secondary antibodies used for this study.

Table S3. Cells annotations for human dataset

Annotation for each cell used in the study of human preimplantation development. Annotations beginning with "Author" are referring to annotation of author's dataset. "Author.lineage": lineage attribution of authors. "Author.TEside": Trophectoderm side attribution from Petropoulos et al. (2016). "Author.pseudoTime": pseudotime of each cell according to Petropoulos et al. (2016). "StirparotEtAl.lineage": lineage attribution of dataset from Petropoulos et al. (2016) by Stirparo et al. (2018).

Table S4. Cells annotations for mouse dataset

Annotation for each cell used in the study of human preimplantation development. Annotations beginning with "Author.lineage" is referring to the lineage attribution of Posfai et al. (2017).

Table S5. Gene modules

Detailed gene list for each WGCNA modules.

References

- Aberkane, A., Essahib, W., Spits, C., De Paepe, C., Sermon, K., Adriaenssens, T., Mackens, S., Tournaye, H., Brosens, J.J., and Van de Velde, H. (2018). Expression of adhesion and extracellular matrix genes in human blastocysts upon attachment in a 2D co-culture system. *Molecular human reproduction* 24, 375-387.
- Ahlstrom, A., Westin, C., Reismer, E., Wikland, M., and Hardarson, T. (2011). Trophectoderm morphology: an important parameter for predicting live birth after single blastocyst transfer. *Human reproduction* 26, 3289-3296.
- Alpha Scientists in Reproductive, M., and Embryology, E.S.I.G.o. (2011). The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Human reproduction* 26, 1270-1283.
- Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* 31, 166-169.
- Aplin, J.D., and Ruane, P.T. (2017). Embryo-epithelium interactions during implantation at a glance. *Journal of cell science* 130, 15-22.
- Barcroft, L.C., Hay-Schmidt, A., Caveney, A., Gilfoyle, E., Overstrom, E.W., Hyttel, P., and Watson, A.J. (1998). Trophectoderm differentiation in the bovine embryo: characterization of a polarized epithelium. *Journal of reproduction and fertility* 114, 327-339.
- Barcroft, L.C., Offenberg, H., Thomsen, P., and Watson, A.J. (2003). Aquaporin proteins in murine trophectoderm mediate transepithelial water movements during cavitation. *Developmental biology* 256, 342-354.
- Bentin-Ley, U., Horn, T., Sjogren, A., Sorensen, S., Falck Larsen, J., and Hamberger, L. (2000). Ultrastructure of human blastocyst-endometrial interactions in vitro. *Journal of reproduction and fertility* 120, 337-350.
- Blakeley, P., Fogarty, N.M.E., del Valle, I., Wamaita, S.E., Hu, T.X., Elder, K., Snell, P., Christie, L., Robson, P., and Niakan, K.K. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* 142, 3151-3165.
- Chazaud, C., and Yamanaka, Y. (2016). Lineage specification in the mouse preimplantation embryo. *Development* 143, 1063-1074.
- Chen, X., Zhang, J., Wu, X., Cao, S., Zhou, L., Wang, Y., Chen, X., Lu, J., Zhao, C., Chen, M., *et al.* (2014). Trophectoderm morphology predicts outcomes of pregnancy in vitrified-warmed single-blastocyst transfer cycle in a Chinese population. *Journal of assisted reproduction and genetics* 31, 1475-1481.
- Ciray, H.N., Campbell, A., Agerholm, I.E., Aguilar, J., Chamayou, S., Esbert, M., Sayed, S., and Time-Lapse User, G. (2014). Proposed guidelines on the nomenclature and annotation of dynamic human embryo monitoring by a time-lapse user group. *Human reproduction* 29, 2650-2660.
- De Paepe, C., Aberkane, A., Dewandre, D., Essahib, W., Sermon, K., Geens, M., Verheyen, G., Tournaye, H., and Van de Velde, H. (2019). BMP4 plays a role in apoptosis during human preimplantation development. *Molecular reproduction and development* 86, 53-62.
- De Paepe, C., Cauffman, G., Verloes, A., Sterckx, J., Devroey, P., Tournaye, H., Liebaers, I., and Van de Velde, H. (2013). Human trophectoderm cells are not yet committed. *Human reproduction* 28, 740-749.
- Deglincerti, A., Croft, G.F., Pietila, L.N., Zernicka-Goetz, M., Siggia, E.D., and Brivanlou, A.H. (2016). Self-organization of the in vitro attached human embryo. *Nature* 533, 251-254.
- Embryology, E.S.o.H.R.a. (2015). Data collection and research (<https://www.eshre.eu/>).
- Fogarty, N.M.E., McCarthy, A., Snijders, K.E., Powell, B.E., Kubikova, N., Blakeley, P., Lea, R., Elder, K., Wamaita, S.E., Kim, D., *et al.* (2017). Genome editing reveals a role for OCT4 in human embryogenesis. *Nature* 550, 67-73.
- Frum, T., Murphy, T.M., and Ralston, A. (2018). HIPPO signaling resolves embryonic cell fate conflicts during establishment of pluripotency in vivo. *Elife* 7.
- Gardner, D.K., Lane, M., Stevens, J., Schlenker, T., and Schoolcraft, W.B. (2000). Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertility and sterility* 73, 1155-1158.
- Grewal, S., Carver, J.G., Ridley, A.J., and Mardon, H.J. (2008). Implantation of the human embryo requires Rac1-dependent endometrial stromal cell migration. *Proceedings of the National Academy of Sciences of the United States of America* 105, 16189-16194.
- Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 32, 2847-2849.
- Hill, M.J., Richter, K.S., Heitmann, R.J., Graham, J.R., Tucker, M.J., DeCherney, A.H., Browne, P.E., and Levens, E.D. (2013). Trophectoderm grade predicts outcomes of single-blastocyst transfers. *Fertility and sterility* 99, 1283-1289 e1281.
- Kilens, S., Meistermann, D., Moreno, D., Chariou, C., Gaignerie, A., Reignier, A., Lelievre, Y., Casanova, M., Vallot, C., Nedellec, S., *et al.* (2018). Parallel derivation of isogenic human primed and naive induced pluripotent stem cells. *Nat Commun* 9, 360.
- Kimmelman, J., Heslop, H.E., Sugarman, J., Studer, L., Benvenisty, N., Caulfield, T., Hyun, I., Murry, C.E., Sipp, D., and Daley, G.Q. (2016). New ISSCR guidelines: clinical translation of stem cell research. *Lancet* 387, 1979-1981.

Kuijk, E.W., van Tol, L.T., Van de Velde, H., Wubbolts, R., Welling, M., Geijssen, N., and Roelen, B.A. (2012). The roles of FGF and MAP kinase signaling in the segregation of the epiblast and hypoblast cell lineages in bovine and human embryos. *Development* *139*, 871-882.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastrioti, M.E., Lonnerberg, P., Furlan, A., *et al.* (2018). RNA velocity of single cells. *Nature* *560*, 494-498.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics* *9*, 559.

Lindenborg, S. (1991). Experimental studies on the initial trophoblast endometrial interaction. *Dan Med Bull* *38*, 371-380.

Liu, X., Nefzger, C.M., Rossello, F.J., Chen, J., Knaupp, A.S., Firas, J., Ford, E., Pflueger, J., Paynter, J.M., Chy, H.S., *et al.* (2017). Comprehensive characterization of distinct states of human naive pluripotency generated by reprogramming. *Nature methods* *14*, 1055-1062.

Liu, Z.P., Wu, C., Miao, H., and Wu, H. (2015). RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. *Database : the journal of biological databases and curation* *2015*.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* *5*, 2122.

Niakan, K.K., and Eggan, K. (2013). Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental biology* *375*, 54-64.

Okabe, H., Toh, H., Sato, T., Hiura, H., Takahashi, S., Shirane, K., Kabayama, Y., Suyama, M., Sasaki, H., and Arima, T. (2018). Derivation of Human Trophoblast Stem Cells. *Cell stem cell* *22*, 50-63 e56.

Petropoulos, S., Edsgard, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* *165*, 1012-1026.

Posfai, E., Petropoulos, S., de Barros, F.R.O., Schell, J.P., Jurisica, I., Sandberg, R., Lanner, F., and Rossant, J. (2017). Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *Elife* *6*.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliener, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature methods* *14*, 979-982.

Reijo Pera, R.A., and Prezzoto, L. (2016). Species-Specific Variation Among Mammals. *Current topics in developmental biology* *120*, 401-420.

Rivron, N.C., Frias-Aldeguer, J., Vrij, E.J., Boisset, J.C., Korving, J., Vivie, J., Truckenmuller, R.K., van Oudenaarden, A., van Blitterswijk, C.A., and Geijssen, N. (2018). Blastocyst-like structures generated solely from stem cells. *Nature* *557*, 106-111.

Roode, M., Blair, K., Snell, P., Elder, K., Marchant, S., Smith, A., and Nichols, J. (2012). Human hypoblast formation is not dependent on FGF signalling. *Developmental biology* *361*, 358-363.

Rossant, J., and Tam, P.P. (2017). New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. *Cell stem cell* *20*, 18-28.

Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*.

Shahbazi, M.N., Jedrusik, A., Vuoristo, S., Recher, G., Hupalowska, A., Bolton, V., Fogarty, N.M., Campbell, A., Devito, L.G., Ilic, D., *et al.* (2016). Self-organization of the human embryo in the absence of maternal tissues. *Nature cell biology* *18*, 700-708.

Stirparo, G.G., Boroviak, T., Guo, G., Nichols, J., Smith, A., and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human pre-implantation epiblast. *Development* *145*.

Theunissen, T.W., Friedli, M., He, Y., Planet, E., O'Neil, R.C., Markoulaki, S., Pontis, J., Wang, H., Iouranova, A., Imbeault, M., *et al.* (2016). Molecular Criteria for Defining the Naive Human Pluripotent State. *Cell stem cell* *19*, 502-515.

Thompson, S.M., Onwubalili, N., Brown, K., Jindal, S.K., and McGovern, P.G. (2013). Blastocyst expansion score and trophoctoderm morphology strongly predict successful clinical pregnancy and live birth following elective single embryo blastocyst transfer (eSET): a national study. *Journal of assisted reproduction and genetics* *30*, 1577-1581.

Trombetta, J.J., Gennert, D., Lu, D., Satija, R., Shalek, A.K., and Regev, A. (2014). Preparation of Single-Cell RNA-Seq Libraries for Next Generation Sequencing. *Curr Protoc Mol Biol* *107*, 4 22 21-17.

White, M.D., Zenker, J., Bissiere, S., and Plachta, N. (2018). Instructions for Assembling the Early Mammalian Embryo. *Developmental cell* *45*, 667-679.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., *et al.* (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature structural & molecular biology* *20*, 1131-1139.

Methods

Human preimplantation embryos

The use of human embryo donated to research as surplus of IVF treatment was allowed by the French embryo research oversight committee: Agence de la Biomédecine, under approval number RE13-010. All human preimplantation embryos used in this study were obtained from and cultured at the Assisted Reproductive Technology unit of the University Hospital of Nantes, France, which are authorized to collect embryos for research under approval number AG110126AMP of the Agence de la Biomédecine. Embryos used were initially created in the context of an assisted reproductive cycle with a clear reproductive aim and then voluntarily donated for research once the patients have fulfilled their reproductive needs or tested positive for the presence of monogenic diseases. Informed written consent was obtained from both parents of all couples that donated spare embryos following IVF treatment. Before giving consent, people donating embryos were provided with all of the necessary information about the research project and opportunity to receive counselling. No financial inducements are offered for donation. Molecular analysis of the embryos was performed in compliance with the embryo research oversight committee and The International Society for Stem Cell Research (ISSCR) guidelines (Kimmelman et al., 2016).

Human preimplantation embryos culture

Human embryos were thawed following the manufacturer's instructions (Cook Medical: Sydney IVF Thawing kit for slow freezing and Vitrolife: RapidWarmCleave or RapidWarmBlast for vitrification). Human embryo frozen at 8-cell stage were loaded in a 12-well dish (Vitrolife: Embryoslide Ibidi) with non-sequential culture media (Vitrolife G2 plus) under mineral oil (Origio: Liquid Paraffin), at 37°C, in 5% O₂/6% CO₂.

Human embryo time-lapse imaging

Embryos were loaded into the Embryoscope® (Vitrolife®), a tri-gas incubator with a built-in microscope allowing time-lapse monitoring of embryo development. Images were captured on seven focal plans every 15-min intervals using Hoffman modulation contrast (HMC) optical setup¹ and a 635 nm LED as light source as provided in the Embryoscope®. The resolution of the camera is 1280x1024 pixels. The development of each embryo was prospectively annotated as described by Ciray et al., by two trained embryologists undergoing regular internal quality control in order to keep inter-operator variability as low as possible (Ciray et al., 2014). Zona Pellucida (ZP) thickness was measured by our novel analysis pipeline (Feyoux, Reignier et al, *bioRxiv* 2018). The term tM refers to a fully compacted morula. At the blastocyst stage, tSB is used to describe the onset of a cavity formation, tB is used for full blastocyst i.e the last frame before the Zona Pellucida (ZP) starts to thin, tEB for expanded blastocyst, i.e. when the ZP is 50% thinned. Blastocyst contractions and the beginning of herniation were also recorded.

Immunofluorescence of human embryos

Embryos were fixed at the morula, B1, B2, B3, B4, B5 or B6 stages according to the grading system proposed by Gardner and Schoolcraft (Gardner et al., 2000). Embryos were fixed with 4% paraformaldehyde for 5 min at room temperature and washed in PBS/BSA. Embryos were permeabilized and blocked in IF Buffer (PBS–0.2% Triton, 10% FBS) at room temperature for 60 min. Samples were incubated with primary antibodies over-night at 4°C. Incubation with secondary antibodies was performed for 2 hours at room temperature along with DAPI counterstaining. Primary and secondary antibodies with dilutions used in this study are listed in **Table S2**.

Imaging

Confocal immunofluorescence pictures were taken with a Nikon A1 confocal microscope and a 20× Mimm Plan Fluor N.A. 0.8 objective. The images were processed using Fiji (<http://fiji.sc>) and Volocity visualization softwares. Volocity software was used to detect and count nuclei. One embryo was photobleached in order to restrain it. Under Nikon A1 confocal, lasers at 561nm and 640nm were set to

100% power for 10min. The NR2F2 (Rabbit / 488) – SOX17 (Goat / 647) – IFI16 (Mouse – 568) stained embryo was restained with NR2F2 (Rabbit / 488) – GATA4 (Rat / 568) – IFI16 (Mouse – 647).

Single-cell RNA sequencing

Single-cell isolation and overall dataset are presented in **Figure S2**. Single-cell RNA-seq libraries were prepared according to the SmartSeq2 protocol with some modifications (Trombetta et al., 2014). Briefly, total RNA was purified using RNA-SPRI beads. Poly(A)+ mRNA was reverse-transcribed to cDNA which was then amplified. cDNA was subject to transposon-based fragmentation that used dual-indexing to barcode each fragment of each converted transcript with a combination of barcodes specific to each sample. In the case of single cell sequencing, each cell was given its own combination of barcodes. Barcoded cDNA fragments were then pooled prior to sequencing. Sequencing was carried out as paired end 2x25bp with an additional 8 cycles for each index. The FASTQ files were mapped with Hisat2 on GRChH38 genome version, downloaded from ensembl.org. HTSeq (Anders et al., 2015) was used to generate raw counts tables from BAM files, using the matching GTF for the reference genome.

Raw count table treatments

Samples were filtered with the use of the R function *isOutlier* from SCRAN (Lun et al., 2016) library. This function tags samples as outliers with a threshold based on median derivation away from the median of the metric. We filtered samples with two metrics: the number of expressed genes with a threshold of 2 median away derivation from median, and the total number of counts in the sample with a threshold of 3 median away derivation from median. Both metrics were used to discard sample in a two-sided way, below and above the median. This two-sided filter was applied to remove samples carrying too little (93876 counts - 5558 genes) or too much (80431229 counts – 18711 genes) information. Indeed, these are considered as potential doublet of cell. Genes that were expressed in less than two cells and with an average expression less than 0.1 were removed.

The four datasets were then normalized together using the *computeSumFactor* function from SCRAN. Logged and non-logged data were collected using the *normalize* function from scater R library. To compute batch effect free expression, we normalized the data as described above but per dataset. We used mutual nearest neighbors correction implemented by the function *mnnCorrect* to achieve the batch correction from the log-normalized data. The reference dataset that were used for *mnnCorrect* is from Petropoulos et al.

Computation of pseudotime model

Monocle2 needs a subset of gene to make the dimensionality reduction and pseudotimes trajectories. To choose the best set of ordering genes we took samples that passed the quality control from Petropoulos et al. to avoid batch effects. We used SCRAN for processing size factors (normalization factor). We then created an R object with the *newCellDataSet* object used by Monocle2 with the raw expression and the expressionFamily parameter set as "negbinomial.size()". Size factors were attributed according to the SCRAN results. The next step consisted of estimating empirical dispersion of each gene in the negative binomial model with the *estimateDispersions* function. We used the dispersion table function to gather the empirical dispersion and the fitted theoretical dispersion for each gene. We made a ratio of empirical dispersion on the theoretical dispersion for each gene. This ratio describes an over-dispersion score. For a given gene *i* the over-dispersion score *S* is calculated as follows:

$$S_i = \frac{\sigma_{emp\ i}^2}{\sigma_{theo\ \mu_i}^2}$$

Genes with an average log expression < 0.5 across samples were filtered out. Remaining genes were ranked based on their overdispersion score.

Pseudotimes models were generated using a range of top ranked ordering genes from the top 500 to the top 5500. This led to 5000 pseudotime models. For each pseudotime model, A new R object was created with the newCellDataSet function, with the batch corrected expression from the four datasets as input and the expressionFamily parameter set as "gaussianff()". Selected ordering genes were then set as input of Monocle2 algorithm, with the number of resulting dimensions set to three dimensions. An automatic classification of pseudotime models was set up following three criteria based on their topology: number of branches populated by mural TE cells, succession of developmental stage and position and number of branching points.

The most common topologies were: (i) all mural TE cells within one branch, (ii) developmental stages succeeding one other, i.e. morula between 8-cell and blastocysts, (iii) two branching points, (iv) first branching point at E5. The chosen pseudotime model belonged systematically to the most abundant topologies and is calculated from 4484 ordering genes. The resulting 3-dimension pseudotime model was rotated to obtain a 2-dimensional projection.

WGCNA

WGCNA(Langfelder and Horvath, 2008) was performed on batch corrected data using a soft power of 10 with signed Pearson correlation. Resulting module were manually curated to choose a set of 8 modules that were well represented in data and that have distinct behaviors. For each module we use the module eigengene metric that is given by WGCNA to infer the global module expression across the samples. A loess regression of eigengenes by pseudotime was used for **Figure 4A**.

Enrichment analysis

Module enrichment analysis was performed with FGSEA (Sergushichev, 2016). Ranking metric for FGSEA was set as WGCNA gene module membership. This score is processed by Pearson correlation of module eigengene and gene expression. Enrichment was made on five databases: Gene ontology (Cellular Component, Molecular Function and Biological Process), Reactom and KEGG. All retained term were enriched below an adjusted Benjamini-Hochberg p-value of 0.05. A list of transcriptional factors (TF) was downloaded from the RegNetwork database (Liu et al., 2015). Pathway eigengene metric (**Figure S5B**) was processed by taking the first component of a principal component analysis of genes from each enriched term.

Loess regressed expression by pseudotime

We used a Locally Weighted Regression (LOESS) to fit expression in the Pseudotime by cell fate with neighbor impact of 0.75. Expression profiles of common segments were fitted to extract global tendencies. A last LOESS was computed with a low neighbor impact to merge segments to obtain continuous expression curves.

Subdivision of pseudotime branches

The original pseudotime model was constituted by five states, as Monocle2 separates states only by branching point. We subdivided pre-specification and trophoctoderm branch samples by using WGCNA module eigengene. For both branches, WGCNA modules with a Pearson correlation higher than 0.75 with the pseudotime were selected. A loess regression of these module eigengene by pseudotime was performed, followed by a hierarchical clustering of regressed module eigengenes. The clustering was then partitioned. For each branch the best partition was determined at three clusters with the greatest relative loss of inertia method.

Usage of expression per scRNAseq counts table type

Raw expression

estimating gene dispersion and select ordering genes

Normalized expression

projection of gene expression on pseudotime (**Figure 1C, Figure 1E**)

Pseudotime User Interface (**Figure S3G-H**)

Loess regressed expression by pseudotime

heatmap (**Figure 4E**)

expression profile curves (**Figure 4C, Figure S5A, Figure S7**)

Batch corrected logged expression

computation of pseudotime model

WGCNA

Data visualization

ComplexHeatmap (Gu et al., 2016), ggplot2 and d3.js. were used for graphical representation. Hierarchical clustering was done using the Ward criteria and from a correlation distance for the gene/pathway eigengenes, or from the euclidean distance for other metrics.

Mouse single-cell RNA-Seq analysis

Mouse dataset were analyzed in a similar way to human datasets, without batch correction. Alignment step was done from the mm10 version of the genome. Timing of blastulation is corroborated by time-lapse (Feyeux, Reignier et al, biorxiv 2018).

RNA velocity

RNA velocity was performed from BAM of samples that have passed all quality control in the final counts table. First, we used velocity.py using the command *velocity run*, with the parameter `--logic` as "SmartSeq2", and the parameter `-m` (RepeatMasker annotations) as a GTF downloaded from the UCSC genome browser. The global GTF was the same that were used for the computation of raw counts table. Resulting loom files were merged using loompy.combine from loompy python package. We used velocity.R for computing Velocity matrix. Loom files were read with the function *read.loom.matrices*. Then we separated spliced reads matrix, unspliced reads matrix and spanning reads matrix. For each of the matrix a gene filtering was done with the function *filter.genes.by.cluster.expression*. The min.max.cluster.average parameters were set for the corresponding matrix as:

spliced reads matrix: 5

unspliced reads matrix: 1

spanning reads matrix: 0.5

Then RNA velocity was estimated using *gene.relative.velocity.estimate*, with the following parameters: `fit.quantile = 0.05`, `deltaT = 1`, `kCells = 5`.

PCA of **Figure 1G-H** were calculated with the function *pca.velocity.plot*.

Data and software availability

All original data have been deposited on European Nucleotide Archive under accession number PRJEB30442.

The source code can be retrieved by following the links below.

scRNA-Seq alignment pipeline:

https://gitlab.univ-nantes.fr/E114424Z/SingleCell_Align

scRNA-Seq preprocessing and normalization:

<https://gitlab.univ-nantes.fr/E114424Z/singlecellnormalize>

Monocle2 workflow:

https://gitlab.univ-nantes.fr/E114424Z/monocle2_workflow

WGCNA workflow:

<https://gitlab.univ-nantes.fr/E114424Z/WGCNA>

Pseudotime User Interface source code:

<https://gitlab.univ-nantes.fr/E114424Z/PseudoTimeUI>

All other parts of the code are available upon request.

3.3 Discussion

L'analyse des données de scRNA-seq utilisées dans cet article a représenté la majeure partie de mon travail de thèse. Cela m'a amené à développer de nombreux pipelines et scripts d'analyse allant de l'alignement des reads à la création d'une interface web (Figure 19, Figure 20).

Mon pipeline d'alignement des reads (<https://gitlab.univ-nantes.fr/E114424Z/SingleCell Align/>) se base sur HISAT (Kim et al., 2015) et HTSeq (Anders et al., 2015), et a été utilisé pour les 5 jeux de données de scRNA-Seq étudiés. Ce pipeline est adapté aux stratégies de RNA-Seq à transcrits morcelés, tel que le SMART-Seq2 (Picelli et al., 2013), utilisé pour le jeu de données créé pour cet article, le jeu de données de Posfai (Posfai et al., 2017) et le jeu de données de Petropoulos (Petropoulos et al., 2016). Les deux autres jeux de données (Fogarty et al., 2017; Yan et al., 2013) adoptent aussi une stratégie basée sur le morcelage des transcrits. Les données de ce type de séquençage single-cell ne nécessitent pas d'étapes de traitement particulier en analyse primaire, le pipeline d'alignement utilisé est donc classique.

La majeure partie du contrôle qualité se fait à l'étape suivant l'alignement, sur la table de comptage. Ainsi les étapes de filtration et de normalisation se sont révélées primordiales pour la qualité des analyses ultérieures. En effet, les modifications de paramètres de filtrage ou de normalisation entraînent de fortes variations dans les résultats d'inférence de trajectoire. La normalisation (<https://gitlab.univ-nantes.fr/E114424Z/singlecellnormalize>) se base sur la librairie R SCRAN (Lun et al., 2016). Cette partie s'est avérée primordiale pour le reste du travail de bioinformatique, en effet lors de mes toutes premières analyses, aucune librairie permettant de normaliser spécifiquement du scRNA-Seq n'était disponible et nous utilisons DESeq2 (Love et al., 2014) à ces fins. Les résultats obtenus par des analyses multivariées sur ces données ne nous avaient pas permis d'avoir des conclusions rigoureuses sur le développement préimplantatoire humain. Il était en effet impossible d'identifier clairement des populations cellulaires. L'étape de filtration de notre pipeline SCRAN se base sur la distribution du nombre de gènes exprimés et de reads alignés par cellule. Le seuil de filtration se trouve au-dessus et en dessous

de la médiane de chaque métrique, de façon à ne garder que les échantillons contenant de l'information exploitable, tout en minimisant le risque de doublets de cellules.

La modélisation par pseudo-temps a été effectuée avec Monocle2, qui n'autorise que des trajectoires de pseudo-temps continues. La présence de doublets peut ainsi sévèrement diminuer les chances d'obtenir un graphe de pseudo-temps représentatif de la réalité. Des cellules supplémentaires ont été retirées car elles avaient un facteur de normalisation aberrant par rapport à leur total de reads alignés. En effet nous pouvons effectuer une régression linéaire du facteur de normalisation en fonction du nombre total de reads. Dans notre cas, certaines cellules (environ une vingtaine) avaient une distance à la droite de régression anormalement élevée : leur retrait permet d'obtenir de meilleures inférences de trajectoires cellulaires. Cette façon de contrôler la qualité de la normalisation n'a, à ma connaissance, pas été automatisée dans un outil et s'est avérée efficace pour améliorer la qualité du modèle de pseudo-temps. La filtration des gènes est simple. Nous retirons les gènes exprimés dans moins de deux cellules et dont la moyenne de reads alignés est inférieure à 0,1. Nos analyses ultérieures éliminent par elles-mêmes les gènes qui leurs sont peu utiles, nous nous autorisons donc à avoir un filtre peu stringent. La correction de batch entre les différents jeux de données s'est fait en utilisant une *mutual neighborhood correction*. Une correction de batch par régression s'est avérée insuffisante et ne permettait pas de reconstruire des modèles de pseudo-temps cohérents.

La construction du modèle de pseudo-temps en lui-même a représenté le plus long travail de recherche. Le principal paramètre faisant varier les résultats de Monocle2 est le choix des gènes à incorporer dans le calcul de la réduction de dimension et du pseudo-temps. Ces gènes sont appelés les *ordering genes*. Pour déterminer une liste d'ordering genes, nous avons utilisé une méthode basée sur le repérage des gènes dont la dispersion est anormalement élevée par rapport à leur moyenne. En effet la moyenne dépend de la dispersion dans une distribution binomiale négative (**Figure 9**). J'ai ainsi calculé un score de surdispersion pour chaque gène, qui correspond à la dispersion observée (empirique) sur la dispersion théorique pour la moyenne

d'expression du gène. Les n premiers gènes classés par le score de surdispersion sont pris en compte pour la réduction. Le calcul de ce score c'est fait à partir du jeu de données de Petropoulos afin de ne pas subir une dispersion due à des effets de batch.

Nous avons remarqué que des différences de quelques ordering genes parmi des milliers peuvent donner des résultats contradictoires, et qu'il est extrêmement difficile de prévoir les résultats d'un ajout ou d'un retrait de gènes parmi les ordering genes. Cependant, nous avons aussi remarqué que nous retombions souvent sur des modèles proches et qu'il existait donc bien moins de solutions que de combinaisons d'ordering genes possibles. Cette propriété de modifications mineures aboutissant à des changements majeurs et l'existence « d'attracteur » autrement dit de solution récurrente nous pousse à qualifier les résultats de Monocle2 de chaotiques (Lorenz, 1972) en présence de millier d'ordering genes. Nous avons donc choisi de résoudre ce problème de la façon la moins subtile par le calcul de tous les pseudo-temps possibles à partir des ensembles de gènes surdispersés, de rang minimum 1 et pour tous les rangs maximums entre 500 et 5500, ce qui nous donne 5000 modèles de pseudo-temps. Les bornes 500-5500 ont été choisies de façon empirique. En dessous de 500 gènes nous avons une trajectoire sans embranchement donc impossible, au-dessus de 5500 gènes le graphe de pseudo-temps commence à ressembler à un « buisson » comportant des dizaines de branches. Cette méthode nous garantis l'obtention d'un des meilleurs pseudo-temps possible avec nos jeux de données, et surtout d'investiguer les résultats de Monocle2 de manière exhaustive. Nous nous assurons ainsi que la majorité des solutions obtenus par monocle sont cohérentes avec les paramètres biologiques connus, indépendants de l'expression des gènes: morphologie, localisation spatiale, nombre de branches. La rationalisation du choix du modèle de Monocle s'est d'abord faite par l'annotation des 5000 solutions calculées. Chaque solution a tout d'abord été classée selon différents 5 critères :

- l'emplacement de l'embranchement principal ;
- la présence d'embranchement additionnel ;

- l'emplacement de la branche de l'endoderme primitive, qui peut être branché soit sur le trophoctoderme, soit sur l'épiblaste ;
- le respect de la chronologie des jours embryonnaires ;
- la présence de cellules de notre jeu de données annotées comme trophoctoderme dans la branche de la masse cellulaire interne.

Nous avons aussi calculé un score de façon automatisée qui prend en compte le nombre de branches, la distance entre endoderme primitif et épiblaste ainsi que les incohérences entre annotation de dissection et emplacement dans le pseudo-temps. Ce score nous a permis de ranger les modèles par ordre de potentiel intérêt. Le pseudo-temps choisi appartient systématiquement à la classe la plus abondante pour les 5 critères ci-dessus et est calculé à partir de 4484 ordering genes. Il est à noter que nous avons calculé tous nos modèles dans un espace à trois dimensions et non pas à deux, afin de limiter les déformations dues à la compression des dimensions. Nous avons ensuite manuellement « aplati » le modèle final par des rotations des branches. Il s'est avéré que laisser une dimension supplémentaire améliorerait de façon notable les modèles obtenus.

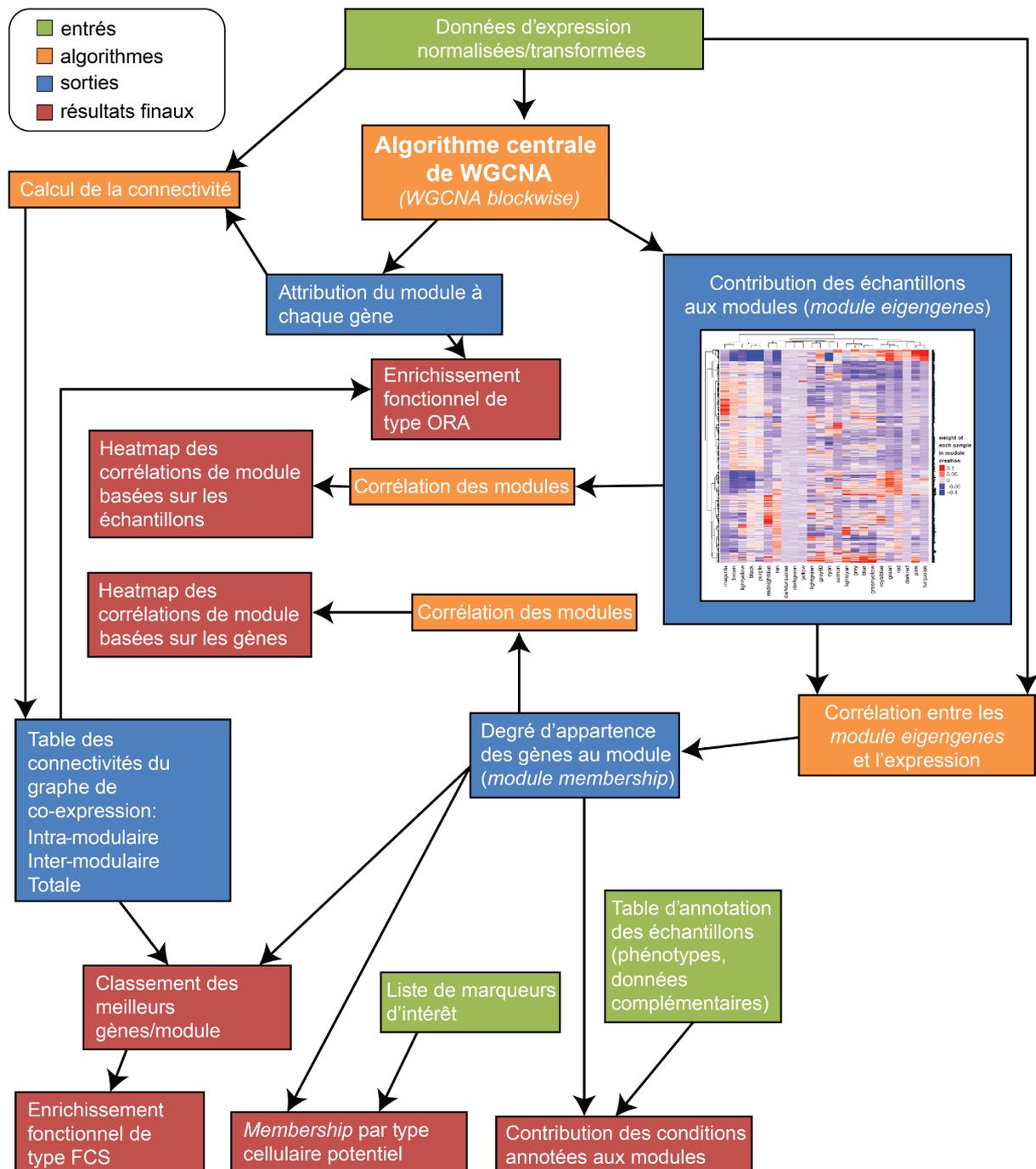


Figure 19 : Description du workflow d'analyse WGCNA. Le code du workflow est disponible à cette adresse : <https://gitlab.univ-nantes.fr/E114424Z/WGCNA/>. Pour le détail sur les différences entre les méthodes d'enrichissement entre ORA et FCS, cf. 1.2.2.9 en page 62.

L'analyse par WGCNA s'est révélée être un excellent complément à l'approche par pseudo-temps (**Figure 19**), malgré le fait que cet outil n'ait pas été écrit pour analyser du single-cell. Le pipeline d'analyse par WGCNA a été écrit pour cette publication et à comme particularité de mettre en relation les variables de la table d'annotation des échantillons et les modules de gènes par le biais d'une matrice de contribution de chaque cellule à chaque module : *les modules eigengenes*. La corrélation entre les

modules eigengenes et l'expression de chaque gène du module nous donne une valeur de représentativité, ce qui permet de hiérarchiser les gènes au sein d'un module. Lorsque cela était possible, nous avons donné un nom à chaque module à partir d'un gène représentatif et connu de la littérature. L'enrichissement fonctionnel par fGSEA (Sergushichev, 2016) a été calculé à partir du score de représentativité. Représenter l'état d'un ensemble de gènes liés à une fonction par une valeur est une tâche ardue. En effet, dans un échantillon une voie de signalisation peut être considérée comme active selon la littérature mais contenir une majorité de gènes sous-régulés. La moyenne d'expression n'est donc pas suffisante pour décrire une voie de signalisation. Nous avons eu l'idée de reprendre la méthode derrière le calcul des modules eigengenes pour représenter l'état des voies de signalisations : nous prenons la première composante d'une analyse en composante principale à partir d'une matrice d'expression centrée, restreinte aux gènes de la voie de signalisation. Nous avons appelé cette métrique le « pathway eigengene » (**Meistermann, Loubersac et al. Supplementary Figure 5B**). Si cette valeur change de signe, cela indique un basculement de la voie de signalisation au sens biologique (active vers inactive ou inactive vers active).

Nous avons effectué une régression des valeurs des modules eigengenes le long du pseudo-temps (**Meistermann, Loubersac et al. Figure 4A**). Avec ces deux informations combinées nous avons pu établir la hiérarchie des vagues d'expression lors du développement préimplantatoire. Le caractère peu supervisé de ces deux analyses est selon nous une vraie force par rapport à des travaux que nous pensons trop susceptibles d'être biaisés par les connaissances injectées dans leurs propres analyses, tel que celle de Stirparo et al., 2018.

L'analyse des vitesses d'ARN nous a permis d'étudier plus précisément la spécification ICM/TE qu'avec le pseudotemps seul. Nous avons utilisé les paramètres recommandés par les auteurs de velocity (La Manno et al., 2018) et obtenus des résultats qui montrent un début de spécification plus précoce que ce que montrent les résultats des analyses transcriptomiques, comme ce que confirme l'analyse par immunofluorescence des jonctions cellulaires. Cependant, cet outil est nouveau et possède un fort potentiel. Nous pouvons penser que ce potentiel ne sera

exploité que dans de futures publications, notamment en termes de modélisation de destin cellulaire.

Une des clefs qui nous a permis de réussir notre analyse est la qualité des annotations de nos cellules, qui nous fait gagner un temps conséquent lors de l'analyse des modèles de pseudo-temps, et nous a permis de réannoter les autres jeux de données (**Meistermann, Loubersac et al. Supplementary Figure 4**). L'apport le plus notable reste la possibilité d'exprimer la temporalité des événements du développement préimplantatoire en stade morpho-cinétique, ce qui résout le problème du développement variable des embryons humains cultivés *in vitro*. D'une façon générale, le manque d'annotation en scRNAseq est un véritable frein à l'interprétation des données.

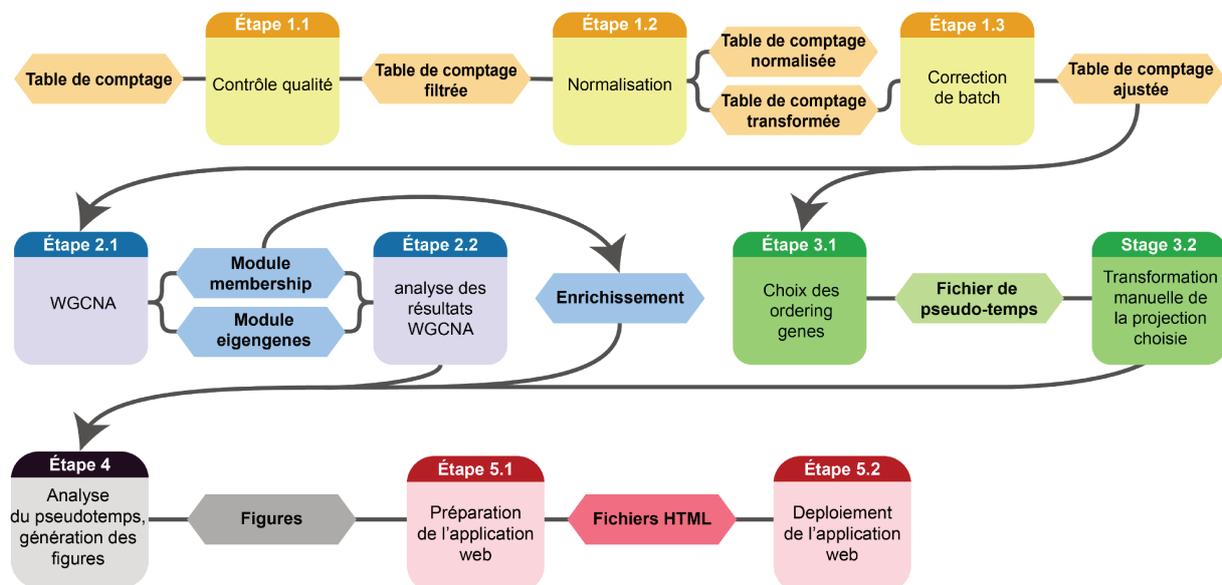


Figure 20 : Vue d'ensemble du workflow d'analyse de Meistermann, Loubersac et al. L'interface web est disponible à l'adresse :

<https://bird2cluster.univ-nantes.fr/demo/PseudoTimeUI/>.

3.4 Conclusion

Notre analyse combinant WGCNA et Monocle2 de quatre jeux de données de scRNA-Seq du développement préimplantatoire a permis de distinguer 9 états cellulaires au cours du développement préimplantatoire humain, dont le transcriptome a été caractérisé par l'activation ou la répression de 8 modules de gènes. Ainsi nous avons fourni une véritable carte temporelle des événements du développement préimplantatoire. Ce travail est donc une ressource de grande valeur pour la communauté scientifique aussi bien en biologie du développement qu'en biologie des cellules souches. Cet aspect est renforcé par le développement d'une interface web permettant de naviguer parmi tous les paramètres de nos analyses, disponible à l'adresse :

<https://bird2cluster.univ-nantes.fr/demo/PseudoTimeUI/>

Ce travail apporte aussi des conclusions nouvelles dans la compréhension du développement préimplantatoire. Nous avons montré qu'il existe trois états de trophoctoderme (TE) successifs : un trophoctoderme exprimant encore des gènes liés à la pluripotence (*5.Early TE*), un trophoctoderme intermédiaire (*8.TE.NR2F2-*) et un trophoctoderme plus avancé marqué par l'expression de *NR2F2* et du module de gènes associé (*9.TE.NR2F2+*). Cet état cellulaire *NR2F2+* est uniquement présent dans les cellules polaires du trophoctoderme après la première spécification (**Meistermann, Loubersac et al. Figure 5**). Une observation importante a été que les blastocystes humains s'attachent à l'épithélium de surface de l'endomètre du côté polaire, c'est-à-dire du pôle du blastocyste qui contient l'ICM (Aberkane et al., 2018; Grewal et al., 2008; Lindenberg, 1991). Cette différence d'état cellulaire entre le TE polaire et mural pourrait expliquer pourquoi le blastocyste s'implante du côté polaire. L'émergence des cellules TE *NR2F2+* du côté polaire pourrait s'expliquer par le dialogue avec l'épiblaste (EPI). Nous avons ainsi mis en évidence une expression restreinte d'IFI16 à l'épiblaste établi (**Meistermann, Loubersac et al. Figure 5D**). IFI16 pourrait faire partie d'un mécanisme de communication entre l'EPI et le TE. Dans ce modèle, les cellules de l'EPI expriment IFI16 et envoient un signal au TE, qui commence ainsi à exprimer *NR2F2* (**Meistermann, Loubersac et al. Figure 5E**). Les cellules de TE ne se spécifient pas en *NR2F2+* et *NR2F2-* : le modèle de pseudo-temps montre que les cellules *NR2F2+* sont plus matures que les *NR2F2-*. Nous pouvons ainsi faire l'hypothèse que l'état *NR2F2+* se propage à tout le trophoctoderme dans le blastocyste postimplantatoire.

Une autre nouveauté concerne la chronologie de la première spécification. Celle-ci avait précédemment été placée à 5 jours après fécondation (Petropoulos et al., 2016), au stade blastocyste. Dans notre modèle, l'embranchement de la première spécification se trouve bien à jour 5 (**Meistermann, Loubersac et al. Figure 2G**), et plus précisément grâce à nos annotations, nous pouvons dire que cet événement se fait entre le stade B2 et B3 du blastocyste. Il existe donc des cellules de blastocyste dans un état non spécifié. Il existe cependant déjà des signes de spécification dans la morula, que ce soit au niveau de la polarisation des jonctions cellulaires (**Meistermann, Loubersac et al. Figure 3a**) ou de la dynamique globale des nouveaux transcrits (**Meistermann, Loubersac et al. Figure 2d**). Nous pensons donc que la première spécification est un processus s'étalant entre la morula et le blastocyste; ce constat permet de réunir la vision transcriptomique et mécanistique du développement préimplantatoire.

3.5 Perspectives

Depuis l'élaboration de ce manuscrit, de nouveaux outils, et des revus d'outils existants ont été publiés dans le domaine du single-cell. Ainsi une revue des méthodes de reconstruction de pseudo-temps (Saelens et al., 2019) a mis en avant des outils comme Slingshot (Street et al., 2018), qui semblent surpasser Monocle2 en terme de performance (**Figure 18**). Monocle lui-même évolue et une troisième version est actuellement en développement avec un algorithme se basant sur la UMAP plutôt que DDRTree pour la réduction de dimension. Du côté de l'analyse de réseaux de gènes, effectuée ici par WGCNA, il semble que les métriques de corrélations soient progressivement abandonnées au profit de l'information mutuelle (Chan et al., 2017; Liang and Wang, 2008). Un outil prometteur est SCENIC (Aibar et al., 2017), qui ajoute une composante fonctionnelle aux clusters de gènes. Ainsi nous n'avons pas des modules mais des régulateurs, dont l'expression est potentiellement régulée par un facteur de transcription.

L'endoderme primitif est un élément qui a particulièrement été difficile à caractériser. Les cellules marquées comme telles dans le modèle final font partie des moins stables dans les différents modèles calculés et ont une plus grande entropie en termes de module de gènes présents (**Meistermann, Loubersac et al. Figure 4A**). De plus le module spécifique du PrE ne contient que 28 gènes. Ces éléments nous montrent que la spécification du PrE n'est potentiellement pas terminée lors du développement

préimplantatoire. Plusieurs éléments pourraient nous aider à résoudre cette question. Le premier est l'incorporation de données plus tardives pour mieux reconstruire les trajectoires cellulaires. Des méthodes pour cultiver des embryons *in vitro* sont maintenant disponibles (Deglincerti et al., 2016; Shahbazi et al., 2016), ce qui permet le séquençage d'embryons postimplantatoires, jusqu'au jour 14. Une équipe a récemment obtenu de telles données (Zhou et al., 2019), de jour 6 à 14. Ces données permettent d'allonger notre vision des trajectoires de destin cellulaire et sont donc susceptibles de répondre à la question du PrE. Des données de spectrométrie de masse, des profils épigénétiques ou encore de nouvelles immunofluorescences de ces cellules pourrait nous confirmer l'existence d'un PrE « bruité », exprimant à la fois des marqueurs du TE et de l'EPI spécifié. La présence de cellules ayant ces caractéristiques est possible. En effet, l'existence d'une expression hautement entropique lors d'un processus de spécification est documenté (Richard et al., 2016). L'utilisation de données postimplantatoires permettrait aussi d'observer la propagation de l'état *TE.NR2F2+* dans tout le trophoctoderme si cette propagation a bien lieu.

Ce travail a permis d'avoir une vision d'ensemble des événements du développement préimplantatoires humain et particulièrement de la spécification ICM/TE. La prochaine étape de notre compréhension de ces événements réside dans des études mécanistiques mettant en évidence les acteurs de la régulation du développement préimplantatoire. L'étude des voies de signalisation actives dans l'embryon par mise en contact avec inhibiteurs/activateurs de ces voies est par exemple une solution envisagée par notre équipe pour aller dans ce sens. Le meilleur candidat pour ce type d'étude concerne le dialogue entre l'EPI et le TE. La première étape consiste à confirmer que c'est bien l'EPI qui cause l'expression de *NR2F2* dans les cellules du TE polaire. La mise en évidence des mécanismes précis impliqués dans ce dialogue serait alors une avancée majeure dans notre compréhension du développement préimplantatoire humain.

Enfin le workflow d'analyse (**Figure 20**) sera au moins en partie implémenté dans une librairie R et fera l'objet d'une future publication. Les futures avancées dans la création de la librairie seront disponibles à l'adresse :

<https://gitlab.univ-nantes.fr/E114424Z/pt2ui>

4 Références

- Aberkane, A., Essahib, W., Spits, C., De Paepe, C., Sermon, K., Adriaenssens, T., Mackens, S., Tournaye, H., Brosens, J.J., and Van de Velde, H. (2018). Expression of adhesion and extracellular matrix genes in human blastocysts upon attachment in a 2D co-culture system. *Mol Hum Reprod* *24*, 375–387.
- Aggarwal, C.C., Hinneburg, A., and Keim, D.A. (2001). On the Surprising Behavior of Distance Metrics in High Dimensional Space. In *Database Theory – ICDT 2001*, J. Van den Bussche, and V. Vianu, eds. (Berlin, Heidelberg: Springer), pp. 420–434.
- Ahlström, A., Westin, C., Reismer, E., Wikland, M., and Hardarson, T. (2011). Trophoctoderm morphology: an important parameter for predicting live birth after single blastocyst transfer. *Hum. Reprod.* *26*, 3289–3296.
- Ahlström, A., Westin, C., Wikland, M., and Hardarson, T. (2013). Prediction of live birth in frozen-thawed single blastocyst transfer cycles by pre-freeze and post-thaw morphology. *Hum. Reprod.* *28*, 1199–1209.
- Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. *Nature Methods* *14*, 1083–1086.
- Alarcon, V.B., and Marikawa, Y. (2018). ROCK and RHO Playlist for Preimplantation Development: Streaming to HIPPO Pathway and Apicobasal Polarity in the First Cell Differentiation. *Adv Anat Embryol Cell Biol* *229*, 47–68.
- Anani, S., Bhat, S., Honma-Yamanaka, N., Krawchuk, D., and Yamanaka, Y. (2014). Initiation of Hippo signaling is linked to polarity rather than to cell position in the pre-implantation mouse embryo. *Development* *141*, 2813–2824.
- Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* *31*, 166–169.
- Andrews, T.S., and Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine* *59*, 114–122.
- Andrews, S., Krueger, F., Segonds-Pichon, A., Biggins, L., Krueger, C., and Wingett, S. (2012). FastQC (Babraham, UK).
- Arce, J.-C., Ziebe, S., Lundin, K., Janssens, R., Helmgaard, L., and Sørensen, P. (2006). Interobserver agreement and intraobserver reproducibility of embryo quality assessments. *Hum. Reprod.* *21*, 2141–2148.
- Assín, R.R. de, Clavero, A., Gonzalvo, M.C., Ramírez, J.P., Zamora, S., Fernández, A., Martínez, L., and Castilla, J.A. (2009). Comparison of methods to determine the assigned value in an external quality control programme for embryo evaluation. *Reproductive BioMedicine Online* *19*, 824–829.

- Assou, S., Boumela, I., Haouzi, D., Monzo, C., Dechaud, H., Kadoch, I.-J., and Hamamah, S. (2012). Transcriptome Analysis during Human Trophectoderm Specification Suggests New Roles of Metabolic and Epigenetic Genes. *PLOS ONE* *7*, e39306.
- Auman, H.J., Nottoli, T., Lakiza, O., Winger, Q., Donaldson, S., and Williams, T. (2002). Transcription factor AP-2 γ is essential in the extra-embryonic lineages for early postimplantation development. *Development* *129*, 2733–2747.
- Bacher, R., Chu, L.-F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. *Nature Methods* *14*, 584–586.
- Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods* *14*, 135–139.
- Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC Bioinformatics* *16*, 334.
- Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour* (Princeton University Press).
- Bengtsson, M., Hemberg, M., Rorsman, P., and Ståhlberg, A. (2008). Quantification of mRNA in single cells and modelling of RT-qPCR induced noise. *BMC Molecular Biology* *9*, 63.
- Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* *57*, 289–300.
- Bessonard, S., De Mot, L., Gonze, D., Barriol, M., Dennis, C., Goldbeter, A., Dupont, G., and Chazaud, C. (2014). Gata6, Nanog and Erk signaling control cell fate in the inner cell mass through a tristable regulatory network. *Development* *141*, 3637–3648.
- Blakeley, P., Fogarty, N.M.E., del Valle, I., Wamaitha, S.E., Hu, T.X., Elder, K., Snell, P., Christie, L., Robson, P., and Niakan, K.K. (2015). Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. *Development* *142*, 3151–3165.
- Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.* *2008*, P10008.
- Bonnaïffoux, A., Herbach, U., Richard, A., Guillemin, A., Gonin-Giraud, S., Gros, P.-A., and Gandrillon, O. (2019). WASABI: a dynamic iterative framework for gene regulatory network inference. *BMC Bioinformatics* *20*, 220.
- Boroviak, T., Stirparo, G.G., Dietmann, S., Hernando-Herraez, I., Mohammed, H., Reik, W., Smith, A., Sasaki, E., Nichols, J., and Bertone, P. (2018). Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common

and divergent features of preimplantation development. *Development* *145*, dev167833.

Boyle, E.I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J.M., and Sherlock, G. (2004). GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* *20*, 3710–3715.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology* *34*, 525–527.

Brons, I.G.M., Smithers, L.E., Trotter, M.W.B., Rugg-Gunn, P., Sun, B., Chuva de Sousa Lopes, S.M., Howlett, S.K., Clarkson, A., Ahrlund-Richter, L., Pedersen, R.A., et al. (2007). Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature* *448*, 191–195.

Buenrostro, J.D., Wu, B., Litzenburger, U.M., Ruff, D., Gonzales, M.L., Snyder, M.P., Chang, H.Y., and Greenleaf, W.J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* *523*, 486–490.

Buettner, F., Natarajan, K.N., Casale, F.P., Proserpio, V., Scialdone, A., Theis, F.J., Teichmann, S.A., Marioni, J.C., and Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat Biotechnol* *33*, 155–160.

Bullard, J.H., Purdom, E., Hansen, K.D., and Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* *11*, 94.

Cacchiarelli, D., Trapnell, C., Ziller, M.J., Soumillon, M., Cesana, M., Karnik, R., Donaghey, J., Smith, Z.D., Ratanasirintrao, S., Zhang, X., et al. (2015). Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. *Cell* *162*, 412–424.

Cannoodt, R., Saelens, W., and Saey, Y. (2016). Computational methods for trajectory inference from single-cell transcriptomics. *European Journal of Immunology* *46*, 2496–2506.

Chan, T.E., Stumpf, M.P.H., and Babbie, A.C. (2017). Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Syst* *5*, 251-267.e3.

Chazaud, C., and Yamanaka, Y. (2016). Lineage specification in the mouse preimplantation embryo. *Development* *143*, 1063–1074.

Chazaud, C., Yamanaka, Y., Pawson, T., and Rossant, J. (2006). Early lineage segregation between epiblast and primitive endoderm in mouse blastocysts through the Grb2-MAPK pathway. *Dev. Cell* *10*, 615–624.

Chen, G., Schell, J.P., Benitez, J.A., Petropoulos, S., Yilmaz, M., Reinius, B., Alekseenko, Z., Shi, L., Hedlund, E., Lanner, F., et al. (2016). Single-cell analyses of X

Chromosome inactivation dynamics and pluripotency during differentiation. *Genome Res.*

Chen, X., Zhang, J., Wu, X., Cao, S., Zhou, L., Wang, Y., Chen, X., Lu, J., Zhao, C., Chen, M., et al. (2014). Trophectoderm morphology predicts outcomes of pregnancy in vitrified-warmed single-blastocyst transfer cycle in a Chinese population. *J. Assist. Reprod. Genet.* *31*, 1475–1481.

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Min* *10*.

Clarke, R., Ressom, H.W., Wang, A., Xuan, J., Liu, M.C., Gehan, E.A., and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat Rev Cancer* *8*, 37–49.

Cole, M.B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., and Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cels* *8*, 315-328.e8.

Costa-Silva, J., Domingues, D., and Lopes, F.M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PLOS ONE* *12*, e0190152.

Croxatto, H.B., Fuentealba, B., Diaz, S., Pastene, L., and Tatum, H.J. (1972). A simple nonsurgical technique to obtain unimplanted eggs from human uteri. *American Journal of Obstetrics & Gynecology* *112*, 662–668.

De Mot, L., Gonze, D., Bessonard, S., Chazaud, C., Goldbeter, A., and Dupont, G. (2016). Cell Fate Specification Based on Tristability in the Inner Cell Mass of Mouse Blastocysts. *Biophys. J.* *110*, 710–722.

De Paepe, C., Cauffman, G., Verloes, A., Sterckx, J., Devroey, P., Tournaye, H., Liebaers, I., and Van de Velde, H. (2013). Human trophectoderm cells are not yet committed. *Human Reproduction* *28*, 740–749.

Deglincerti, A., Croft, G.F., Pietila, L.N., Zernicka-Goetz, M., Siggia, E.D., and Brivanlou, A.H. (2016). Self-organization of the in vitro attached human embryo. *Nature* *533*, 251–254.

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-Cell RNA-Seq Reveals Dynamic, Random Monoallelic Gene Expression in Mammalian Cells. *Science* *343*, 193–196.

Dietrich, J.-E., and Hiiragi, T. (2007). Stochastic patterning in the mouse pre-implantation embryo. *Development* *134*, 4219–4231.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* *29*, 15–21.

- Du, Q.-Y., Wang, E.-Y., Huang, Y., Guo, X.-Y., Xiong, Y.-J., Yu, Y.-P., Yao, G.-D., Shi, S.-L., and Sun, Y.-P. (2016). Blastocoele expansion degree predicts live birth after single blastocyst transfer for fresh and vitrified/warmed single blastocyst transfer cycles. *Fertility and Sterility* *105*, 910-919.e1.
- Duò, A., Robinson, M.D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res* *7*.
- Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *KDD*, p.
- Fanchin, R., Ayoubi, J.M., Righini, C., Olivennes, F., Schönauer, L.M., and Frydman, R. (2001). Uterine contractility decreases at the time of blastocyst transfers. *Hum. Reprod.* *16*, 1115–1119.
- Fardin, P., Moretti, S., Biasotti, B., Ricciardi, A., Bonassi, S., and Varesio, L. (2007). Normalization of low-density microarray using external spike-in controls: analysis of macrophage cell lines expression profile. *BMC Genomics* *8*, 17.
- Fell, D.A., and Small, J.R. (1986). Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem J* *238*, 781–786.
- Fisher, R.A. (1992). Statistical Methods for Research Workers. In *Breakthroughs in Statistics: Methodology and Distribution*, S. Kotz, and N.L. Johnson, eds. (New York, NY: Springer New York), pp. 66–70.
- Fogarty, N.M.E., McCarthy, A., Snijders, K.E., Powell, B.E., Kubikova, N., Blakeley, P., Lea, R., Elder, K., Wamaitha, S.E., Kim, D., et al. (2017). Genome editing reveals a role for OCT4 in human embryogenesis. *Nature* *550*, 67–73.
- Frankenberg, S., Gerbe, F., Bessonard, S., Belville, C., Pouchin, P., Bardot, O., and Chazaud, C. (2011). Primitive Endoderm Differentiates via a Three-Step Mechanism Involving Nanog and RTK Signaling. *Developmental Cell* *21*, 1005–1013.
- Froussios, K., Schurch, N.J., Mackinnon, K., Gierliński, M., Duc, C., Simpson, G.G., and Barton, G.J. (2019). How well do RNA-Seq differential gene expression tools perform in a complex eukaryote? A case study in *Arabidopsis thaliana*. *Bioinformatics* *35*, 3372–3377.
- Galán, A., Montaner, D., Póo, M.E., Valbuena, D., Ruiz, V., Aguilar, C., Dopazo, J., and Simón, C. (2010). Functional genomics of 5- to 8-cell stage human embryos by blastomere single-cell cDNA analysis. *PLoS ONE* *5*, e13615.
- Gardner, R.L. (1982). Investigation of cell lineage and differentiation in the extraembryonic endoderm of the mouse embryo. *J Embryol Exp Morphol* *68*, 175–198.
- Gardner, D.K., and Schoolcraft, W.B. (1999). Culture and transfer of human blastocysts. *Curr. Opin. Obstet. Gynecol.* *11*, 307–311.

Gardner, D.K., Schoolcraft, W.B., Wagley, L., Schlenker, T., Stevens, J., and Hesla, J. (1998). A prospective randomized trial of blastocyst culture and transfer in in-vitro fertilization. *Hum. Reprod.* *13*, 3434–3440.

Gerbe, F., Cox, B., Rossant, J., and Chazaud, C. (2008). Dynamic expression of Lrp2 pathway members reveals progressive epithelial differentiation of primitive endoderm in mouse blastocyst. *Developmental Biology* *313*, 594–602.

Glujovsky, D., Farquhar, C., Quinteiro Retamar, A.M., Alvarez Sedo, C.R., and Blake, D. (2016). Cleavage stage versus blastocyst stage embryo transfer in assisted reproductive technology. *Cochrane Database Syst Rev* CD002118.

Goto, Y., and Takagi, N. (1998). Tetraploid embryos rescue embryonic lethality caused by an additional maternally inherited X chromosome in the mouse. *Development* *125*, 3353–3363.

Grabarek, J.B., Zyzyńska, K., Saiz, N., Piliszek, A., Frankenberg, S., Nichols, J., Hadjantonakis, A.-K., and Plusa, B. (2012). Differential plasticity of epiblast and primitive endoderm precursors within the ICM of the early mouse embryo. *Development* *139*, 129–139.

Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B., and Pierce, E.A. (2011). Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics* *27*, 2518–2528.

Grewal, S., Carver, J.G., Ridley, A.J., and Mardon, H.J. (2008). Implantation of the human embryo requires Rac1-dependent endometrial stromal cell migration. *Proc. Natl. Acad. Sci. U.S.A.* *105*, 16189–16194.

Guo, G., Huss, M., Tong, G.Q., Wang, C., Li Sun, L., Clarke, N.D., and Robson, P. (2010). Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Developmental Cell* *18*, 675–685.

Guo, G., von Meyenn, F., Santos, F., Chen, Y., Reik, W., Bertone, P., Smith, A., and Nichols, J. (2016). Naive Pluripotent Stem Cells Derived Directly from Isolated Cells of the Human Inner Cell Mass. *Stem Cell Reports* *6*, 437–446.

Haghverdi, L., Buettner, F., and Theis, F.J. (2015). Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* *31*, 2989–2998.

Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology* *36*, 421–427.

Heard, E., Chaumeil, J., Masui, O., and Okamoto, I. (2004). Mammalian X-chromosome inactivation: an epigenetics paradigm. *Cold Spring Harb. Symp. Quant. Biol.* *69*, 89–102.

- Hermitte, S., and Chazaud, C. (2014). Primitive endoderm differentiation: from specification to epithelium formation. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* *369*.
- Hirate, Y., and Sasaki, H. (2014). The role of angiotensin phosphorylation in the Hippo pathway during preimplantation mouse development. *Tissue Barriers* *2*, e28127.
- Hirate, Y., Hirahara, S., Inoue, K., Suzuki, A., Alarcon, V.B., Akimoto, K., Hirai, T., Hara, T., Adachi, M., Chida, K., et al. (2013). Polarity-Dependent Distribution of Angiotensin Localizes Hippo Signaling in Preimplantation Embryos. *Current Biology* *23*, 1181–1194.
- Hirate, Y., Hirahara, S., Inoue, K., Kiyonari, H., Niwa, H., and Sasaki, H. (2015). Par-aPKC-dependent and -independent mechanisms cooperatively control cell polarity, Hippo signaling, and cell positioning in 16-cell stage mouse embryos. *Development, Growth & Differentiation* *57*, 544–556.
- Hou, Y., Guo, H., Cao, C., Li, X., Hu, B., Zhu, P., Wu, X., Wen, L., Tang, F., Huang, Y., et al. (2016). Single-cell triple omics sequencing reveals genetic, epigenetic, and transcriptomic heterogeneity in hepatocellular carcinomas. *Cell Res* *26*, 304–319.
- Islam, S., Zeisel, A., Joost, S., Manno, G.L., Zajac, P., Kasper, M., Lönnerberg, P., and Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nat Methods* *11*, 163–166.
- Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* *8*, 118–127.
- Jr, J.H.W. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* *58*, 236–244.
- Kang, M., Piliszek, A., Artus, J., and Hadjantonakis, A.-K. (2013). FGF4 is required for lineage restriction and salt-and-pepper distribution of primitive endoderm factors but not their initial expression in the mouse. *Development* *140*, 267–279.
- Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* *12*, 357–360.
- Korotkevich, E., Niwayama, R., Courtois, A., Friese, S., Berger, N., Buchholz, F., and Hiiragi, T. (2017). The Apical Domain Is Required and Sufficient for the First Lineage Segregation in the Mouse Embryo. *Developmental Cell* *40*, 235–247.e7.
- Krawchuk, D., Honma-Yamanaka, N., Anani, S., and Yamanaka, Y. (2013). FGF4 is a limiting factor controlling the proportions of primitive endoderm and epiblast in the ICM of the mouse blastocyst. *Developmental Biology* *384*, 65–71.
- Krupa, M., Mazur, E., Szczepańska, K., Filimonow, K., Maleszewski, M., and Suwińska, A. (2014). Allocation of inner cells to epiblast vs primitive endoderm in the mouse embryo is biased but not determined by the round of asymmetric divisions (8→16- and 16→32-cells). *Dev. Biol.* *385*, 136–148.

- Kuckenberger, P., Buhl, S., Woynecki, T., Fürden, B. van, Tolkunova, E., Seiffe, F., Moser, M., Tomilin, A., Winterhager, E., and Schorle, H. (2010). The Transcription Factor TCFAP2C/AP-2 γ Cooperates with CDX2 To Maintain Trophectoderm Formation. *Molecular and Cellular Biology* *30*, 3310–3320.
- Kurimoto, K., Yabuta, Y., Ohinata, Y., Ono, Y., Uno, K.D., Yamada, R.G., Ueda, H.R., and Saitou, M. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Res* *34*, e42–e42.
- L. Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol* *17*.
- La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastrioti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. *Nature* *560*, 494–498.
- Lall, S., Sinha, D., Bandyopadhyay, S., and Sengupta, D. (2018). Structure-Aware Principal Component Analysis for Single-Cell RNA-seq Data. *J. Comput. Biol.*
- Lam, T.W., Sung, W.K., Tam, S.L., Wong, C.K., and Yiu, S.M. (2008). Compressed indexing and local alignment of DNA. *Bioinformatics* *24*, 791–797.
- Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* *9*, 559.
- Langfelder, P., Zhang, B., and Horvath, S. (2008). Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* *24*, 719–720.
- Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* *9*, 357–359.
- Lederer, A.R., and La Manno, G. (2020). The emergence and promise of single-cell temporal-omics approaches. *Current Opinion in Biotechnology* *63*, 70–78.
- Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* *25*, 1754–1760.
- Li, Y., and Parast, M.M. (2014). BMP4 regulation of human trophoblast development. *Int J Dev Biol* *58*, 239–246.
- Liang, K.-C., and Wang, X. (2008). Gene regulatory network reconstruction using conditional mutual information. *EURASIP J Bioinform Syst Biol* 253894.
- Lindenberg, S. (1991). Ultrastructure in human implantation: transmission and scanning electron microscopy. *Baillieres Clin Obstet Gynaecol* *5*, 1–14.
- Liu, X., Nefzger, C.M., Rossello, F.J., Chen, J., Knaupp, A.S., Firas, J., Ford, E., Pflueger, J., Paynter, J.M., Chy, H.S., et al. (2017). Comprehensive characterization of distinct

states of human naive pluripotency generated by reprogramming. *Nature Methods* *14*, 1055–1062.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* *15*, 550.

Lun, A.T.L., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* *5*, 2122.

Luo, W., Friedman, M.S., Shedden, K., Hankenson, K.D., and Woolf, P.J. (2009). GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* *10*, 161.

Lyon, M.F. (1961). Gene Action in the X-chromosome of the Mouse (*Mus musculus* L.). *Nature* *190*, 372–373.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research* *9*, 2579–2605.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* *161*, 1202–1214.

Mansour, R., Ishihara, O., Adamson, G.D., Dyer, S., de Mouzon, J., Nygren, K.G., Sullivan, E., and Zegers-Hochschild, F. (2014). International Committee for Monitoring Assisted Reproductive Technologies world report: Assisted Reproductive Technology 2006. *Hum. Reprod.* *29*, 1536–1551.

Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M., and Gilad, Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* *18*, 1509–1517.

Maza, E. (2016). In Papyro Comparison of TMM (edgeR), RLE (DESeq2), and MRN Normalization Methods for a Simple Two-Conditions-Without-Replicates RNA-Seq Experimental Design. *Front. Genet.* *7*.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv:1802.03426 [Cs, Stat].

Meilhac, S.M., Adams, R.J., Morris, S.A., Danckaert, A., Le Garrec, J.-F., and Zernicka-Goetz, M. (2009). Active cell movements coupled to positional induction are involved in lineage segregation in the mouse blastocyst. *Dev. Biol.* *331*, 210–221.

Meistermann, D., Loubersac, S., Reigner, A., Firmin, J., Francois, V., Kilens, S., Lelievre, Y., Lammers, J., Feyeux, M., Hulin, P., et al. (2019). Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophectoderm. *BioRxiv* 604751.

- Mihajlović, A.I., Thamodaran, V., and Bruce, A.W. (2015). The first two cell-fate decisions of preimplantation mouse embryo development are not functionally independent. *Sci Rep* *5*, 1–16.
- Morris, S.A., Teo, R.T.Y., Li, H., Robson, P., Glover, D.M., and Zernicka-Goetz, M. (2010). Origin and formation of the first two distinct cell types of the inner cell mass in the mouse embryo. *Proc. Natl. Acad. Sci. U.S.A.* *107*, 6364–6369.
- Morris, S.A., Graham, S.J.L., Jedrusik, A., and Zernicka-Goetz, M. (2013). The differential response to Fgf signalling in cells internalized at different times influences lineage segregation in preimplantation mouse embryos. *Open Biol* *3*, 130104.
- Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* *5*, 621–628.
- Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* *502*, 59–64.
- Newport, J., and Kirschner, M. (1982). A major developmental transition in early *Xenopus* embryos: II. Control of the onset of transcription. *Cell* *30*, 687–696.
- Niakan, K.K., and Eggan, K. (2013). Analysis of human embryos from zygote to blastocyst reveals distinct gene expression patterns relative to the mouse. *Developmental Biology* *375*, 54–64.
- Nichols, J., and Smith, A. (2009). Naive and Primed Pluripotent States. *Cell Stem Cell* *4*, 487–492.
- Nichols, J., Silva, J., Roode, M., and Smith, A. (2009). Suppression of Erk signalling promotes ground state pluripotency in the mouse embryo. *Development* *136*, 3215–3222.
- Ohnishi, Y., Huber, W., Tsumura, A., Kang, M., Xenopoulos, P., Kurimoto, K., Oleś, A.K., Araúzo-Bravo, M.J., Saitou, M., Hadjantonakis, A.-K., et al. (2014). Cell-to-cell expression variability followed by signal reinforcement progressively segregates early mouse lineages. *Nat Cell Biol* *16*, 27–37.
- Okabe, H., Toh, H., Sato, T., Hiura, H., Takahashi, S., Shirane, K., Kabayama, Y., Suyama, M., Sasaki, H., and Arima, T. (2018). Derivation of Human Trophoblast Stem Cells. *Cell Stem Cell* *22*, 50-63.e6.
- Paternot, G., Wetsels, A.M., Thonon, F., Vansteenbrugge, A., Willems, D., Devroey, J., Debrock, S., D’Hooghe, T.M., and Spiessens, C. (2011). Intra- and interobserver analysis in the morphological assessment of early stage embryos during an IVF procedure: a multicentre study. *Reprod Biol Endocrinol* *9*, 127.
- Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* *2*, 559–572.

- Pera, M.F., de Wert, G., Dondorp, W., Lovell-Badge, R., Mummery, C.L., Munsie, M., and Tam, P.P. (2015). What if stem cells turn into embryos in a dish? *Nature Methods* *12*, 917–919.
- Petropoulos, S., Edsgård, D., Reinius, B., Deng, Q., Panula, S.P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R., and Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell* *165*, 1012–1026.
- Petryszak, R., Burdett, T., Fiorelli, B., Fonseca, N.A., Gonzalez-Porta, M., Hastings, E., Huber, W., Jupp, S., Keays, M., Kryvych, N., et al. (2014). Expression Atlas update—a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments. *Nucleic Acids Res* *42*, D926–D932.
- Picarda, E., Bézie, S., Boucault, L., Autrusseau, E., Kilens, S., Meistermann, D., Martinet, B., Daguin, V., Donnart, A., Charpentier, E., et al. (2017). Transient antibody targeting of CD45RC induces transplant tolerance and potent antigen-specific regulatory T cells. *JCI Insight* *2*.
- Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat Methods* *10*, 1096–1098.
- Plusa, B., Piliszek, A., Frankenberg, S., Artus, J., and Hadjantonakis, A.-K. (2008). Distinct sequential cell behaviours direct primitive endoderm formation in the mouse blastocyst. *Development* *135*, 3081–3091.
- Pollard, M.O., Gurdasani, D., Mentzer, A.J., Porter, T., and Sandhu, M.S. (2018). Long reads: their purpose and place. *Hum Mol Genet* *27*, R234–R241.
- Posfai, E., Petropoulos, S., de Barros, F.R.O., Schell, J.P., Jurisica, I., Sandberg, R., Lanner, F., and Rossant, J. (2017). Position- and Hippo signaling-dependent plasticity during lineage segregation in the early mouse embryo. *ELife* *6*, e22906.
- Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature Methods*.
- Qiu, X., Rahimzamani, A., Wang, L., Mao, Q., Durham, T., McFaline-Figueroa, J.L., Saunders, L., Trapnell, C., and Kannan, S. (2018). Towards inferring causal gene regulatory networks from single cell expression Measurements (Genomics).
- Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* *135*, 216–226.
- Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Gao, N.P., Gunawan, R., Cosette, J., et al. (2016). Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLOS Biology* *14*, e1002585.

- Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* *43*, e47–e47.
- Rivron, N., Pera, M., Rossant, J., Arias, A.M., Zernicka-Goetz, M., Fu, J., Brink, S. van den, Bredenoord, A., Dondorp, W., Wert, G. de, et al. (2018a). Debate ethics of embryo models from stem cells. *Nature* *564*, 183–185.
- Rivron, N.C., Frias-Aldeguer, J., Vrij, E.J., Boisset, J.-C., Korving, J., Vivié, J., Truckenmüller, R.K., Oudenaarden, A. van, Blitterswijk, C.A. van, and Geijssen, N. (2018b). Blastocyst-like structures generated solely from stem cells. *Nature* *557*, 106–111.
- Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* *11*, R25.
- Robinson, M.D., and Smyth, G.K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* *23*, 2881–2887.
- Rossant, J., and Tam, P.P.L. (2017). New Insights into Early Human Development: Lessons for Stem Cell Derivation and Differentiation. *Cell Stem Cell* *20*, 18–28.
- Rotem, A., Ram, O., Shores, N., Sperling, R.A., Goren, A., Weitz, D.A., and Bernstein, B.E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat Biotechnol* *33*, 1165–1172.
- Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2019). A comparison of single-cell trajectory inference methods. *Nature Biotechnology* *1*.
- Sagan, A., and Singer, P. (2007). The Moral Status of Stem Cells. *Metaphilosophy* *38*, 264–284.
- Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* *33*, 495–502.
- Schnabel, L.V., Abratte, C.M., Schimenti, J.C., Southard, T.L., and Fortier, L.A. (2012). Genetic background affects induced pluripotent stem cell generation. *Stem Cell Res Ther* *3*, 30.
- Schrode, N., Saiz, N., Di Talia, S., and Hadjantonakis, A.-K. (2014). GATA6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst. *Dev. Cell* *29*, 454–467.
- Schröter, C., Rué, P., Mackenzie, J.P., and Arias, A.M. (2015). FGF/MAPK signaling sets the switching threshold of a bistable circuit controlling cell fate decisions in embryonic stem cells. *Development* *142*, 4205–4216.
- Schulz, K.N., and Harrison, M.M. (2019). Mechanisms regulating zygotic genome activation. *Nat Rev Genet* *20*, 221–234.

- Sergushichev, A. (2016). An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *BioRxiv* 060012.
- Shahbazi, M.N., Jedrusik, A., Vuoristo, S., Recher, G., Hupalowska, A., Bolton, V., Fogarty, N.N.M., Campbell, A., Devito, L., Ilic, D., et al. (2016). Self-organization of the human embryo in the absence of maternal tissues. *Nat. Cell Biol.* *18*, 700–708.
- Singh, A.M., Hamazaki, T., Hankowski, K.E., and Terada, N. (2007). A Heterogeneous Expression Pattern for Nanog in Embryonic Stem Cells. *STEM CELLS* *25*, 2534–2542.
- Song, L., Langfelder, P., and Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC Bioinformatics* *13*, 328.
- Soumillon, M., Cacchiarelli, D., Semrau, S., Oudenaarden, A. van, and Mikkelsen, T.S. (2014). Characterization of directed differentiation by high-throughput single-cell RNA-Seq. *BioRxiv* 003236.
- Stephen, E.H., and Chandra, A. (1998). Updated projections of infertility in the United States: 1995-2025. *Fertil. Steril.* *70*, 30–34.
- Stirparo, G.G., Boroviak, T., Guo, G., Nichols, J., Smith, A., and Bertone, P. (2018). Integrated analysis of single-cell embryo data yields a unified transcriptome signature for the human preimplantation epiblast. *Development* dev.158501.
- Street, K., Risso, D., Fletcher, R.B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* *19*.
- Subira, J., Craig, J., Turner, K., Bevan, A., Ohuma, E., McVeigh, E., Child, T., and Fatum, M. (2016). Grade of the inner cell mass, but not trophectoderm, predicts live birth in fresh blastocyst single transfers. *Hum Fertil (Camb)* *19*, 254–261.
- Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* *102*, 15545–15550.
- Suter, D.M., Molina, N., Gatfield, D., Schneider, K., Schibler, U., and Naef, F. (2011). Mammalian genes are transcribed with widely different bursting kinetics. *Science* *332*, 472–474.
- Suzuki, R., and Shimodaira, H. (2006). Pvcust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* *22*, 1540–1542.
- Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nat Protoc* *13*, 599–604.
- Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* *126*, 663–676.

- Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* *131*, 861–872.
- Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting transcription factor control circuitry toward ground-state pluripotency in human. *Cell* *158*, 1254–1269.
- Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* *6*, 377–382.
- Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* *448*, 196–199.
- Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic identification of culture conditions for induction and maintenance of naive human pluripotency. *Cell Stem Cell* *15*, 471–487.
- Thompson, S.M., Onwubalili, N., Brown, K., Jindal, S.K., and McGovern, P.G. (2013). Blastocyst expansion score and trophectoderm morphology strongly predict successful clinical pregnancy and live birth following elective single embryo blastocyst transfer (eSET): a national study. *J. Assist. Reprod. Genet.* *30*, 1577–1581.
- Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic stem cell lines derived from human blastocysts. *Science* *282*, 1145–1147.
- Traag, V.A., Waltman, L., and Eck, N.J. van (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* *9*, 1–12.
- Trapnell, C. (2015). Defining cell types and states with single-cell genomics. *Genome Res.* *25*, 1491–1498.
- Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* *7*, 562–578.
- Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods* *14*, 565–571.
- Van den Abbeel, E., Balaban, B., Ziebe, S., Lundin, K., Cuesta, M.J.G., Klein, B.M., Helmgaard, L., and Arce, J.-C. (2013). Association between blastocyst morphology and outcome of single-blastocyst transfer. *Reprod. Biomed. Online* *27*, 353–361.
- Wamaitha, S.E., del Valle, I., Cho, L.T.Y., Wei, Y., Fogarty, N.M.E., Blakeley, P., Sherwood, R.I., Ji, H., and Niakan, K.K. (2015). *Gata6* potently initiates reprogramming

of pluripotent and differentiated cells to extraembryonic endoderm stem cells. *Genes Dev.* *29*, 1239–1255.

Wang, D., and Gu, J. (2018). VASC: Dimension Reduction and Visualization of Single-cell RNA-seq Data by Deep Variational Autoencoder. *Genomics, Proteomics & Bioinformatics* *16*, 320–331.

Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the Waddington landscape and biological paths for development and differentiation. *PNAS* *108*, 8257–8262.

Wang, L., Feng, Z., Wang, X., Wang, X., and Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics* *26*, 136–138.

Wang, T., Li, B., Nelson, C.E., and Nabavi, S. (2019). Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* *20*, 40.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology* *20*, 59.

Wong, K.H., Jin, Y., and Moqtaderi, Z. (2013). Multiplex Illumina sequencing using DNA barcoding. *Curr Protoc Mol Biol Chapter 7*, Unit 7.11.

Xenopoulos, P., Kang, M., Puliafito, A., Di Talia, S., and Hadjantonakis, A.-K. (2015). Heterogeneities in Nanog Expression Drive Stable Commitment to Pluripotency in the Mouse Blastocyst. *Cell Reports* *10*, 1508–1520.

Yamanaka, Y., Lanner, F., and Rossant, J. (2010). FGF signal-dependent segregation of primitive endoderm and epiblast in the mouse blastocyst. *Development* *137*, 715–724.

Yan, L., Yang, M., Guo, H., Yang, L., Wu, J., Li, R., Liu, P., Lian, Y., Zheng, X., Yan, J., et al. (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nature Structural and Molecular Biology* *20*, 1131.

Zhao, Y.-Y., Yu, Y., and Zhang, X.-W. (2018). Overall Blastocyst Quality, Trophectoderm Grade, and Inner Cell Mass Grade Predict Pregnancy Outcome in Euploid Blastocyst Transfer Cycles. *Chin Med J (Engl)* *131*, 1261–1267.

Zhou, F., Wang, R., Yuan, P., Ren, Y., Mao, Y., Li, R., Lian, Y., Li, J., Wen, L., Yan, L., et al. (2019). Reconstituting the transcriptome and DNA methylome landscapes of human implantation. *Nature* *572*, 660–664.

(2011). The Istanbul consensus workshop on embryo assessment: proceedings of an expert meeting. *Hum Reprod* *26*, 1270–1283.

Tableau annexe : résumé des outils utilisés

Outil	Utilisé pour	Utilisable en single-cell	Qualités	Défauts	Référence	Implémentation ou exemple d'utilisation
HISAT2	Alignement	Oui	Bon compromis rapidité/qualité.	Lent pour les larges jeux de données.	Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. <i>Nature Methods</i> 12, 357–360.	https://gitlab.univ-nantes.fr/E114424Z/SingCell_Align
HTSeq	Comptage	Oui	Facilement paramétrable permet d'aligner à partir des SAM/BAM.	Pas de procédure de régularisation statistique du comptage.	Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. <i>Bioinformatics</i> 31, 166–169.	https://gitlab.univ-nantes.fr/E114424Z/SingCell_Align
SCRAN	Normalisation	Single-cell uniquement	Rapide, facile et efficace.	Pas de problème rencontré.	L. Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. <i>Genome Biol</i> 17.	https://gitlab.univ-nantes.fr/E114424Z/SingCellNormalize
DESEQ2	Normalisation & gènes différentiellement exprimé	Non	Méthode fiable et largement utilisée.	Lent pour les larges jeux de données.	Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. <i>Genome Biol.</i> 15, 550.	https://gitlab.univ-nantes.fr/E114424Z/BulkRNAseq
ComBat	Correction de batch	Oui	Paramétrisation du design de l'expérience dans la correction de batch.	Difficile à prendre en main, moins efficace que MnnCorrect pour le single-cell.	Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. <i>Biostatistics</i> 8, 118–127.	Fonction <i>ComBat</i> de la librairie R <i>sva</i>
MnnCorrect	Correction de batch	Oui	Peu corriger des jeux de données très différents.	Argument (k) à paramétrer manuellement. Des cellules de même type cellulaires doivent être présents entre jeu de données.	Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. <i>Nature Biotechnology</i> 36, 421–427.	https://gitlab.univ-nantes.fr/E114424Z/SingCellNormalize
Over-dispersion	Sélection de variable	Oui	Rapide et facile.	Restreint aux données suivant une loi négative binomiale. Sensible à l'influence des populations surreprésentés dans le jeu de données. Besoin de déterminer un seuil.	Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. <i>Genome Biol.</i> 15, 550.	https://gitlab.univ-nantes.fr/E114424Z/veneR/NASeq_R/getMostVariableGenes4
WGCNA	Clustering de gènes par co-expression, clustering de cellules à partir des modules eigengenes	Oui	Efficace sur des jeux de données complexes, permet de simplifier radicalement et efficacement le transcriptome en module de gène.	Lent et gourmand en mémoire. Tri des modules à effectuer a posteriori.	Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. <i>BMC Bioinformatics</i> 9, 559.	https://gitlab.univ-nantes.fr/E114424Z/WGCNA

Monocle2	Réduction de dimension, pseudo-temps, clustering de cellules	Single-cell uniquement	Résultat facilement interprétable.	Paramétrage fastidieux, pas de cycle dans les trajectoires, résultats peu reproductibles.	Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nature Methods.	https://gitlab.univ-nantes.fr/E114424Z/monocle2_workflow
Clustering ascendant hiérarchique sur critère de Ward	Clustering d'échantillon (à partir d'une matrice de distance euclidienne), ou clustering de gène (partir d'une distance de corrélation)	Oui	Méthode universelle, possibilité d'obtenir un dendrogramme pouvant être partitionné.	Impossibilité de déterminer des clusters avec une topologie complexe.	Duò, A., Robinson, M.D., and Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Res 7.	https://gitlab.univ-nantes.fr/E114424Z/veneR/NASeq_R/unsupervisedClustering
Clustering de densité	Clustering de cellule	Oui	Méthode intuitive, possibilité d'obtenir des clusters à la topologie complexe, mise en évidence des échantillons « outliers ».	Paramétrage fastidieux, résultats à affiner à posteriori.	Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In KDD, p.	Fonction <i>dbscan</i> de la librairie R <i>dbscan</i>
Analyse par composante principale	Réduction de dimension linéaire	Oui	Simple, pas de paramètre à déterminer, possibilité de passer de l'espace réduit à l'espace du transcriptome à volonté.	Les deux premières dimensions ne suffisent pas à résumer un jeu de données complexe.	Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572.	https://gitlab.univ-nantes.fr/E114424Z/veneR/projection.R/ACP
UMAP	Réduction de dimension non linéaire	Oui	Efficace sur des jeux de données complexes, possibilité de faire du clustering ou de l'inférence de pseudo-temps sur le résultat. Possibilité d'ajouter des échantillons à la projection.	Paramétrage : en fonction de celui-ci les résultats peuvent être mal interprétés.	McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv:1802.03426 [Cs, Stat].	https://gitlab.univ-nantes.fr/E114424Z/monocle2_workflow
Procédure de Benjamini-Hochberg	Ajustement de p-valeur	Oui	Rapide, simple.	Parfois trop conservateur.	Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 57, 289–300.	Fonction <i>p.adjust</i> de R, avec <i>method="BH"</i>
AUROC (aire sous la courbe ROC)	Détermination de gènes marqueurs d'une population d'échantillons	Oui	Méthode universelle, rapide et efficace.	Sensible aux étapes de contrôle qualités. Le score change en fonction du nombre de type cellulaire présents et de leurs relations.	Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.	https://gitlab.univ-nantes.fr/E114424Z/veneR/NASeq_R/getMarkers2
Over Representation Analysis (ORA)	Enrichissement fonctionnel	Oui	Utilisable avec une liste de gène d'intérêt.	Résultats peu fiables, pas de hiérarchie dans les gènes d'intérêt.	Rao, P.V. (Pejaver V., and 1935- (1998). Statistical research methods in the life sciences (Duxbury Press).	https://gitlab.univ-nantes.fr/E114424Z/veneR/NASeq_R/EnrichFisher
fGSEA	Enrichissement fonctionnel	Oui	Résultats plus fiables qu'une ORA.	Besoin de déterminer un score d'intérêt pour chaque gène.	Rao, P.V. (Pejaver V., and 1935- (1998). Statistical research methods in the life sciences (Duxbury Press).	https://gitlab.univ-nantes.fr/E114424Z/veneR/NASeq_R/Enrich

Liste des publications

Articles en premier auteur

*Kilens, S.**, *Meistermann, D.**, Moreno, D., Chariou, C., Gaignerie, A., Reignier, A., Lelièvre, Y., Casanova, M., Vallot, C., Nedellec, S., et al. (2018). Parallel derivation of isogenic human primed and naive induced pluripotent stem cells. *Nat Commun* 9, 1–13.

*Meistermann, D.**, *Loubersac, S.**, Reigner, A., Firmin, J., Francois, V., Kilens, S., Lelievre, Y., Lammers, J., Feyeux, M., Hulin, P., et al. (2019). Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophectoderm. *BioRxiv* 604751. **Prépublication, en cours de révision dans *Cell Stem Cell*.**

**Co-premier auteur*

Autres articles en co-auteur

Picarda, Elodie, Séverine Bézie, Laetitia Boucault, Elodie Autrusseau, Stéphanie Kilens, Dimitri Meistermann, Bernard Martinet, et al. « Transient antibody targeting of CD45RC induces transplant tolerance and potent antigen-specific regulatory T cells ». *JCI insight* 2, no 3 (2017).

Reignier, A, D Meistermann, D Moreno, S Kilens, J Lammers, P Barriere, L David, et T Freour. « Markers of the first lineage separations in human blastocysts ». In *HUMAN REPRODUCTION*, 32:220–220. OXFORD UNIV PRESS GREAT CLARENDON ST, OXFORD OX2 6DP, ENGLAND, 2017.

Bézie, S., Picarda, E., Boucault, L., Autrusseau, E., Kilens, S., Meistermann, D., Martinet, B., Daguin, V., Donnart, A., Charpentier, E., et al. (2017). Transient Anti-CD45RC mAb Treatment Induces Specific Transplant Tolerance Through Potentiation of Tregs. *Transplantation* 101, S33.

Bézie, S., Meistermann, D., Boucault, L., Kilens, S., Zoppi, J., Autrusseau, E., Donnart, A., Nerrière-Daguin, V., Bellier-Waast, F., Charpentier, E., et al. (2018). Ex vivo

expanded human non-cytotoxic CD8⁺ CD45RC^{low}/- Tregs efficiently delay skin graft rejection and GVHD in humanized mice. *Frontiers in Immunology* 8, 2014.

Feyeux, M., REIGNIER, A., Mocaer, M., Lammers, J., Meistermann, D., Vandormael-Pournin, S., Cohen-Tannoudji, M., Barriere, P., Paul-Gilloteaux, P., David, L., et al. (2018). Development of a robust automated tool for the annotation of embryo morphokinetic parameters. *BioRxiv* 445288. **Prépublication**

Picarda, E., Bézie, S., Boucault, L., Autrusseau, E., Kilens, S., Meistermann, D., Martinet, B., Daguin, V., Donnart, A., Charpentier, E., et al. (2017). Transient antibody targeting of CD45RC induces transplant tolerance and potent antigen-specific regulatory T cells. *JCI Insight* 2.

Rapports préalables à la soutenance



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Nathalie Pinçonnet

Gestionnaire de site ED BS
Direction de la Recherche, des Partenariats et de l'Innovation
Tél : 02.40.41.11.02 (N°interne : 311102)
OUVERTURE au public : lundi mardi et jeudi
9h00-12h00 / 14h00-16h00
Bureau : 202 (2ème étage - UFR Médecine)

Zürich 20th February 2020,

Thesis report for Dimitri MEISTERMANN (David lab, Université de Nantes).

It is my great pleasure to provide my comments and my opinions of the research carried out by Dimitri Meistermann towards his doctoral thesis in the group of Dr. Laurent DAVID, at the University of Nantes.

In his thesis, Dimitri has established a new computational pipeline allowing the analysis of single cell RNA-seq data and used it to analyse published and lab-generated mouse and human scRNA-seq from human preimplantation embryos. He was also able to use his computational skills to compare such scRNA-seq to bulk RNA-seq data from *in vitro* generated Induced pluripotent stem cells. Such analysis allowed him in a first paper to determine different state of pluripotency of human Induced Pluripotent Stem Cells (hiPSCs) and in a second study to identify novel markers for human embryo trophectoderm and epiblast lineages.

Content of thesis: The thesis consists of an “introduction” explaining the state of the art for two important fields of research: human and mouse early development and stem cells Biology in one hand and, the available tools and methods currently used for the analysis of bulk and single cell transcriptomic datasets. Then, Dimitri divided his thesis into two important chapters which are both based on a scientific paper (one published and one in revision). Each of these chapters contains a more specialized introduction part dedicated to the description of the rationale behind the work performed and, an additional discussion and perspective parts on the results obtained. Finally, references are indicated at the end of the thesis.

Presentation, Structure & Coherence: Overall, the thesis is well written in clear, well-formulated French and he has conveyed both his results and ideas in a clear and interesting way.



IMHS Institute of Molecular
Health Sciences

ETH Zurich
HPL G32.1
Otto-Stern Weg 7
CH-8093 Zurich
Switzerland

Dr. Constance Ciaudo
Assistant Professor and Chair of
RNAi and Genome Integrity

Tel. +41-44-633 0858
e-mail: cciaudo@ethz.ch
[http://www.mhs.biol.ethz.ch/
research/ciaudo](http://www.mhs.biol.ethz.ch/research/ciaudo)

The introduction part is well documented with references and illustrations, and complete. The first part concisely describes the first steps of the preimplantation development in human. Then, Dimitri depicts the importance of a better understanding of human early development in the context of the *in vitro* fertilization procedures for couples who faced issues to have kids. The second part of the introduction largely focussed on the description of the computational softwares and statistical models allowing the analysis of transcriptomic data from population to single cell.

The result part is very complete with two full scientific papers with introduction, Material and methods, results and discussion part with in addition a discussion on his contribution and an interesting perspectives part, which will, I assume, open interesting discussion during his defence. All experiments are well presented and controlled, and in my opinion very convincing. Finally, all results and computational approaches have been deposited and are publicly available.

I would like to suggest minor additions to this thesis in order to strength it:

- In my opinion, it will be important to add in the introduction section a part on the ethic consideration according to the fact that the material used here are human embryos and stem cells, as well as the frame allowed by French laws for such work.
- In the second part of the introduction, a paragraph on the library preparation protocols for scRNA-seq will be useful to better comprehend the limitation of the approach, which are then translated at the computational levels. Perhaps some comparison tables of all the tools mentioned in some sections might be useful also to summarise.
- Finally, a general discussion of the thesis will be also of great interest. Several points might be discussed like: the extraction of more information from the currently available scRNA-seq? the correlation between RNA and proteomic information? The integration of other omics approaches in the frame of human early development? the usage of extracted information for IVF patient?

Originality of work and contribution to knowledge: Whilst Dimitri clearly had to face several computational challenges during his thesis, he has managed to produce good and interesting data, which contribute to a better knowledge of the identification of novel markers for human preimplantation embryos. His work provides an important novel resource for the community and his of great importance. His computational pipeline for scRNA-seq analysis and comparison with bulk RNA-seq approach might be considered for one or two additional method papers according to the solid validation presented in this thesis.

Finally, through his PhD work Dimitri was able to acquire a very consequent computational knowledges in a young and developing field, proving his dedication and quality as a computational biologist. This training should be a great asset for his future scientific career.

Summary and Recommendation: The thesis and the work produced by Dimitri are very impressive. I fully support the PhD defence of Dimitri, which should be awarded by a PhD degree after the oral presentation exam.

With very best wishes,

Constance Ciaudo, Ph.D.

A handwritten signature in blue ink, consisting of a large, sweeping loop above the name "ciaudo" written in a cursive, lowercase font.

Université de Nantes
Comue universite Bretagne Loire
Ecole doctorale n° 605
Biologie Santé
Spécialité : Bioinformatique

Antonio Rausell, PhD
Group leader
Clinical Bioinformatics Lab
Institut Imagine
24, boulevard du Montparnasse
75015 Paris
Tlf: +33(0)142754575
antonio.rausell@institutimagine.org

Paris, February 23rd 2020

Object: Avis favorable soutenance de thèse M. Dimitri MEISTERMANN

I am sending my report on the thesis manuscript of M. **Dimitri MEISTERMANN** titled "**Modélisation du développement préimplantatoire humain à partir de données de transcriptome de cellule unique**", co-supervised by Prof Jérémie BOURDON, Laurent DAVID.

The thesis addressed first the establishment and transcriptional characterization of naive human induced pluripotent stem cell lines (hiPSC). Second, pseudotime computational models of cell trajectories of human and mouse embryonic preimplantation were built from single cell RNA-Seq data, followed by experimental validation.

The thesis report is structured in 3 major chapters: first a global introduction, followed by two chapters each corresponding to a specific manuscript, for which an additional introduction, discussion, conclusions and perspectives are provided. The introduction provides the necessary background for the two major components of the thesis work: (i) the cell and molecular biology aspects of human embryonic development, both at pre- and post-implantation stages; here the associated pluripotency characteristics of human Embryonic Stem Cells -as compared to induced Pluripotent Stem Cells (iPSC)- are introduced. And (ii) the bioinformatics aspects of both bulk and single-cell RNA-seq analysis, where a thorough review of the computational pipeline and statistical challenges is provided.

In the second chapter, the manuscript *Parallel derivation of isogenic human primed and naive induced pluripotent stem cells* is presented. Here, the generation of naive human induced pluripotent stem cells (hiPSC) directly from somatic cells (fibroblasts) is reported, notably without transitioning through a primed intermediary state. In these regards, a main aim of the thesis is the transcriptional characterization of such naïve hiPSCs through bulk RNA-seq (DGE-seq) as compared to primed hiPSC. The correspondence of naïve hiPSC with the preimplantation naive epiblast is shown through a comparative analysis of the bulk libraries against single-cell RNA-seq libraries obtained from human preimplantation epiblast. Extensive computational analyses and experimental validation is provided supporting such correspondence.

In the third chapter, the manuscript *Spatio-temporal analysis of human preimplantation development reveals dynamics of epiblast and trophoctoderm specification* is presented. Both human and mouse preimplantation models were derived from pseudotime bioinformatics analysis from human embryos single-cell RNA-seq data. The bioinformatics analysis recovered the hierarchy of differentiation events associated to embryonic preimplantation. Important discoveries were made such as that the first specification in humans is completed only at the blastocyst stage and that the polar trophoctoderm evolves faster than the wall trophoctoderm.

The thesis manuscript is well written, reflects a thorough knowledge of the start-of-the-art in the field as well as an extraordinary scientific quality. The research work presented through the thesis is original and timely, and addresses fundamental questions of human embryonic pre-implantation and of stem cell biology. The computational developments and bioinformatics results presented are methodologically well grounded, and potential follow-up venues are discussed. Overall, the thesis presents important novelties in the competitive area of human stem cell biology and the understanding of the pre-implantation development, and is expected that this work has a wide impact in the research areas of assisted reproduction and regenerative medicine

In terms of scientific production, the thesis translated so far in: (i) one peer-reviewed high-impact publication as co-first author (Kilens, S.*, Meistermann, D.*, et al Nat Commun 9, 1–13), (ii) one BioRxiv preprint as co-first author, currently under review in Cell Stem Cell (Meistermann, D.*, Loubersac, S*., et al. BioRxiv 604751), (iii) 6 other co-authored publications. Notably, commitment to reproducible and open science has been consistently shown throughout the thesis work with the release of source code and pipelines for both bulk and single-cell RNA-seq in different Github repositories.

For all these reasons, I give a favourable recommendation for the defence of Dimitri MEISTERMANN

I remain at your disposal for any further inquiry

With best regards,



Antonio Rausell

Titre : Modélisation du développement préimplantatoire humain à partir de données de transcriptome de cellule unique

Mots clés : Développement préimplantatoire, Bioinformatique, Destin cellulaire, Transcriptome, Cellule unique, Pluripotence

Résumé : Le développement préimplantatoire humain s'étend de la fécondation à la nidation de l'embryon dans la paroi utérine. C'est au cours de cette période que les cellules embryonnaires font leur premier choix de destin cellulaire, passant d'une cellule à un embryon stratifié par trois types cellulaires. Cependant, la séquence d'événements au cours de ces toutes premières spécifications reste inconnu. Pour la comprendre, nous avons tout d'abord établi des lignées induites de cellules souches pluripotente naïves humaines. Grâce à des analyses transcriptomiques, nous avons montré que ces lignées *in vitro* étaient un modèle représentatif de l'épiblaste humain préimplantatoire. Nous avons ensuite construit des modèles *in silico* du développement préimplantatoire humain et murin à partir de données transcriptomiques en cellules uniques.

Le cœur de l'analyse consiste en une inférence des trajectoires cellulaires par un algorithme de pseudo-temps. Nous montrons que la première spécification chez l'homme n'est achevée qu'à partir du stade blastocyste, et non au stade morula comme chez la souris. Enfin le partitionnement du transcriptome en modules de gènes couplé au pseudo-temps permet de décrire les vagues d'expression qui rythment le développement préimplantatoire. Ceci nous a permis de démontrer que le trophoctoderme polaire évolue plus rapidement que le trophoctoderme mural. Ces approches ont contribué de manière significative à notre compréhension du développement préimplantatoire, ouvrant de nouvelles voies de recherche dans les domaines de l'assistance à la procréation et de la médecine régénérative.

Title: Inferring human preimplantation development fate trajectories from single-cell RNA-Seq

Keywords: Preimplantation development, Bioinformatics, Cell fate, Transcriptome, Single cell, Pluripotency

Abstract: Human preimplantation development extends from fertilization to the implantation of the embryo in the uterine wall. It is during this period that embryonic cells make their first choice of cell fate, moving from one cell to an embryo stratified by three cell types in the mature blastocyst. However, the sequence of events during these very first specifications remains unknown. We first established naive human induced pluripotent stem cell lines, and through analyses of transcriptome data, we showed that these lines were a representative cell model of the preimplantation human epiblast. We then constructed models of human and mouse preimplantation development from single cell RNA-Seq from five datasets. The core of the analysis of these data consists of an

inference of cell trajectories by pseudotime analysis. In contrast to the mouse, we show that the first specification in humans is completed at the transcriptomic level only at the blastocyst stage. Finally, the clustering of the transcriptome into gene modules coupled with pseudotime has allowed us to precisely describe the waves of expression that paces pre-implantation development. This allowed us to show that the polar trophoctoderm evolves faster than the wall trophoctoderm. These approaches have contributed significantly to our understanding of pre-implantation development, opening new avenues of research in the fields of assisted reproduction and regenerative medicine.