Développement du séquençage ARN ciblé sur cellules uniques en microfluidique de gouttes et applications

Sophie Foulon

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'Ecole supérieure de physique et de chimie industrielles de la ville de Paris

Développement du séquençage ARN ciblé sur cellules uniques en microfluidique de gouttes et applications

Soutenue par
Sophie FOULON
Le 10 Octobre 2019

Ecole doctorale n° 388
Chimie physique et chimie analytique de Paris centre

Spécialité
Chimie Physique

ESPCI PARIS | PSL

Composition du jury :

Mme Marie-Claude, POTIER Institut de la Moelle et du Cerveau CNRS – UPMC Directeur de Recherche	*Président du jury*
M. Linas, MAZUTIS Vilnius University – Institute of Biotechnology Professor	*Rapporteur*
M. Pascal, BARBRY Université Côte d'Azur – IPMC - CNRS Directeur de Recherche	*Rapporteur*
Mme Céline, VALLOT Institut Curie - CNRS Chargé de Recherche	*Examinateur*
M. Andrew, GRIFFITHS ESPCI Paris Professeur	*Examinateur*
M. Philippe, NGHE ESPCI Paris Maître de Conférences	*Directeur de thèse*

Remerciements

Je remercie chaleureusement Andrew et Philippe pour m'avoir permis de me lancer dans l'aventure de la thèse et pour m'avoir soutenu pendant ces 4 ans, en endossant la lourde responsabilité de directeur de thèse.

Je tiens également à remercier nos partenaires, sur les projets auxquels j'ai pris part. Je remercie l'équipe de Pierre Gressens, pour leur aide et leur soutien, et tout particulièrement Juliette, pour sa patience et sa gentillesse. Je remercie également l'équipe de Nadia Naffakh. J'ai été très heureuse de participer à ce projet prometteur et j'ai pris beaucoup de plaisir à étroitement collaborer avec Kuang-Yu et Mathieu.

Je tiens à remercier mes rapporteurs, Mr Pascal Barbry et Mr Linas Mazutis, d'avoir accepté la tâche fastidieuse de lire ce long manuscrit. Je remercie également Mme Céline Vallot et Mme Marie-Claude Potier d'avoir accepté de faire partie de mon jury. Je remercie enfin notre collaborateur Pierre Gressens d'avoir trouvé le temps de prendre part à la soutenance de thèse en tant que membre invité.

J'ai évolué pendant 5 ans dans ce laboratoire et j'y ai rencontré de très nombreuses personnes fabuleuses. Je tiens à remercier Yacine, avec qui tout a commencé, et bien sûr Baptiste, qui m'a initié à la biologie moléculaire et à bien plus, avec qui j'ai partagé des joies, des frustrations, des fous rires, et finalement beaucoup de merveilleux moments. Je garde en souvenir de cette période faste tous les sages, avec qui nous avons vécu tant d'aventures, à travers le monde, la France, et la rue Mouffetard ; merci à Verouille, Dine, AP, Baba, Yaya, Kev&Lod, Samichel, Rafiki, Toinou et Danio, pour l'énorme impact que vous avez eu sur ma vie.

J'ai évolué dans un laboratoire aussi hétérogène que le plus complexe des tissus. Grâce à vous tous, j'ai pu vivre à l'étranger en plein centre de Paris. Merci *My Dear*, tu es comme une évidence dans ma vie. Je ne me rappelle même plus le laboratoire sans toi. Merci encore pour la Grèce, les cigares au chocolat, ton soutien de tous les jours et ton visage rayonnant quoiqu'il arrive. Merci à Simon, que j'ai adoré embêter et qui me le rendait bien. Cette thèse n'aurait pas non plus été la même sans toi, et pas seulement parce que tu m'as aidé à écrire la plupart de mes scripts d'analyse. Qui dit Simon, dit Kevin, qui s'est bien amusé à me

torturer. Je réfléchis à mettre mon haut napperon juste pour toi le jour de la thèse, et ce même si tu m'as abandonné pour le ménage de culture cellulaire.

Je n'oublie pas les sages du laboratoire : Pablo, toujours là pour régler les problèmes dont personne ne veut s'occuper et m'écouter pleurnicher, Sandeep, qui n'a jamais abandonné l'idée de me faire courir et a toujours été, entre deux moqueries, d'un grand soutien. Merci à Stefie, la plus sage d'entre tous, pour sa gentillesse, sa patience et pour avoir essayé de m'initier au yoga. Merci à Roberta, ma Mamma du LBC, pour ses conseils, sa gentillesse et sa bienveillance. Merci à Andrea, pour son empathie et son soutien, et pour avoir pris part au combat contre les billes. Merci à Matt, qui m'a un jour apporté des chocolats alors que je piquais une crise de nerfs intra-manuscrit. Et question crise de nerfs, le B230, Antoine Rafiki et Stef, n'est pas en reste. Et aujourd'hui Karolis, avec qui je suis si heureuse de partager le bureau, toujours positif, alors que j'ai tendance à voir le négatif tout le temps, tu es mon Yang du B230. Merci pour votre soutien et votre humour à toute épreuve !

Merci à mes stagiaires, Charlotte et Antoine, qui ont été exemplaires, ont su s'adapter à ma désorganisation la plus complète et se montrer travailleur, curieux et extrêmement intelligents et dégourdis.

Bien évidemment, je remercie Isa et Hélène. Le LBC sans vous, c'est un catamaran sans voile, une voiture sans roues, un scRNAseq sans billes d'hydrogel !

Je remercie mes relecteurs de dernière minute pour leur temps et leur efficacité, mes amies de l'UTC, Nath, Alice, Elo, Lulu et Mumu, et à mes super collègues, Antoine, Cyrille, Robin, Pablo, Kévin et Simon.

Je remercie finalement mes proches. Merci à mes parents qui seront toujours pour moi le premier des soutiens. Merci à Jérémy et Tiffany, merci à bébé Lucas qui parvenait à me redonner le sourire même dans les pires moments. Merci à Morgane pour m'avoir aidé à tout oublier sur des danses folles à la Bodega et avoir été toujours à l'écoute. Merci Mai-Thu et Minh-Quyen, d'avoir cru en mon pouvoir de voyance. Et enfin, merci à toi mon Panda. La thèse est réputée pour séparer les couples. Tu es l'exception qui confirme la règle. Je n'aurai pas tenu le choc sans toi. Tu m'as prouvé mille fois combien tu tenais à moi, en ne me faisant jamais de reproches, malgré ma mauvaise humeur et mon indisponibilité. J'espère pouvoir te rendre tout ce que tu m'as donné. Je t'aime.

Liste des abréviations

ADN = Acide Désoxyribo-Nucléique

ADNc = Acide Désoxyribo-Nucléique complémentaire

ARN = Acide Ribo-Nucléique

ARNm = Acide Ribo-Nucléique messager

BC = code-barres

CDR = Région Complémentaire Déterminante

CELseq = *Cell Expression by Linear amplification and Sequencing*

Ct = threshold cycle = cycle seuil

CTC = Cellule Tumorale Circulante

CyToF = *Mass cytometry by time-of-flight* = Cytométrie de Masse par temps-de-vol

DARTseq = *Droplet-Assisted RNA Targeting by single-cell sequencing*

ERCC = External RNA Controls Consortium

FACS = *Fluorescence Activated Cell Sorting* = Tri cellulaire actif par fluorescence

G C = Guanosine :Cytosine

HRP = Horse Radish Peroxydase

IAV = virus Influenza de type A

IVT = Transcription *in vitro*

LNA = *Locked Nucleic Acid* = Acide Nucléique vérouillé

LPS = Lipopolysaccharide

MARS-seq = Massively Parallel RNA single-cell sequencing

M-MLV = *Moloney murine leukemia virus*

NGS = *Next Generation Sequencing* = Technologies de Séquençage de Nouvelle Génération

PA = Polyacrylamide

PBMC = cellules périphériques mononuclées du sang

PBS = Tampon Phosphate Salin

PCR = Polymérisation par Réaction en Chaîne

PDMS = Poly(diméthylsiloxane)

PEG = polyéthylène glycol

PEG DA = Polyéthylène glycol - di acrylate

PFPE = perfluoropolyéther

qPCR = Polymérisation par Réaction en Chaîne quantitative en temps réel

RNAseq = Technologie de séquençage de l'Acide Ribo-Nucléique

RPKM = *Read Per Kilobase of exon model per Million mapped reads*

RT = Rétrotranscription

scATACseq = *single-cell Assay Transposase-Accessible Chromatin using Sequencing*

scChIPseq = *single-cell Chromatin Immuno-Precipitation* = Immunoprécipitation de la Chromatine à l'échelle de la cellule unique

scRNAseq = *single-cell RNA sequencing* = Séquençage de l'ARN à l'échelle de la cellule unique

smRNA FISH = *single-molecule RNA Fluorescence in situ hybridization*

SNC = Système Nerveux Central

SNP = *Single Nucleotide Polymorphism* = Polymorphisme d'un Seul Nucléotide

SNP = Système Nerveux Périphérique

SNS = *Single Nucleus Sequencing* = Séquençage du Noyau Unique

SPLIT-seq = Split Pool Ligation-based transcriptome sequencing

STRT-seq = *Single-cell Tagged Reverse Transcription*

Tm = *Melting Temperature* = Température de fusion

UMI = Identifiant Moléculaire Unique

VH =Variable Heavy = Lourde variable

VL = Variable Light segment = Légère variable

vRNP = ribonucléoprotéine virale

Liste des figures

Liste des tableaux

Table des matières

CHAPITRE 1 : INTRODUCTION

CHAPITRE 2 : MATERIEL ET METHODES

CHAPITRE 3 : OPTIMISATIONS

CHAPITRE 4 : MICROGLIE

CHAPITRE 5 : LES VIRUS

Chapitre1 : Introduction

I. Les technologies cellules uniques

Il y a 200 ans, la cellule a été définie par Schleiden, Schwann et Virchow comme l'unité fonctionnelle et structurale de tout organisme. Dès lors, les biologistes ont identifié des sous-groupes de cellules pour caractériser des processus biologiques distincts. Ces études se faisaient à l'échelle de la population de cellules et ne tenaient pas compte des éventuelles différences pouvant exister au sein même d'une population (Proserpio & Mahata, 2016).

Les analyses populationnelles donnent une image moyennée d'un ensemble de cellules parfois très hétérogène. Pour comprendre un système biologique dans toute sa complexité, les différences entre cellules, leurs interactions, entre elles ou avec l'environnement, et leur évolution face à un changement extérieur, il est nécessaire d'observer chaque cellule individuellement. Chattopadhyay propose une analogie pour comprendre la nécessité d'utiliser des technologies cellules uniques, en combinaison avec une analyse spatio-temporelle, pour la surveillance d'un système cellulaire complexe, le système immunitaire ; imaginons un match de football. La position moyenne du ballon au cours de la partie est le milieu du terrain. Cette mesure nous fait perdre une information essentielle sur la partie, à savoir les moments où le ballon s'est retrouvé au fond d'un but. De même, une observation statique du terrain au moment d'un but ne permet pas de comprendre comment est arrivé cet évènement (Chattopadhyay, Gierahn, Roederer, & Love, 2014).

L'émergence des technologies cellules uniques fut une révolution qui permit de découvrir de nouveaux sous-types cellulaires dans des système complexes tels que le système immunitaire (Jaitin & Kenigsberg & Keren-Shaul *et al.*, 2014) ou le système nerveux (Mrdjen *et al.*, 2018)(Saunders *et al.*, 2018)(Poulin, Tasic, Hjerling-Leffler, Trimarchi, & Awatramani, 2016). Elles ont également amélioré la compréhension des processus biologiques et réseaux d'expression en mettant en évidence des états transitoires d'activation (Karaiskos *et al.*,

2017)(Matcovitch-Natan, Winter, Giladi, Vargas Aguilar, Spinrad, Sarrazin, Ben-Yehuda, David, Zelada González, *et al.*, 2016)(Mathys *et al.*, 2017). Elles se sont finalement révélées d'une grande aide dans le domaine de la recherche biomédicale, en améliorant la connaissance sur la naissance d'une maladie et sa prolifération, en permettant le suivi de son évolution de manière fine et l'établissement de nouvelles stratégies thérapeutiques plus efficaces. En cancérologie, l'hétérogénéité intra-tumorale a pu être déchiffrée (Patel *et al.*, 2014), de même que la dynamique d'émergence de cellules cancéreuses résistantes (Shaffer *et al.*, 2017). Des populations rares, telles que les Cellules Tumorales Circulantes (CTC) peuvent être analysées grâce à ces technologies, permettant d'assurer un suivi non invasif de l'évolution d'un myélome multiple (Lohr *et al.*, 2016). En neurologie, les études cellules uniques ont notamment permis d'identifier un sous-type de microglie, constituant le système immunitaire du cerveau, associé à des maladies neuro-dégénératives telles que la maladie d'Alzheimer, avec un potentiel protecteur (Keren-Shaul *et al.*, 2017).

A. Les technologies -omiques

L'information génétique d'une cellule est contenue dans son ADN sous forme de gènes. Seule une portion de ces gènes est exprimée. Ils sont alors transcrits en ARN puis traduits en protéines. Chaque cellule exprime un panel de gènes différents à un temps donné, définissant son identité. Cette expression n'est pas figée et varie dans le temps, en fonction des interactions avec les autres cellules ou de signaux extérieurs. Il y a donc un contrôle très fin de l'expression des gènes, de sorte à pouvoir répondre rapidement à un besoin (Figure 1). Cette régulation peut se faire à plusieurs niveaux ;

- sur l'ADN via l'apparition de mutations ou la variation du nombre de copies des gènes ;
- via des modifications chimiques au niveau de l'ADN, permettant d'augmenter ou au contraire d'inhiber la transcription de certains gènes ;
- en variant le taux de transcription des gènes ;
- lors de la sélection des variants d'épissage ;
- du transfert des ARN hors du noyau ;
- de la dégradation des ARN ;
- de la traduction des ARN en protéines ;
- ou de l'activation de ces dernières et leur sous-compartimentalisation (Alberts *et al.*, 2002).

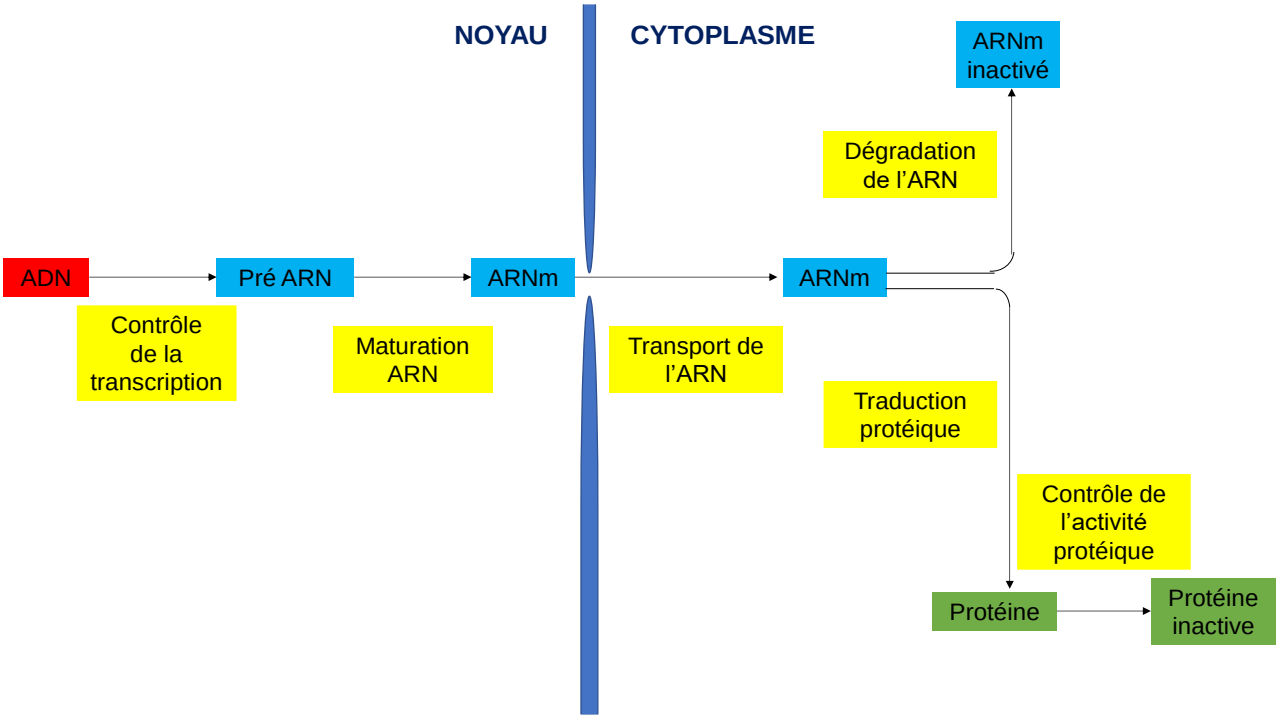


Figure 1 Les différents niveaux de régulation de l'expression d'un gène ; de l'ADN à la protéine. Figure extraite et traduite à partir de Alberts B, Johnson A, Lewis J, et al. Molecular Biology of the Cell. 4th edition. New York: Garland Science; 2002. An Overview of Gene Control.Available from: https://www.ncbi.nlm.nih.gov/books/NBK26885/)

Des technologies à l'échelle de la cellule unique ont été développées à chacun de ces niveaux. La tendance actuelle est d'ailleurs de combiner ces différentes technologies pour obtenir un profil multimodal de chaque cellule, et ainsi améliorer la compréhension des mécanismes de régulation au sein des cellules (Stuart & Satija, 2019)(Hu *et al.*, 2018).

Les analyses génomiques s'intéressent à l'ADN génomique. La recherche de mutations entre échantillon contrôle et échantillon de patients atteints d'un cancer permet de mettre en évidence des réarrangements somatiques à l'origine de la maladie et aide à la conception de traitements thérapeutiques (Campbell *et al.*, 2008). A l'échelle de la cellule unique, des technologies d'analyse du génome entier, telles que la technologie SNS (Single Nucleus Sequencing) ont été mises au point pour l'identification de mutations ponctuelles et la mesure de la variation du nombre de copies des gènes. Ce niveau de résolution à l'échelle de la cellule unique est particulièrement intéressant pour l'étude de l'évolution des tumeurs, qui correspondent à des populations génétiquement très hétérogènes (Navin *et al.*, 2011).

Chez les cellules eucaryotes, l'ADN génomique est présent dans le noyau sous forme de chromatine, une structure ultra compacte formée à partir de l'interaction de l'ADN avec des protéines, les histones. Le nucléosome représente la sous-unité structurelle et fonctionnelle de la chromatine (Kornberg & Thonmas, 1974). Des modifications de la chromatine vont avoir un impact direct sur la régulation de l'expression génique. Les études épigénomiques vont justement s'intéresser à ces modifications et apporter une information phénotypique prédictive permettant de classifier les cellules, tout en aidant à la compréhension des mécanismes de régulation de l'expression des gènes (Bernstein, Meissner, & Lander, 2007). On peut par exemple citer la technologie scATAC-seq (single-cell Assay Transposase-Accessible Chromatin using Sequencing), qui permet de séquencer les partie ouvertes et accessibles de la chromatine à l'échelle de la cellule unique, et ainsi prédire les zones activement transcrites (Buenrostro *et al.*, 2015)(Cusanovich *et al.*, 2018). La technologie scChIPseq quant à elle (pour single-cell Chromatin Immuno-Precipitation soit Immunoprécipitation de la Chromatine à l'échelle de la cellule unique), permet d'analyser les modifications chimiques sur les histones, qui peuvent avoir un impact positif ou négatif sur la transcription, selon le type de modifications (Rotem *et al.*, 2015)(Grosselin *et al.*, 2019).

Les études transcriptomiques s'intéressent aux transcrits, soit à l'expression des gènes sous forme d'ARN. Ce sont celles sur lesquelles je vais me focaliser au cours de cette thèse. Les technologies transcriptomiques à l'échelle de la cellule unique constituent la référence actuelle pour fonctionnellement caractériser les cellules, les différencier et comprendre la dynamique des interactions entre cellules (Wu, Wang, Streets, & Huang, 2017).

Enfin, le protéome correspond à l'ensemble des protéines, soit le produit fonctionnel final de l'expression génique. De ce fait très intéressantes, ces études sont néanmoins limitées à l'analyse d'un nombre restreint de molécules spécifiques, conduisant à une analyse biaisée, et les technologies transcriptomiques à l'échelle de la cellule unique leur sont donc préférées (Svensson *et al.*, 2017). Des progrès ont néanmoins été faits pour augmenter le nombre de protéines analysables en parallèle à l'échelle de la cellule unique. La technologie de cytométrie de masse CyTOF (Kay, Strauss-Albee, & Blish, 2016) permet d'analyser près de 40 marqueurs de surfaces protéiques en parallèle. Elle fait appel à des anticorps porteurs d'isotopes métalliques. L'analyse au niveau protéique ne nécessite pas de récupérer le

matériel génétique contenu à l'intérieur de la cellule (ARN, ADN), impliquant une étape de lyse, et est donc compatible avec une analyse temporelle sur cellule vivante. Le suivi de la sécrétion des protéines en temps réel à l'échelle de la cellule unique, en isolant les cellules dans des compartiments, est par exemple possible (Chattopadhyay *et al.*, 2014)(Eyer *et al.*, 2017)(Han *et al.*, 2012).

B. Utilisation de la microfluidique pour l'analyse à l'échelle de la cellule unique

Les études à l'échelle de la cellule unique ont révolutionné la biologie. Cependant, il convient d'être prudent quant à l'interprétation des données obtenues. En effet, celles-ci sont influencées par les problèmes techniques et les biais biologiques inhérents à l'analyse d'une cellule seule (Prakadan, Shalek, & Weitz, 2017). Si l'on prend l'exemple d'une étude transcriptomique où les ARN messagers de chaque cellule sont quantifiés, le bruit technique est notamment lié à la variabilité dans l'efficacité de capture de l'ARN d'une cellule à une autre, entraînant un résultat différent pour 2 cellules qui sont pourtant identiques. Le bruit biologique peut quant à lui être dû à la variation de la transcription d'un ARN messager dans le temps (éclatement de la transcription) qui fait que l'image d'une cellule prise à un temps T et à un temps T+1, et sans réel changement fonctionnel, sera pour autant différente (Elowitz, Levine, Siggia, & Swain, 2002) (Raj, van den Bogaard, Rifkin, van Oudenaarden, & Tyagi, 2008). On parle de bruit intrinsèque ou encore stochastique (phénomène qui relève du hasard et non déterministe). Augmenter le nombre de cellules étudiées permet de réduire l'impact de ces bruits. Un autre moyen de limiter le poids du bruit est d'effectuer des études sur plusieurs variables en parallèle (en s'intéressant à la co-expression de plusieurs gènes) ou encore à plusieurs niveaux en même temps, transcriptomique et protéomique par exemple.

De nouveaux outils ont été mis au point pour permettre d'augmenter le nombre de cellules analysées en parallèle tout en réduisant les coûts, les outils microfluidiques. La microfluidique se définit comme la science de la manipulation des fluides à l'échelle micrométrique (Whitesides, 2006). Les outils microfluidiques permettent d'isoler des cellules uniques physiquement, dans un volume réduit de l'ordre du picolitre au nanolitre, soit d'un facteur 1000 à 1 000 000 par rapport aux micro-puits classiques (Mazutis *et al.*, 2013). Cela entraîne une diminution des réactifs et consommables et donc des coûts (Agresti *et al.*, 2010). La cellule, ainsi compartimentée, peut-être lysée et libérer son contenu pour procéder à des analyses génomiques et transcriptomiques de cellules uniques. De même, on peut procéder à des études protéomiques telles que le suivi de la sécrétion de protéines par une cellule unique, contrairement au FACS qui ne permet de s'intéresser qu'aux marqueurs de surfaces.

De plus, du fait du très faible volume de chaque compartiment, les molécules sont concentrées, permettant une détection plus fine et plus précoce d'analytes d'intérêt (Köster *et al.*, 2008). Ce volume réduit offre aussi l'avantage d'augmenter l'efficacité réactionnelle, notamment pour l'amplification de molécules ADN ou la rétrotranscription de transcrits (même peu abondants) en ADNc (Streets *et al.*, 2014)(Marcy *et al.*, 2007). Enfin, ces outils permettent de réaliser des expériences miniaturisées en parallèle en très grand nombre. Le débit est très fortement augmenté, ce qui accroît la confiance dans les interprétations faites (Prakadan *et al.*, 2017).

Intéressons-nous à 3 outils microfluidiques majeurs ; les valves (Figure 2 A), les nano puits (Figure 2 C) et les microgouttes (Figure 2 B) et aux avantages et inconvénients offerts par chacun de ces systèmes.

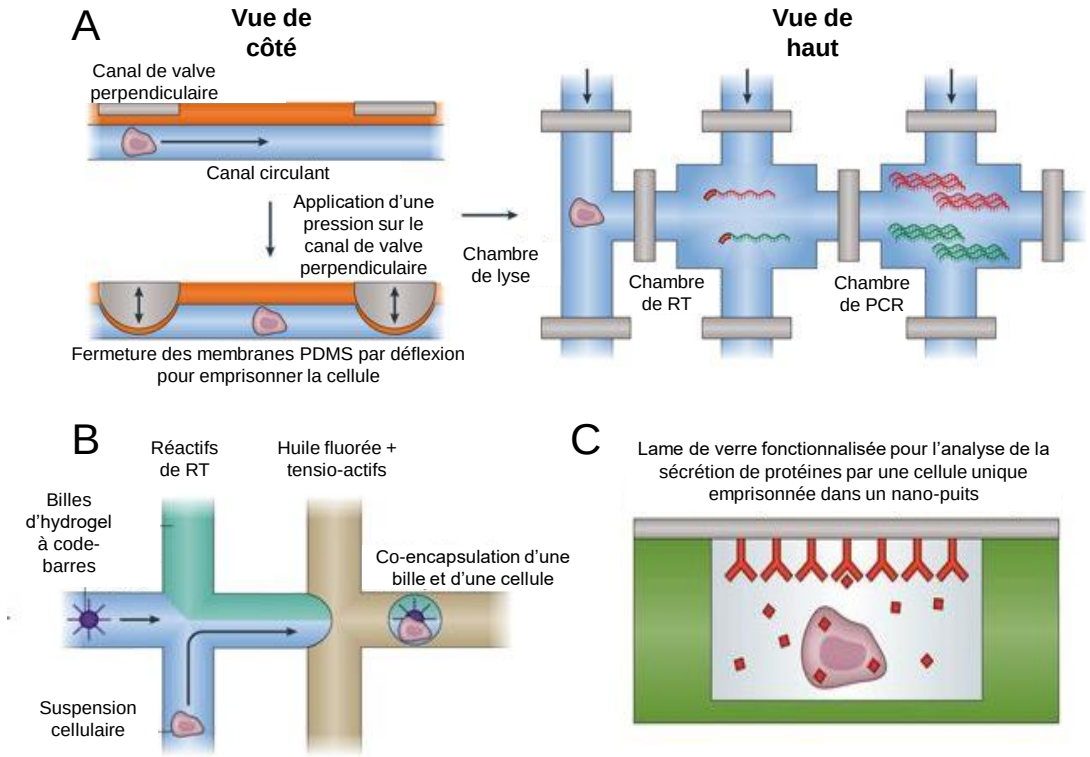


Figure 2 Exemple d'applications des 3 dispositifs microfluidiques majeurs ; A= système de circuits microfluidiques intégrés à valves pour l'amplification du transcriptome complet. Des valves sous le contrôle de pompes à pression définissent différents compartiments où vont successivement s'effectuer la capture de cellules uniques et leur lyse, la rétrotranscription des ARN libérés puis l'amplification par PCR des ADNc néosynthétisés, B= Séquençage de l'ARN à l'échelle de la cellule unique en utilisant la microfluidique en gouttes. Des cellules et des billes d'hydrogel porteuses d'amorces à code-barres sont encapsulées dans des gouttes, dispersées dans de l'huile fluorée. La cellule est lysée une fois dans la goutte et ses ARN messagers sont capturées par les amorces apportées par la bille. C= Nano-puits dans lesquelles sont confinées des cellules par gravité, et sellés par une lame de verre ; Cette dernière peut être fonctionnalisée avec des anticorps, par exemple, pour permettre la détection et quantification de protéines sécrétées par une cellule. Figure adaptée et traduite à partir de (Prakadan *et al.*, 2017)

1. Les valves

Il s'agit des premiers dispositifs microfluidiques développés. D'une grande complexité et faits d'un élastomère souple (tel que le Poly(diméthylsiloxane) ou PDMS), ils permettent d'emprisonner des cellules uniques, dans des compartiments de quelques nanolitres de volume, délimités par des valves qui s'ouvrent et se ferment sous le contrôle de pompes à pression. Grâce à cette capacité, de nombreuses opérations peuvent être réalisées dans chaque compartiment, telles que l'ajout de réactifs, l'incubation, le lavage et la détection d'analytes et enfin la récupération des molécules synthétisées. Cet outil est tout particulièrement indiqué quand il est nécessaire d'avoir un contrôle très précis des paramètres externes. En revanche, il n'est pas adapté à l'étude d'un échantillon large, le nombre de compartiments étant limité à quelques dizaines de milliers au maximum, et en pratique le plus souvent à quelques centaines (Mazutis *et al.*, 2013) (Chattopadhyay *et al.*, 2014). La fabrication de ces dispositifs est également assez complexe. Elle fait appel à des techniques de lithographies souples en couches multiples (Unger, Chou, Thorsen, Scherer, & Quake, 2000). A cause du phénomène de bio-encrassement des parois, ils ne sont réutilisables qu'un nombre limité de fois.

Ces outils ont été développés pour des études aussi bien génomiques que transcriptomiques ou protéomiques. Mis au point par le groupe de Steven Quake dans les années 90, ces systèmes de circuits intégrés microfluidiques à valves sont aujourd'hui commercialisés par Fluidgm et offrent un large panel d'applications.

On peut par exemple citer le système C1 de Fluidgm qui permet de capturer 96 cellules dans des compartiments, et de les observer au microscope avant de les lyser pour procéder à une analyse de leur transcriptome. Ce système a ainsi permis en 2014 d'étudier la signalisation paracrine entre oligodendrocytes (Shalek *et al.*, 2014).

Le système des valves offre donc de nombreux avantages, tels que la compartimentalisation dans des petits volumes, un contrôle précis des valves et la possibilité de procéder à de nombreuses opérations complexes, ou encore la possibilité d'observer au microscope les cellules capturées pour s'assurer qu'il n'y ait bien qu'une seule cellule.

En revanche, les circuits fluidiques intégrés sont coûteux du fait de leur grande complexité de fabrication. De même, la capture des cellules à l'intérieur des compartiments nécessite de grandes quantités de cellules (1000 cellules sont nécessaires pour en capturer une), ce qui n'est pas compatible avec des travaux sur des échantillons rares. Pour autant, le nombre de compartiments est limité, ne rendant pas non plus possible une étude à très haut débit.

Enfin, la capture ne sera réellement efficace que pour un échantillon dont la taille des cellules est relativement homogène, étant donné que les compartiments sont proposés dans 3 gammes de taille assez restreinte ; les cellules trop adhérentes ou à la forme non sphérique présentent aussi des taux de capture relativement faibles.

Malgré ces nombreuses qualités, du fait des inconvénients listés précédemment (Kolodziejczyk, Kim, Svensson, Marioni, & Teichmann, 2015b), cette technologie a été peu à peu délaissée par les utilisateurs au profit d'autres systèmes microfluidiques, à savoir les nano-puits et les gouttelettes.

2. Les nano-puits

Les nano-puits (également appelés micro-puits) sont des dispositifs en PDMS, ou encore verre, contenant des milliers de puits aux dimensions microscopiques. Les cellules sont déposées au fond de ces compartiments par gravité selon une loi statistique de Poisson (Ahrberg, Lee, & Chung, 2018) ;

$$P_\lambda(k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

Où P est la probabilité d'avoir k cellules dans un puits et λ est le nombre moyen de cellules par puits.

Tous les compartiments ne possèdent pas de cellules, et la densité cellulaire de la solution déposée doit être calculée, de sorte à limiter les compartiments avec plus d'une cellule par puits. Ces puits ne sont normalement pas fermés mais ils peuvent néanmoins être scellés à l'aide d'une lame en verre fonctionnalisée par exemple, ajoutant une dimension protéomique à l'étude, ou encore avec de l'huile, limitant les risques de contamination d'un puits à l'autre.

Ce dispositif microfluidique a pour principal avantage sa simplicité. Il ne nécessite pas l'utilisation d'équipements coûteux. De plus, ils sont compatibles avec des échantillons rares, comptant très peu de cellules.

Enfin, il est possible d'imager les puits pour vérifier la présence des cellules et d'éventuels doublets. En revanche, ils offrent un débit moins grand que les systèmes en gouttes, et permettent moins de manipulations et de contrôle que les systèmes de valves (Prakadan *et al.*, 2017).

3. Les microgouttelettes

La microfluidique en gouttes permet d'encapsuler des cellules et réactifs dans des gouttes de quelques picolitres à quelques nanolitres, en poussant à l'aide de seringues ou de pompes à pression, des phases aqueuses et huileuses dans des puces microfluidiques.

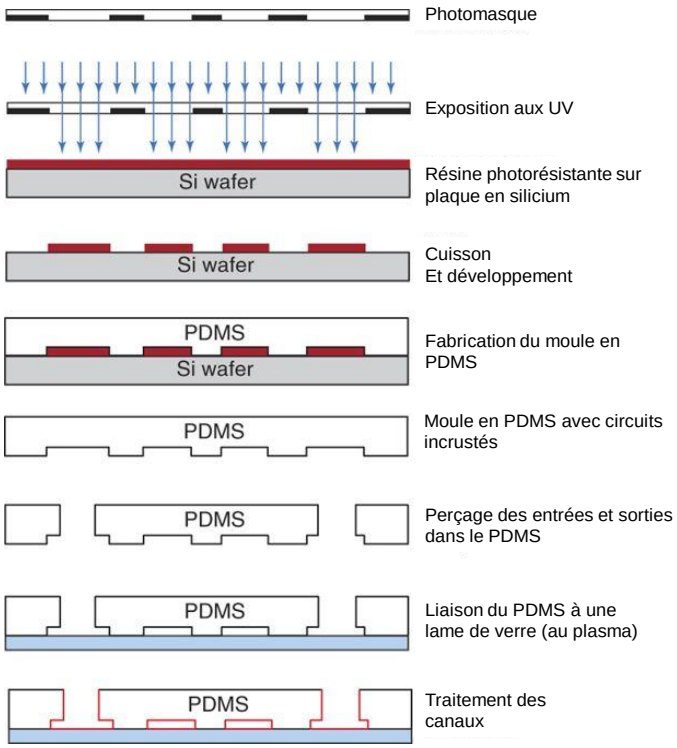


Figure 3 Schéma de la fabrication d'une puce microfluidique par lithographie souple, extrait et traduit à partir de (Mazutis *et al.*, 2013)

Les puces microfluidiques sont faites en utilisant des techniques de photolithographie. Des circuits microfluidiques sont ainsi incrustés dans un support souple en PDMS, qui est collé à une lame de verre par un traitement au plasma pour permettre la circulation des fluides dans les canaux micrométriques, après perçage préalable des entrées et sorties dans le PDMS. Ces circuits sont dessinés au préalable sur logiciel de conception assisté par ordinateur, en suivant des règles précises, et sont imprimées en relief, à une hauteur micrométrique précise, sur un disque de silicium en utilisant les techniques de lithographie souple. Plusieurs couches successives peuvent être alignées et imprimées pour assurer des hauteurs différentes au sein d'un même dispositif dans certains cas particuliers. Le PDMS est coulé sur ce disque ainsi modifié, permettant d'incruster les circuits dans l'élastomère (Figure 3) (Mazutis *et al.*, 2013).

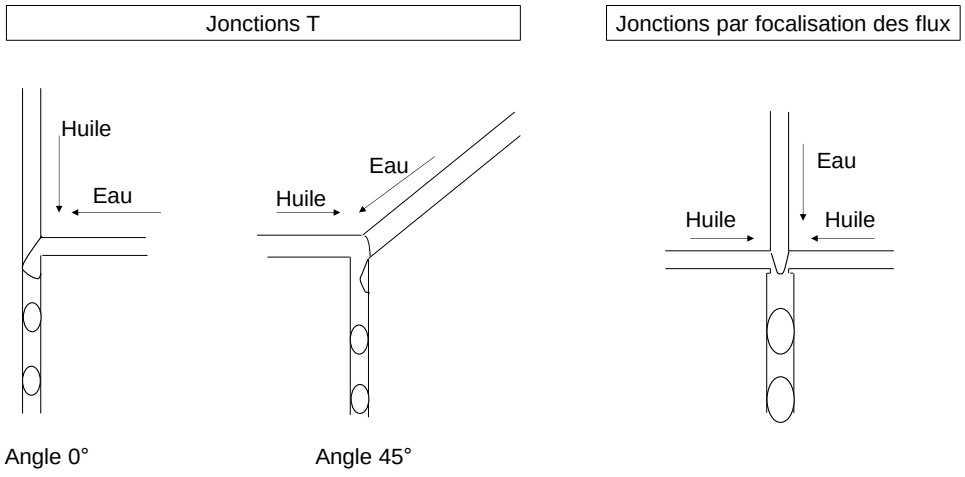


Figure 4 Schéma de différentes géométries pour la formation de gouttes d'eau dans l'huile (Gauche = Jonction T, Droite = Jonction par focalisation des flux), inspiré à partir de (A. R. Abate *et al.*, 2009)

Il existe plusieurs géométries permettant la formation de gouttes d'eau dans l'huile (Figure 4), telles que la jonction T, où les flux aqueux et huileux se croisent selon un certain angle, ou encore la jonction par focalisation des flux, où l'huile est subdivisée en 2 canaux symétriques qui croisent le flux aqueux de part et d'autre.

Cette dernière va souvent être préférée car elle permet de produire des émulsions mono-disperses (toutes les gouttes sont produites à la même taille) stables à des ratios flux huileux/flux aqueux modérés et à haute fréquence (A. R. Abate *et al.*, 2009). La faculté de produire des émulsions mono-disperses est essentielle pour assurer des analyses quantitativement comparables d'une goutte à l'autre.

En utilisant ces puces, des microgouttelettes sont produites à des fréquences de l'ordre du kilohertz. Ce très haut débit rend possible la parallélisation de millions d'expériences. Cela fait de ce dispositif microfluidique celui au plus grand débit et donc le dispositif de choix lorsqu'un large échantillon doit être analysé.

Chaque goutte constitue un compartiment à part entière, au même titre que le puits d'une plaque. Ces gouttes aqueuses sont dispersées dans une huile fluorée, une huile biocompatible qui offre l'avantage de dissoudre l'oxygène 20 fois plus vite que l'eau et qui permet donc de maintenir des cellules encapsulées dans des gouttes dispersées dans l'huile, en vie pendant plusieurs jours (K. C. Lowe, 2002). Très hydrophobe, l'huile fluorée représente une barrière pour les molécules organiques solubles en phase aqueuse, ce qui permet d'éviter les fuites d'une goutte à l'autre (Simons & Linevsky, 1952).

La stabilité des gouttes dans le temps et face aux changements de température est assurée par un tensio-actif, une molécule amphiphile constituée d'une queue hydrophobe et d'une tête hydrophile, qui se place à l'interface huile-eau pour en réduire les tensions de surface (Baret, 2012). Ce tensio-actif permet de limiter les phénomènes de coalescence entre gouttes et d'ainsi conserver la compartimentation initiale des réactions. Dans le cas des huiles fluorées, il convient d'utiliser des tensio-actifs à queue fluorée ; une queue de type perfluoropolyéther (PFPE) montre une grande solubilité dans l'huile fluorée. Bien que ces queues soient potentiellement non biocompatibles, l'ajout d'une tête hydrophile de type polyéthylène glycol (PEG) par exemple, qui se place au contact de la phase aqueuse, assure la biocompatibilité. Des cellules HEK293T ont été déposées sur une couche de d'huile fluorée en présence ou non de différents tensio-actifs et incubées pendant 48h. Les tensio-actifs de type PFPE à tête PEG montrent une bonne biocompatibilité, n'affectant pas l'intégrité des membranes cellulaires ni la prolifération des cellules (Clausell-Tormos *et al.*, 2008).

La production de microgouttes de façon mono-disperse à ultra-haut débit permet donc d'isoler des réactions, d'effectuer de très nombreuses études en parallèle de manière quantitative, et de travailler dans des volumes réduits, à l'échelle de la cellule unique voire de la molécule unique (Theberge *et al.*, 2010).

L'encapsulation des cellules ou particules en gouttes est passive et suit une loi de Poisson selon l'équation ;

$$P_\lambda(k) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

Où P est la probabilité d'avoir k cellules par goutte et λ est le nombre moyen de cellules dans le volume d'une goutte (Figure 5).

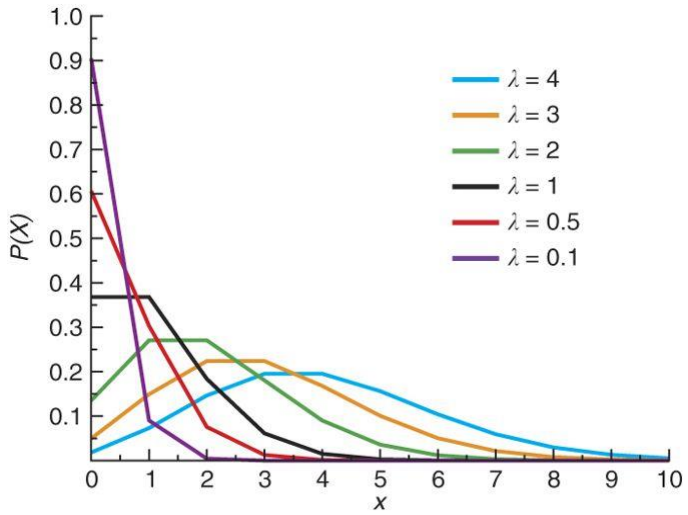


Figure 5 Probabilité d'avoir X cellules par goutte à différentes valeurs de lambda, extrait de (Mazutis *et al.*, 2013)

En suivant cette équation, le λ peut être ajusté en contrôlant la densité de cellules de sorte à avoir une probabilité d'avoir plus de 1 cellule par goutte très faible. Typiquement, les études à l'échelle de la cellule unique se place à un λ de 0.1, avec 9% des gouttes contenant une cellule unique, 90% des gouttes étant vides, et moins de 1% des gouttes contenant plus

d’une cellule (Köster *et al.*, 2008). Etant donné les fréquences de production des gouttes et le taux d’encapsulation, 10 000 cellules peuvent être encapsulées en quelques minutes. Ce chiffre est cohérent, étant donné les contraintes liées aux performances des séquenceurs.

Des stratégies existent néanmoins pour augmenter la proportion de gouttes contenant une cellule unique, par exemple en ordonnant les cellules avant qu’elles arrivent à la zone de génération des gouttes et en synchronisant les flux (Theberge *et al.*, 2010).

Dans le cadre des études transcriptomiques à l’échelle de la cellule unique, les transcrits provenant d’une même cellule sont habillés d’une séquence code-barres unique, de sorte à retrouver leur origine après avoir regroupés les transcrits de toutes les cellules. Ce code-barres unique de goutte est apporté dans les différentes gouttes par une bille d’hydrogel. Chaque bille est recouverte par des millions de molécules à code-barres, et ce dernier est le même pour une bille donné mais différent d’une bille à l’autre.

De sorte à éviter une double encapsulation selon la loi de Poisson à la fois des billes et des cellules, les billes utilisées sont déformables et hyper-concentrées. Ces dernières sont distribuées de manière contrôlée, et en ajustant la fréquence de formation des gouttes à ce flux de billes, de sorte à aboutir à un taux d’encapsulation de billes uniques bien plus grand que celui attendu selon une loi de Poisson (Adam R Abate, Chen, Agresti, & Weitz, 2009). Cette stratégie est employée par les technologies inDrop et 10x Chromium, aboutissant à un taux d’encapsulation des billes à codes-barres de l’ordre de 80% (X. Zhang *et al.*, 2019).

Ces gouttes sont manipulables (Kintses, van Vliet, Devenish, & Hollfelder, 2010)(Mazutis *et al.*, 2013), permettant de réaliser des expériences complexes en plusieurs étapes successives (Figure 6).

Les gouttes peuvent ainsi être :

- fusionnées, passivement (Mazutis, Baret, & Griffiths, 2009) ou activement par électro-coalescence (Chabert, Dorfman, & Viovy, 2005) ;
- divisées (D. R. Link, Anna, Weitz, & Stone, 2004) passivement ou activement, avec l'utilisation d'un champ électrique (Darren R Link *et al.*, 2006) ;
- des réactifs peuvent être ajoutés à chaque goutte par pico-injection (Adam R Abate, Hung, Mary, Agresti, & Weitz, 2010) ;
- incubées ;
    - dans la puce microfluidique ; à l'intérieur d'une ligne de délai (Frenz, Blank, Brouzes, & Griffiths, 2009), ou dans des chambres (Hatch *et al.*, 2011) ;
    - hors de la puce, en tube scellé (pour limiter tout contact avec l'air pouvant augmenter le phénomène de coalescence) ou encore en chambre, permettant le suivi des gouttes en temps réel (Eyer *et al.*, 2017) ;
- analysées en temps réel avec mesure de la fluorescence émise dans chaque goutte passant sous une ligne laser (Mazutis *et al.*, 2013) ;
- triées ;
    - passivement, sur la base de leurs propriétés physiques (Chabert & Viovy, 2008) ;
    - activement en utilisant la di-électrophorèse (Agresti *et al.*, 2010) ou les ondes acoustiques (Franke, Abate, Weitz, & Wixforth, 2009)

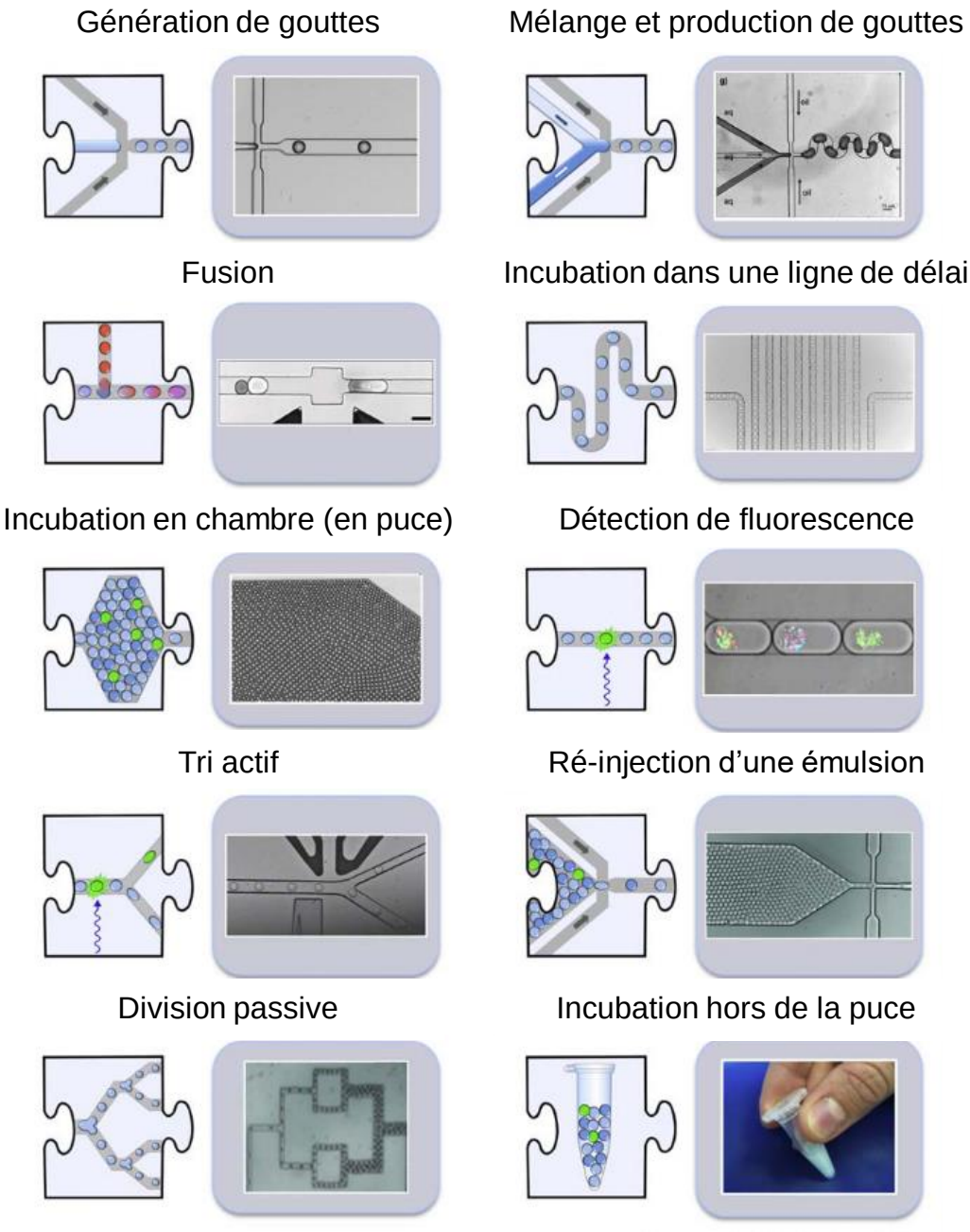


Figure 6 Manipulations de microgouttes, extrait et traduit à partir de (Kintses *et al.*, 2010)

Un exemple d'expérience à ultra haut-débit faisant appel à ces différentes manipulations des gouttes microfluidiques est une étude d'évolution dirigée, pour l'optimisation des capacités catalytiques d'une protéine d'intérêt, l'enzyme HRP (Horse Radish Peroxydase). Dans cette étude, la goutte constitue un compartiment à part pour chaque protéine à tester et créée un lien direct entre phénotype et génotype (Tawfik & Griffiths, 1998). Une librairie de 10^8 variants est ainsi analysée à très haut débit de sorte à identifier, après plusieurs cycles de sélection, des versions mutantes optimisées de l'enzyme de départ (Agresti *et al.*, 2010). Les gouttes sont produites et permettent l'encapsulation des levures transformées et d'un substrat non fluorescent dans des gouttes de 6 picolitres à une fréquence de 2kHz selon une loi de Poisson. Celles-ci sont ensuite incubées pendant 5 minutes dans la puce microfluidique dans une ligne de délai, et l'enzyme HRP va convertir le substrat en un produit d'oxydation coloré, avec une efficacité catalytique plus ou moins grande selon les variants. La fluorescence de chaque goutte est enfin mesurée et les gouttes dépassant un seuil fixé, correspondant aux enzymes les plus efficaces, sont activement triées par diélectrophorèse.

II. Le transcriptome à l'échelle de la cellule unique

A. Introduction à certains concepts de base

1. Puces à ADN et séquençage de l'ARN (RNAseq)

Le terme « transcriptome » apparaît à la fin des années 90 (Velculescu *et al.*, 1997) (R. Lowe, Shirley, Bleackley, Dolan, & Shafee, 2017). Il se définit par l'ensemble des transcrits et leur abondance, présents dans une cellule à un instant t, et représentant les gènes exprimés par cette cellule (Wang, Gerstein, & Snyder, 2009). Le génome constitue une représentation statique de la cellule. Le transcriptome, lui, évolue dans le temps, en fonction de signaux internes et externes, et est une représentation du phénotype de la cellule.

L'analyse du transcriptome dans différents tissus, à différents temps ou dans différentes conditions permet de comprendre la fonction des gènes, comment ils sont régulés et également l'impact de certains transcrits dans l'émergence de maladies.

Les technologies en transcriptomique sont en constante évolution depuis les années 90, permettant d'accéder à des informations de plus en plus riches et précises. Au début limitée à des études ciblées (puces à ADN), la technologie a évolué vers des études du transcriptome entier (RNAseq), puis, plus récemment, vers des études à l'échelle de la cellule unique.

Les deux technologies modernes majeures en transcriptomique sont la technologie des puces à ADN et la technologie RNAseq.

La technologie des puces à ADN (R. Lowe *et al.*, 2017) se base sur l'utilisation d'amorces d'intérêt fixées sur un support solide (tel que du verre) sur lesquelles vont se fixer par complémentarité les transcrits, auxquels a été ajouté préalablement un fluorophore. Elle permet de mesurer l'intensité de quelques milliers de transcrits en parallèle. L'intensité de fluorescence à un point d'ancrage, c'est-à-dire pour un transcrit d'intérêt, est fonction du nombre de transcrits s'étant fixé et donc du nombre de transcrits présents dans l'échantillon

de départ. La technologie a l'avantage de ne pas nécessiter d'étape de préparation complexe. Les avancées technologiques ont permis d'augmenter la résolution en ajoutant plus de points d'ancrage, soit plus de transcrits analysables en parallèle. Brièvement, les ARN messagers sont extraits et purifiés puis retro-transcrits en ADN complémentaires. Après fragmentation, ils sont étiquetés avec un fluorophore et déposés sur la puce à ADN recouverte d'amorces spécifiques. De nombreuses puces sont aujourd'hui commercialisées. On peut par exemple citer les puces GeneChip d'Affymetrix. Ces puces à très haute densité sont fabriquées en couplant la synthèse chimique d'oligonucléotides et des techniques de lithographies.

Cette technologie, très populaire au début des années 2000, présente quelques inconvénients cependant ; elle nécessite de grandes quantités de matériel de départ (de l'ordre du µg d'ARN) ainsi qu'une connaissance préalable du transcriptome à étudier pour choisir les amorces adaptées. L'efficacité de l'hybridation ne sera pas la même d'une amorce à l'autre, entraînant des biais dans l'analyse. De plus, du fait d'un fort bruit de fond dû à l'hybridation croisée entre différentes amorces, la gamme dynamique d'analyse n'est que de 3 ordres de grandeurs. Enfin, la technologie des puces à ADN permet d'étudier à nombre limité d'échantillons, à l'échelle de la population et non de la cellule unique.

Le RNAseq a été développé au début des années 2000 et à très vite connu un très fort engouement, lié notamment à l'émergence des technologies de séquençage de nouvelles générations (technologies NGS), qui délivrent une information à l'échelle de la molécule unique. Celles-ci ont connu une évolution très rapide, offrant des capacités de séquençage toujours plus grandes et des coûts de plus en plus bas. Wang et al parle d'un outil révolutionnaire pour les études transcriptomiques (Wang *et al.*, 2009), avec tout un panel d'applications. Il peut être utilisé pour quantifier des différences d'expression entre échantillons (Mortazavi, Williams, McCue, Schaeffer, & Wold, 2008), mais pour la découverte de nouveaux transcrits, variants d'épissage ou de variations génétiques, telles que de SNPs (Chen, Zhou, Wang, & Ning, 2016)(Y. Zhao *et al.*, 2019) (Polymorphisme d'un Seul Nucléotide) ou encore pour la construction de novo de transcriptome (VERA *et al.*, 2008), grâce à la capacité de pouvoir séquencer les transcrits sur toute leur longueur. Cet outil a permis de révéler toute la complexité des transcriptomes eucaryotes (Sultan *et al.*, 2008)(Nagalakshmi *et al.*, 2008).

La préparation d'une librairie de RNAseq consiste à extraire et purifier les ARN polyadénylés, puis à les transformer en courts fragments ADN via une étape de retro transcription puis de fragmentation. Des adaptateurs nécessaires au séquençage sont ajoutés aux extrémités des fragments puis les différentes molécules sont séquencées. Les séquences obtenues sont alors alignées contre un génome de référence (ou assemblés de novo) et les différents transcrits sont quantifiés, en utilisant généralement l'unité RPKM (pour *Read Per Kilobase of exon model per Million mapped reads*). Cette unité reflète la concentration d'un transcrit dans l'échantillon de départ en normalisant par rapport à la longueur de l'ARN et le nombre total de *reads* mesurés dans l'échantillon (Mortazavi *et al.*, 2008). La profondeur de séquençage, définie par le nombre de molécules qui peuvent être lues par le séquenceur, a un impact sur la détection des transcrits faiblement exprimés et donc sur la limite de détection. Dans des conditions optimales de séquençage (*i.e.* avec une profondeur suffisante), la technologie RNAseq a montré une large gamme dynamique de l'ordre de 5 ordres de grandeurs, en adéquation avec la gamme dynamique d'expression d'une cellule eucaryote typique. Une étude comparative d'expression différentielle chez des cellules T activées(S. Zhao, Fung-Leung, Bittner, Ngo, & Liu, 2014) a montré une grande reproductibilité entre les technologies RNAseq et des puces à ADN, avec un avantage de la technologie RNAseq quant à sa sensibilité, sa gamme dynamique plus large et sa capacité à détecter des isoformes d'épissage ou des variants génétiques.

2. Le séquençage de seconde génération : exemple de la technologie Illumina

Plusieurs technologies de séquençage existent et diffèrent selon la méthode de préparation des libraires et la stratégie de séquençage et d'imagerie utilisés (Metzker, 2010). On peut par exemple citer la technologie Pacific Bioscience, qui n'inclut pas d'étape d'amplification du matériel initial et est adaptée au séquençage de longues molécules. La technologie la plus utilisée aujourd'hui est la technologie Illumina, dont le fonctionnement est présenté dans la Figure 7.

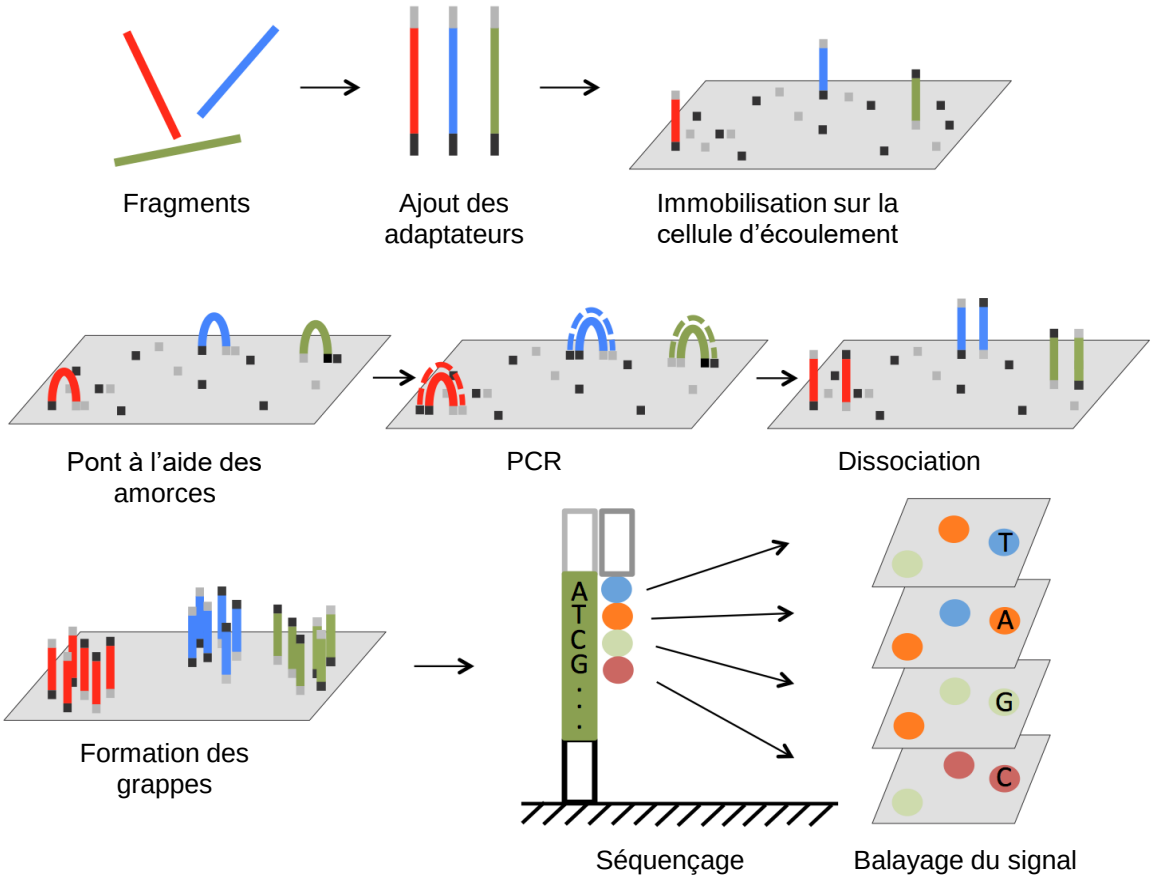


Figure 7 Principe de séquençage avec la technologie illumina, extrait et traduit de (Lu, Shen, Warren, & Walter, 2016)

Les molécules à séquencer sont préparées de sorte à ajouter à chacune d'elles des adaptateurs pour l'immobilisation sur un support solide par complémentarité (les amorces P5 et P7) ainsi que des séquences universelles, site d'hybridation des amorces Rd1 et Rd2 à partir desquelles seront lues les séquences. Les molécules sont ensuite déposées sur une cellule d'écoulement, et elles s'y immobilisent par capture des adaptateurs P5, sur des sites suffisamment espacés. Il est primordial de correctement quantifier l'échantillon de sorte à déposer un nombre adapté de molécules sur le support solide. Si trop de molécules s'attachent à la cellule, les signaux de chaque molécule risquent de se confondre, aboutissant à un séquençage de mauvaise qualité.

S'ensuit une amplification par PCR en pont, aboutissant à la génération de grappes, étant le fruit de l'amplification d'une seule molécule. Cette étape permet d'amplifier le signal généré par chaque molécule, de sorte à ce qu'il soit compatible avec le système de détection optique. La lecture de chaque grappe se fait ensuite selon une technique de séquençage par synthèse réversible à terminaison cyclique, faisant appel à une ADN polymérase et à des bases

fluorescentes (4 couleurs, une par base) modifiées à leur extrémité 3’ (ou bloqués) de sorte à bloquer l’addition d’une seconde base. Le déroulement est le suivant ; une base modifiée vient se fixer par complémentarité en face de la matrice, c’est-à-dire la molécule à séquencer, en 3’ de l’amorce Illumina. Les bases ne s’étant pas fixées sont lavées puis la cellule est imagée afin de voir quelle base a été attachée au niveau de chaque grappe. Le site bloqueur en 3’ est alors clivé et une nouvelle base peut venir d’incorporer à la suite.

Les molécules peuvent être ou non retournées, offrant la possibilité de lire les molécules d’intérêt à chaque extrémité ; on parle respectivement de séquençage Single-*Read* ou Paired-End.

Les données sont fournies sous forme d’un fichier texte de type fastq, avec autant de lignes qu’il y avait de molécules attachées à la cellule d’écoulement. A chaque molécule sont associées les informations suivantes ; un identifiant correspondant à la position de la molécule sur la cellule, la séquence lue, et la qualité de séquençage, reflet du degré de confiance quant à l’exactitude de la base lue à chaque position.

Ces séquences sont ensuite alignées à des génomes de référence, pour identification. Chaque occurrence est définie comme un *read* (une lecture). Si la séquence associée à un gène X a été retrouvée 100 fois dans le fichier, cela signifie que 100 molécules correspondant à ce gène étaient présentes dans l’échantillon séquencé.

B. L'histoire du séquençage de l'ARN à l'échelle de la cellule unique

Très rapidement après le développement du RNAseq, celui-ci a été adapté à l'étude sur des cellules uniques (Tang *et al.*, 2009). Le but initial était d'améliorer la sensibilité de la technologie RNAseq afin de rendre possible les études transcriptomiques sur un nombre restreint de cellules, dans le cadre d'études sur le développement embryonnaire précoce. De nombreuses améliorations techniques sont rapidement apparues, rendant possible des études à très haut débit sur des dizaines voire des centaines de milliers de cellules, avec une quantification plus précise et des méthodes d'analyse plus poussées (Kolodziejczyk *et al.*, 2015b).

Le RNAseq à l'échelle de la cellule unique, ou scRNAseq, est devenu un outil essentiel pour les biologistes, leur permettant d'identifier de nouveaux sous-groupes de cellules dans une population ainsi que des populations rares, de comprendre les transitions entre différents états cellulaires en découvrant des profils transcriptionnels intermédiaires, ou encore d'étudier les interactions entre gènes au sein des cellules, reflet de la complexité des réseaux de régulation.

Dans cette section, je vais présenter les premières grandes innovations mises au point et faisant du scRNAseq ce qu'il est aujourd'hui.

Le scRNAseq a été introduit par Tang *et al.* en 2009. S'inspirant de la technologie des puces à ADN (Kurimoto *et al.*, 2006), ils ont développé une méthode de RNAseq pour caractériser le transcriptome complet d'un blastomère unique en utilisant les nouveaux outils de séquençage. Contrairement aux techniques de RNAseq de l'époque, nécessitant une grande quantité de matériel de départ (quelques centaines de milliers de cellules mammaires), ils sont parvenus à obtenir le signal transcriptomique d'une seule cellule, avec une grande profondeur. Ils ont ainsi découvert de nouveaux transcrits, de nouvelles jonctions d'épissage et surtout ils ont pu quantifier l'abondance des différents transcrits, avec une large gamme dynamique de 5 ordres de grandeur, permettant la détection de gènes faiblement exprimés. Cependant, leur technologie, basée sur l'utilisation d'amorces polyT, bien qu'optimisée, ne leur permet pas d'obtenir le signal en 5' de longs transcrits de plus de 3kb (36% de perte

environ) ni d'ARN messagers dépourvus de queue polyA (tels que les ARN messagers d'histone). Leur travail est particulièrement intéressant pour la recherche sur le développement embryonnaire ou encore l'étude des cellules souches, où la quantité de matériel de départ est très faible. En revanche, ce travail à très bas débit (une seule cellule est étudiée) n'inclut pas encore la notion de code-barres moléculaires, permettant de multiplexer cette analyse à plusieurs cellules.

Cette notion clé apparait en 2011 (Islam *et al.*, 2011) avec l'introduction de la technologie STRT-seq (pour *Single-cell Tagged Reverse Transcription*). Pour la première fois, le transcriptome de 96 cellules uniques a été analysé. Les cellules sont préalablement isolées dans une plaque 96 puits contenant du tampon de lyse. Les ADN complémentaires (ADNc) sont synthétisés et ils incorporent un code-barres de puits (96 code-barres différents) par *template switching* à l'extrémité 3' des ADNc néosynthétisés. Le *template switching* est un mécanisme faisant appel à la capacité des retro-transcriptases de type Moloney murine leukemia virus (M-MLV) d'ajouter des paires de bases de type Cytosine en 3' de l'ADNc synthétisé, une fois arrivées à la fin de la matrice ARN (Zhu *et al.*, 2001). Un oligonucléotide contenant une extrémité 5' GGG et une séquence d'intérêt (ici, un code-barres de puits et une séquence universelle) s'accroche par complémentarité aux Cytosines et sert d'amorce pour générer le second brin complémentaire (Figure 9). Les molécules d'ADNc, une fois identifiées par leur code-barres, peuvent alors être mélangées, ce qui facilite grandement les étapes de préparation des librairies de séquençage et limite le nombre de cycles d'amplification, qui peut générer de nombreux biais et introduire des erreurs.

Cette technique d'amplification de l'ADNc par *template switching* est proposée dans une version améliorée avec les protocoles SmartSeq (Ramsköld *et al.*, 2012) puis SmartSeq2 (Picelli *et al.*, 2014) et va s'avérer être l'une des principales méthodes d'amplification de l'ADNc employée par les différents protocoles de scRNAseq.

La technologie CEL-seq (pour *Cell Expression by Linear amplification and Sequencing*) (Hashimshony, Wagner, Sher, & Yanai, 2012) voit le jour en 2012 et reprend la notion de code-barres mais diffère de la technologie STRT-seq au niveau de la méthode d'amplification

des ADNc, la transcription *in vitro (IVT)*. L'IVT est une technique d'amplification linéaire, qui n'introduit pas de biais, d'où son intérêt. En revanche, contrairement à la PCR, qui n'est pas limitée par la quantité de matériel de départ, cette technique nécessite un minimum de 400pg d'ARN total pour être efficace. Une cellule eucaryote typique est composée d'environ 10 à 30pg d'ARN total, dont 1 à 5% sont des ARN messagers. L'ARN provenant d'une seule cellule n'est donc pas suffisant. Eberwine *et al.* propose d'appliquer cette technique d'amplification à des cellules uniques mais le protocole en 3 étapes d'amplification est très long et laborieux (Eberwine *et al.*, 1992). Les auteurs reprennent ici la notion de code-barres évoquée précédemment pour permettre d'augmenter la quantité de matériel de départ en regroupant les ARN provenant de plusieurs cellules tout en gardant l'information cellule unique et pouvoir ainsi réaliser une amplification linéaire efficace. Le protocole CEL-seq est le suivant ; dans un premier temps les cellules sont isolées dans des microtubes et lysées, puis une réaction de RT est réalisée à partir d'une amorce contenant ; un promoteur T7 pour initier la transcription, une séquence universelle (un adaptateur illumina), un code-barres unique (un différent dans chaque tube) et enfin une queue polyT pour capturer les ARNm polyadénylés. Les ADNc néosynthétisés portent l'information cellule unique grâce au code-barres et sont regroupés dans un seul tube. Après synthèse du brin complémentaire des ADNc, la réaction d'IVT permet d'amplifier le matériel de départ. L'ARN amplifié est alors fragmenté et un adaptateur est ligué à l'extrémité 3' des molécules, permettant une étape finale d'amplification en 12 cycles de PCR pour sélectionner les fragments porteurs des 2 adaptateurs. Ils sont alors prêts pour être soumis au séquençage (Figure 9). Les auteurs ont comparé leur méthode à la technologie STRT-seq en utilisant les mêmes échantillons que ceux utilisés par Islam et son équipe (9 cellules souches embryonnaires de souris ainsi que 7 fibroblastes embryonnaires de souris) et des ARN synthétiques (ERCC Spike In) et la technologie CEL-seq a montré une meilleure reproductibilité entre les différents réplicas ainsi qu'une distinction plus efficace des 2 types cellulaires. Ils ont défini la limite de détection de leur technologie à 50% de chance d'identifier un transcrit possédant 5 copies ARN. Finalement, la technologie CEL-seq est une technologie qui offre plusieurs avantages ; elle permet l'analyse d'un échantillon large grâce à sa capacité de multiplexage et d'une amplification linaire permettant de réduire les biais techniques d'amplification entre différents transcrits et ainsi d'augmenter la reproductibilité. Cette technologie conserve également l'information de brin. On peut énoncer certaines limites également ; cette technologie souffre d'un fort biais en 3', n'offrant pas la possibilité d'analyser des variants d'épissage ni de séquencer les transcrits sur toute leur longueur. De plus, cette technologie est limitée à

l'analyse des ARN polyadénylés, avec un mode de capture par une amorce polyT. Enfin, sa limite de détection rend l'analyse de gènes faiblement exprimés, de l'ordre de 1 à 30 copies, compliquée. Ses 2 derniers inconvénients sont aussi valables pour la plupart des technologies RNAseq en cellule unique utilisant une amorce polyT.

Une autre notion majeure à introduire est l'identifiant moléculaire unique ou UMI.

Le scRNAseq est une technique précise et profonde qui permet d'avoir accès à de nouvelles informations sur les cellules. Elle est définie par une certaine limite de détection, dépendant de l'efficacité de la capture des ARN et de la réaction de RT, et par la précision de la quantification à l'issue du séquençage, dépendant des biais d'amplification du matériel de départ, à savoir les ADNc néosynthétisés (Islam *et al.*, 2014).

Les premières méthodes d'analyse de données RNAseq se basaient sur le comptage des *reads* à l'issue du séquençage pour déterminer l'abondance de différentes espèces/ARN. Ainsi en 2009, Tang *et al.* quantifient les différents transcrits présents dans leur blastomère en utilisant la méthode de comptage RPKM (*reads per kilobase of exon model per million mapped reads*). Cette méthode de calcul ne corrige pas les éventuels biais pouvant survenir lors des étapes d'amplification du matériel de départ, (en particulier si l'amplification se fait par PCR) ; ainsi, une séquence qui est amplifiée plus efficacement sera surreprésentée au séquençage quand bien même elle était présente en même quantité initiale qu'une autre séquence avec une faible efficacité d'amplification. De même, cette méthode de quantification est relative, elle dépend de l'échantillon total. Kivioja *et al.* proposent une solution (Kivioja *et al.*, 2012) pour améliorer la précision de quantification et s'affranchir des biais d'amplification en introduisant un identifiant moléculaire unique qui sera fixé à chaque molécule présente dans l'échantillon avant amplification. L'UMI est une séquence de quelques paires de bases aléatoires, tarditionnellement 5bp, soit 1024 (4^5) séquences possibles, ce qui permet de couvrir toutes les molécules uniques issues d'une cellule (Islam *et al.*, 2014). Une cellule contient environ 10^5 à 10^6 molécules d'ARN, correspondant à l'expression d'environ 10 000 gènes différents soit environ 100 molécules d'ARNm identiques en moyenne (bien moins que les 1024 possibilités d'UMI). Islam et son équipe ont d'ailleurs vérifié cette théorie en analysant le nombre d'UMI comptés pour chaque gène d'une cellule, ce nombre était bas dans la plupart des cas et n'a jamais excédé 1000 UMI.

Si l'on reprend l'exemple des 2 molécules présentes en même quantité initiale mais avec des efficacités d'amplification différentes, et que l'on marque chaque molécule avec un UMI, celles-ci auront un nombre de *reads* différents en fin de séquençage mais le même nombre d'UMI. Un échantillon de RNAseq ayant été amplifié 15 fois ou 25 fois a été comparé avec les 2 méthodes de quantification (RPKM et UMI). Quand la première méthode montre des différences de quantification entre les 2 amplifications, la méthode UMI permet de corriger les biais et donne les mêmes résultats dans les 2 cas.

Cette méthode n'a de sens que si l'échantillon est séquencé avec suffisamment de profondeur ; chaque UMI doit être observé plusieurs fois (plusieurs *reads*), garantissant qu'aucun UMI n'a disparu lors de l'échantillonnage.

L'UMI est donc un outil essentiel pour déterminer le nombre de transcrits converti en ADNc et associés à un gène et donc son niveau d'expression, en s'affranchissant des biais d'amplification survenant au cours des étapes de préparation des librairies de séquençage, source importante de variabilité technique.

Il est cependant important de noter que cet outil est classiquement adapté à des protocoles où seulement une extrémité du transcrit est séquencée et où la normalisation en fonction de la taille du transcrit n'est pas applicable. Il ne peut donc pas apporter d'information quantitative pour étudier la proportion de différents isoformes ou l'expression d'allèle spécifique, qui sont possibles uniquement avec des protocoles séquençant le transcrit sur toute sa longueur est nécessitant une étape de fragmentation.

Enfin, l'UMI seul ne permet pas de corriger les biais techniques liés à des différences d'efficacité de RT d'une cellule à l'autre. Pour cela, d'autres outils tels que des ARN synthétiques externes peuvent être utilisés (Stegle, Teichmann, & Marioni, 2015).

C. Le séquençage de l'ARN à l'échelle de la cellule unique en 2019 : Etat de l'art

De nombreuses méthodes de scRNAseq existent aujourd'hui. On ne peut pas dire qu'une méthode est meilleure qu'une autre, mais plutôt qu'une méthode est plus adaptée à une application.

Grâce aux avancées techniques présentées précédemment, de plus en plus de cellules peuvent être analysées en parallèle. Alors que seulement quelques dizaines de cellules étaient étudiées en 2009, à peine 10 ans plus tard, ce sont de dizaines de milliers de cellules qui sont séquencées en parallèle (Svensson, Vento-Tormo, & Teichmann, 2018).

Cette augmentation de débit phénoménale est liée à l'introduction du code-barres cellulaire et aux grandes avancées dans les technologies de capture et d'isolement des cellules uniques, et notamment à l'avènement de la microfluidique.

Plus récemment, une nouvelle méthode d'ajout de code-barres en plusieurs étapes à l'intérieur même de cellules fixées ou de noyaux a vu le jour. Les technologies sci-RNA-seq (Cao *et al.*, 2017) et dernièrement, SPLIT-seq (Rosenberg *et al.*, 2018), emploient cette approche. La technologie SPLIT-seq, par exemple, permet de multiplexer jusqu'à 100 000 cellules. Celles-ci sont fixées et perméabilisées puis réparties dans une plaque 96 puits, où va s'opérer une première étape de RT *in situ* avec une amorce polyT contenant un index de puits. La réaction se fait donc à l'intérieur de la cellule, qui sert de compartiment aux ADNc synthétisés. Les cellules sont regroupées puis redistribuées dans une nouvelle plaque 96 puits, où va se faire une réaction de ligation pour ajouter un second index de puits sur les ADNc. Cette opération est répétée une troisième fois, avec un nouvel index de puits incrémenté d'un UMI. Finalement, les cellules, contenant leurs ADNc maintenant étiquetés avec code-barres et UMI, sont regroupées puis réparties de nouveau dans 24 puits où elles sont lysées avant amplification par PCR des ADNc à l'aide d'amorces indexées. En combinant 4 étapes d'indexation, jusqu'à 20 millions de code-barres peuvent être générés, permettant de multiplexer en théorie jusqu'à 1 million de cellules. Cette méthode offre l'avantage d'un débit encore jamais atteint, et surtout ne nécessite pas d'équipements de pointes telles qu'un FACS ou une station microfluidique.

La Figure 8 présente l'évolution des technologies scRNAseq en termes de nombre de cellules analysables. Ce gain de débit sans précédent permet aujourd'hui d'étudier la complexité de tout un organe voire d'un organisme entier.

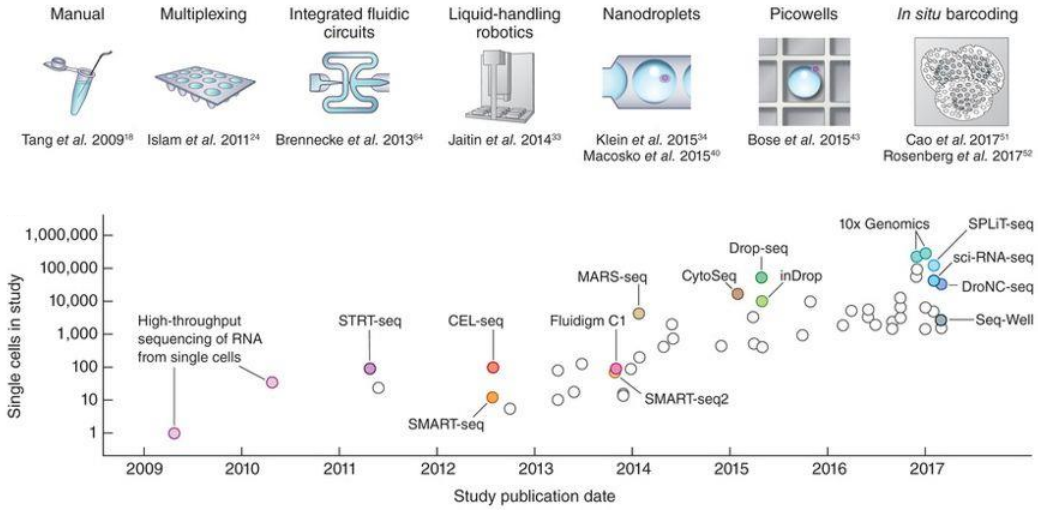


Figure 8 Evolution des technologies scRNAseq au cours de ces 10 dernières années. Image extraite de (Svensson *et al.*, 2018)

3. Classification des différentes technologies scRNAseq

Les différentes technologies scRNAseq suivent pour la plupart le même schéma en 4 étapes clés à savoir :

- la capture des cellules uniques ;
- la lyse des cellules et la rétrotranscription des ARN messagers à partir d'amorces polyT ;
- l'amplification des ADNc néosynthétisés ;
- la préparation des librairies de séquençage ;

Chacune de ces étapes va varier d'une méthode à l'autre, en voici les différentes possibilités.

a) La capture des cellules uniques

La capture des cellules uniques peut se faire manuellement, à l'aide d'une micropipette, méthode employée par Tang *et al.* en 2009 pour isoler le blastomère. Cette technique précise permet de s'affranchir des doublets mais présente un très bas débit ; elle est adaptée aux études sur des échantillons de quelques dizaines de cellules.

Afin d'augmenter le débit de capture, les cellules d'un tissu peuvent être dissociées et mises en suspension avant d'être capturées. Cette dissociation préalable est employée par les différentes techniques exposées ci-après. Cette étape fait appel à des enzymes telles que la trypsine, la papaïne ou encore des collagénases et à des traitements mécaniques, pouvant provoquer une augmentation de la mort cellulaire ainsi que des changements transcriptionnels (Olsen & Baryawno, 2018)(Kolodziejczyk *et al.*, 2015b).

Le tri cellulaire actif par fluorescence faisant appel à la cytométrie en flux (FACS) pour *Fluorescence Activated Cell Sorting*) est une technique de choix pour sélectionner et trier des cellules uniques. L'échantillon peut être préalablement marqué avec des anticorps fluorescents spécifiques de marqueurs de surface d'intérêt, permettant d'ajouter une dimension protéomique à l'étude. Il est ensuite soumis au cytomètre qui distingue chaque évènement unique et va déposer chaque cellule d'intérêt dans le puits d'une plaque 96 puits ou 384 puits. Le calibrage de la machine doit se faire de manière très précise pour assurer qu'une cellule unique soit déposée bien au fond du puits, contenant le tampon de lyse. Cette technique de tri est celle employée par la technologie MARS-seq (Jaitin & Kenigsberg & Keren-Shaul *et al.*, 2014).

Ces 2 techniques offrent l'avantage de pouvoir précisément sélectionner les cellules, en fonction de leur morphologie ou de leurs marqueurs de surface en cas d'utilisation d'anticorps fluorescents. Elles permettent d'avoir un grand degré de confiance quant au fait qu'une seule cellule est bien présente dans chaque compartiment. En revanche, le volume réactionnel est grand, de l'ordre du microlitre pour chaque compartiment, rendant les coûts des réactifs importants. Le débit est très limité dans le cas de la micromanipulation, et l'est dans une moindre mesure pour un tri par FACS (quelques milliers de cellules en multiplexant les plaques). Enfin, ces 2 méthodes sont assez lentes (en particulier la micromanipulation), ce qui peut accroître le risque de mort cellulaire ou de dégradation de l'ARN dans les cellules. La cytométrie est une technique à ultra haut débit, mais les paramètres de tri, très stricts, et

les étapes de mises au point, de même que le nécessité de multiplexer le tri en plusieurs plaques, peuvent nécessiter plusieurs heures de travail et ralentissent le débit.

Des outils microfluidiques ont été mis en place pour permettre d'isoler des cellules dans des compartiments au volume réactionnel très faible et à très haut débit. Elles sont présentées plus en détail dans la section I.B.

Ainsi, les technologies inDrop (Klein *et al.*, 2015)(Zilionis *et al.*, 2017), Drop-seq (Macosko *et al.*, 2015) et le système commercial 10x Chromium, employant la technologie GemCode (Zheng *et al.*, 2017), font tous 3 appels à la microfluidique en gouttes pour compartimenter des cellules uniques avec des billes porteuses d'amorces à code-barres cellulaires, dans des gouttes d'un volume de l'ordre du nanolitre. Il s'agit de méthodes à ultra-haut débit, permettant de séquencer le transcriptome de dizaines à centaines de milliers de cellules en quelques dizaines de minutes.

La technologie des circuits fluidiques intégrés à valves proposée par Fluidigm permet d'isoler une centaine de cellules dans une puce, visualisable au microscope, avec certaines contraintes quant à la taille et la forme des cellules, pour assurer une capture efficace dans les différents compartiments, d'un volume de l'ordre du nanolitre. Cette méthode a donc un débit modéré mais un très grand contrôle. Le volume réactionnel est faible mais le prix du dispositif n'est pas négligeable.

Enfin, la technologie des micro-puits, telles que la technologie CytoSeq (H. C. Fan, Fu, & Fodor, 2015) permet de répartir dans les différents compartiments aux dimensions microscopiques, des cellules uniques et des billes magnétiques porteuses d'une amorce polyT à code-barres et possédant un UMI. Comme décrit plus tôt dans cette introduction, la distribution des cellules dans les puits suit une loi de Poisson tandis que les dimensions des puits et des billes permettent de charger quasiment tous les puits avec une seule bille magnétique. Il y a ainsi une bille dans chaque puits. Les puits offrent un débit de l'ordre de la dizaine de milliers de cellules et la possibilité d'observer chaque compartiment au microscope pour repérer ceux contenant des doublets de cellules.

Bien que permettant une étude à beaucoup plus bas débit, les techniques de type FACS ou Fluidigm, ne subissant pas de loi de Poisson pour le chargement des cellules dans les différents compartiments, peuvent bénéficier d'un contrôle externe de l'efficacité de capture des ARNm et de leur rétrotranscription dans chaque compartiment. Pour cela, des ARN synthétiques sont chargés dans chaque compartiment ; il s'agit d'un mélange de 92 ARN polyadénylés de 500 à 2500 nucléotides, développés par un consortium, et baptisé ERCC Spike-In mix (Baker *et al.*, 2005). De concentrations connues et réparties sur une gamme dynamique de 6 ordres de grandeur, leurs séquences ont été blastées contre le génome de référence de la souris ou encore de l'humain. Ils servent ainsi à normaliser les données provenant des différents compartiments et réduire l'impact du bruit technique, particulièrement important lorsque la quantité de matériel de départ est faible, comme c'est le cas pour une cellule unique (Brennecke *et al.*, 2013)(Grün, Kester, & Van Oudenaarden, 2014). Enfin, ils servent d'outil de normalisation permettant de réduire l'effet « batch », à savoir les différences entre échantillons opérés séparément et dû à des biais techniques. D'autres spikes ont également été développés, en suivant le même principe, par la suite. On peut citer les SIRVs (Spike-In RNA Variants)(Paul et al., 2016), commercialisés par Lexogen, qui permettent l'ajout d'une nouvelle dimension d'analyse avec l'introduction de variants d'épissage.

Cet outil a également été utilisé pour caractériser les performances des différentes technologies scRNAseq et notamment les technologies en gouttes, et les comparer entre elles (Wu *et al.*, 2014)(Svensson *et al.*, 2017)(Ziegenhain *et al.*, 2017). En revanche, ces ARN synthétiques ne peuvent pas être employés en tant que contrôles externes en routine pour les technologies en gouttes (Hwang, Lee, & Bang, 2018). Du fait de la loi de Poisson suivant laquelle les cellules sont chargées, jusqu'à 90% des compartiments sont dénués de cellules mais possèdent une bille à amorces polyT. Les ARN synthétiques sont, quant à eux, présents dans toutes les gouttes et ils occuperaient donc 90% du signal au séquençage. Il est à noter que ces ARN synthétiques présentent des différences non négligeables avec des ARN messagers provenant d'une cellule eucaryote tels que l'absence de coiffe en 5', pouvant affecter les performances de la réaction de *template switching*, tout comme une queue 3' polyadénylés plus courte (20 nucléotides), qui peut entraîner une dégradation plus forte de ces ARN (Kolodziejczyk & Lönnberg, 2017)(Svensson *et al.*, 2017).

b) La capture des ARNs et la synthèse en ADNc

Différents protocoles de capture des ARNm, puis de leur synthèse en ADNc, existent. Les ARNm polyadénylés sont capturés au niveau de leur queue polyA par une amorce polyT, sélectionnant ainsi préférentiellement les ARN messagers et non les ARN de transfert et ribosomaux, non informatifs quant au profil d'expression de la cellule mais bien plus abondants. Cette capture peut se faire sur des amorces de RT en solution dans le compartiment ou sur un support solide, tel qu'une bille d'hydrogel ; dans ce cas, le nombre total d'amorces délivrées est dépendant de la capacité fonctionnelle du support solide. Ce type d'approche est utilisé par les technologies en gouttes, où la bille permet d'amener dans chaque goutte un code-barres unique, précédent la séquence polyT. Les amorces sont en général libérées une fois dans le compartiment, par photo clivage dans le cas du protocole inDrop, ou par dissolution de la bille d'hydrogel dans le cas du protocole 10x Chromium, ce qui permet d'augmenter l'efficacité de capture.

Dans le cas de la technologie Drop-seq ou CytoSeq, en revanche, les ARNm sont capturés sur les différentes billes à code-barres uniques et y restent accrochés, permettant de réaliser l'étape de RT de toutes les cellules dans un seul tube, et facilitant ainsi le workflow général.

c) L'amplification des ADNc

Le nombre d'ARNm est de quelques centaines de milliers par cellule (Marinov *et al.*, 2014), soit une échelle de l'ordre du picogramme. Il a donc fallu mettre au point des méthodes pour amplifier de très faibles quantités de matériel.

Tout d'abord, les différents ADNc synthétisées à partir des différentes cellules uniques sont préalablement étiquetés avec un code-barres moléculaire pour identifier leur appartenance à une cellule précise. Cela permet de regrouper tous les ADNc et ainsi augmenter la quantité de matériel initial tout en diminuant les risques de pertes ainsi que les manipulations en parallèle. Ce code-barres peut être ajouté à différents niveaux ; dans le cas des technologies inDrop, Drop-seq, 10x chromium ou encore CytoSeq, il est délivré par une bille dans le compartiment où a été capturée la cellule. Chaque bille est porteuse de millions d'oligonucléotides dont la séquence contient une séquence universelle, un code-barres

moléculaire unique pour chaque bille, mais différent d'une bille à l'autre, un identifiant moléculaire unique (UMI) et une queue polyT pour la capture des ARNm polyadénylés. Comme expliqué précédemment, l'UMI est une courte séquence aléatoire de quelques bases qui permet d'améliorer la quantification des transcrits. Une bille ne possède qu'un code-barres, en revanche, elle possède une multitude d'UMI. Dans le cas des technologies STRT-seq, Quartz-seq, ou encore SMART-seq, le code-barres est ajouté en 3' de l'ADNc *par template switching*. Cela permet de lire toute la séquence du transcrit ou en tout cas, son extrémité 5'.

Une fois regroupées après ajout du code-barres, les ADNc sont amplifiés selon 2 stratégies principales ; la PCR ou la transcription *in vitro* (Figure 9).

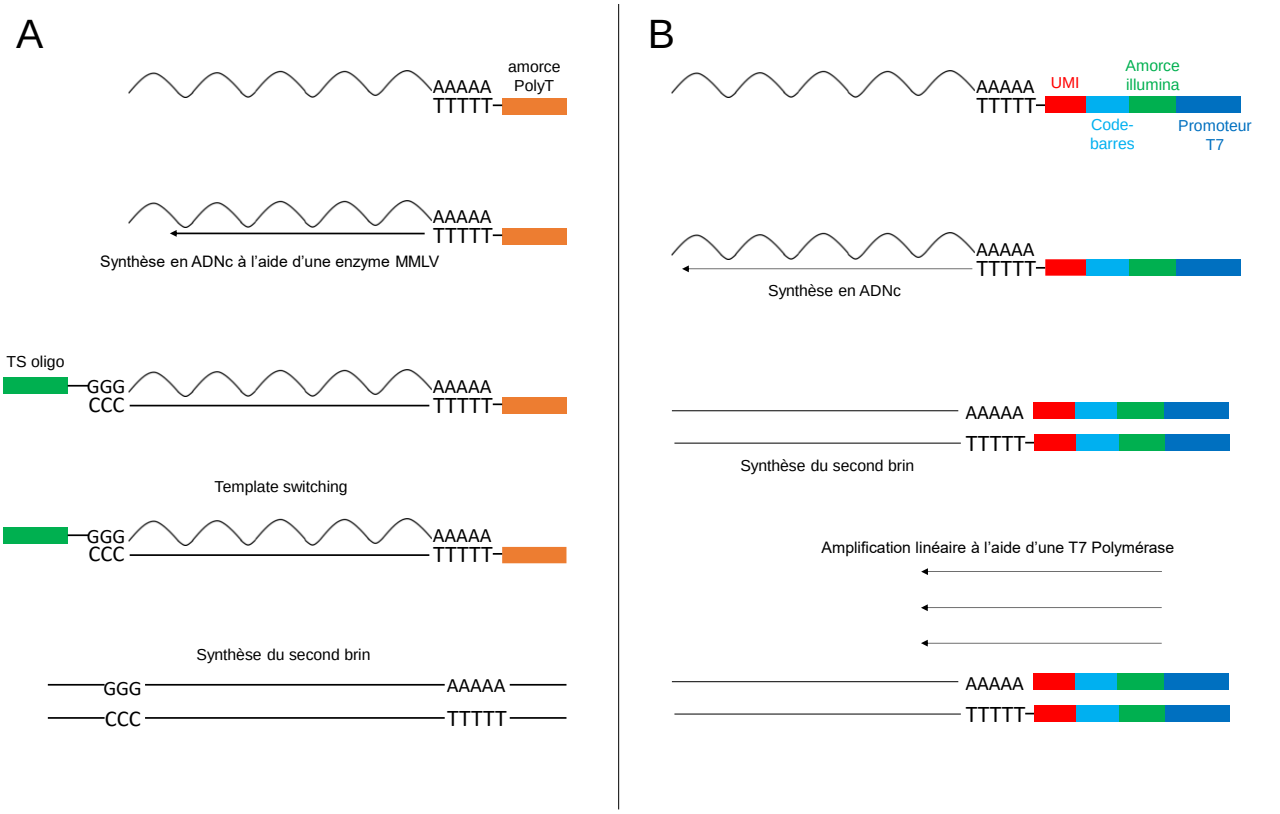


Figure 9 Amplification de l'ADNc selon une stratégie de A= Template switching et PCR ou B= Synthèse du second brin puis IVT (extrait et traduit à partir de (Olsen & Baryawno, 2018))

La première stratégie nécessite au préalable l'ajout d'une séquence universelle aux deux extrémités de l'ADNc pour permettre la fixation des 2 amorces de PCR. Tang *et al.*, en 2009,

ont employé une stratégie d'ajout d'une seconde queue polyA en 3' de l'ADNc néosynthétisé à l'aide de l'enzyme Terminal desoxynucléotidyl Transférase. Cette nouvelle extension sert d'ancre pour l'hybridation d'une amorce sens polyT, possédant une séquence universelle en 5' et à partir de laquelle peut être synthétisé le brin complémentaire, rendant l'ADNc double brin. Celui-ci est alors amplifié par PCR, avec deux amorces reconnaissant les 2 séquences universelles situées de part et d'autre de l'ADNc.

L'autre possibilité est d'utiliser la méthode du *template switching*, faisant appel à la capacité des enzymes RT de type M-MLV, d'ajouter 3 à 6 cytosines en fin de synthèse de l'ADNc (Figure 9). Une séquence d'intérêt, flanquée de 3 riboguanosines en 3', peut alors s'hybrider et servir de point de départ pour la synthèse du brin complémentaire. Picelli *et al.* proposent une version optimisée, baptisée SmartSeq2, de cette réaction. Les 3 riboguanosines de la version SmartSeq (Ramsköld *et al.*, 2012) sont remplacées par 2 riboguanosines classiques et une riboguanosine modifiée de type LNA (Locked Nucleic Acid) permettant d'augmenter la force d'hybridation entre les cytosines de l'ADNc et les 3 riboguanosines complémentaires (dont celle modifiée augmente la stabilité thermale du complexe) de l'oligonucléotide initiateur de la synthèse du second brin. Une autre amélioration est le changement de tampon avec l'ajout de bétaïne, donneur de groupements méthyles, permettant de réduire les structures secondaires des ARN et d'augmenter la stabilité thermale des protéines, accompagné d'une augmentation de la teneur en Mg2+, le tout entraînant une augmentation du rendement de la synthèse d'ADNc. Enfin, en éliminant une étape de purification par rapport au protocole initial, ils diminuent les pertes et assurent un meilleur rendement réactionnel et facilitent le workflow général (Picelli *et al.*, 2013)(Picelli *et al.*, 2014). L'ADNc double brin est flanqué de deux séquences universelles (l'une apportée par l'amorce polyT de RT et l'autre par l'oligonucléotide de *template swiching*) est prêt à être amplifié par PCR, en prenant garde à augmenter le temps d'extension lors de chaque cycle, considérant la longueur des ADNc générés (quelques kilopaires de bases).

Cette méthode de *template switching* a été adoptée par de nombreuses technologies de scRNAseq telles que STRT-seq, la version SmartSeq2 de C1 ou encore 10x chromium, car elle permet de couvrir toute la longueur des ARN et d'ainsi réduire les biais 3' tout en offrant des informations supplémentaires.

Cependant, l'amplification par PCR est non linéaire et apporte donc des biais, liés à une efficacité variable en fonction de la teneur en G :C et de la longueur de la séquence amplifiée, d'autant plus que le nombre de cycles est important (Wu *et al.*, 2017).

La seconde stratégie fait appel à un promoteur T7 pour initier une réaction de transcription, nécessitant une étape préalable de synthèse du second brin des ADNc néosynthétisés. Les enzymes utilisées pour la réaction de rétrotranscription sont dépourvues de leur activité RNAse H, critère nécessaire pour assurer une réaction efficace malgré les très faibles quantités de matériel apportées par une cellule unique. De ce fait, à l'issue de la synthèse des ADNc, le brin d'ARN peut être dégradé par une RNAse H, et les fragments générés peuvent servir d'amorces à une polymérase de type DNA Polymérase I pour synthétiser le second brin. Une ligase permet de fermer les espaces entre deux sites de polymérisation (Gubler, 1987)(Wu *et al.*, 2017). L'ADNc, maintenant double-brin, et possédant code-barres cellulaire et UMI, est amplifié selon une réaction d'IVT. Cette méthode d'amplification est linaire et permet de s'affranchir des biais générés lors d'une amplification exponentielle (Hashimshony *et al.*, 2012). En revanche, il est nécessaire de passer par une nouvelle étape de rétrotranscription avec des amorces aléatoires contenant un adaptateur de séquençage, générant un biais en 3'. Cette stratégie est celle employée par les technologies CEL-seq, MARS-seq ou encore inDrop.

d) Séquençage des librairies

Une question importante à se poser lors du séquençage d'un échantillon est la profondeur à laquelle celui-ci doit être séquencé. Cela va dépendre du nombre de cellules collectées, de la couverture transcriptomique associée à la technologie utilisée et enfin du degré de précision souhaitée.

Le premier point est lié à la complexité de l'échantillon analysé ; plus celui-ci est complexe, plus le nombre de cellules à séquencer devra être grand. De même, l'identification d'une population rare (par exemple à moins de 1%) nécessite un débit plus important. Le degré de complexité d'un échantillon peut être prédit en tenant compte d'autres données cellules

uniques déjà publiées telles que des données d'analyses protéomiques en cytométrie en flux ou en cytométrie de masse. Celui-ci peut également être réduit en présélectionnant une population dans l'échantillon de départ si cela est possible, par FACS ou encore par MACS (*Magnetic-Activated Cell Sorting*), mais cela nécessite une connaissance préalable de l'échantillon de départ et l'existence de marqueurs spécifiques permettant la sélection.

Le second point, à savoir la couverture transcriptomique, est inhérent à la technologie scRNAseq choisie. Tandis que certaines technologies permettent de couvrir les transcrits sur toute leur longueur, telle que la technologie SmartSeq, la plupart des protocoles ne couvrent qu'une extrémité des transcrits (l'extrémité 3' pour inDrop, ou encore CEL-seq, l'extrémité 5' pour STRT-seq). Alors que le séquençage d'une seule extrémité suffit pour définir les différentes sous-populations d'un échantillon (Macosko *et al.*, 2015), séquencer les transcrits sur toute leur longueur apporte des informations supplémentaires telles que la présence d'isoformes d'épissage voire la découverte de nouveaux variants. Cela implique bien évidemment une profondeur de séquençage plus grande.

Finalement, un point essentiel est la profondeur de séquençage nécessaire par cellule. Ce critère va fortement influencer la sensibilité de la mesure effectuée, affectant la limite de détection (Wu *et al.*, 2014). Des études ont montré qu'une profondeur de 1 à 2 millions de *reads* par cellule est nécessaire pour ne pas perdre d'information, soit de quoi couvrir 10 fois le nombre total de transcrits polyadénylés d'une cellule eucaryote typique. La sensibilité est alors dépendante de la technologie elle-même, de sa limite de détection (Grün *et al.*, 2014)(Svensson *et al.*, 2017). *A contrario,* Pollen *et al.* ont mené une étude systématique en utilisant le système SmartSeq C1 pour définir un seuil de profondeur limite dans l'analyse de différents échantillons. Ils ont ainsi séquencé à haute ou faible profondeur (1) un échantillon comprenant des cellules de différentes natures ou (2) un échantillon avec des cellules de même type en phase de différenciation. Ils ont montré que respectivement 10 000 *reads* et 50 000 *reads* par cellule étaient suffisants pour retrouver les différentes sous populations dans les 2 types d'échantillons. Bien que les transcrits faiblement exprimés ne soient pas détectés, la couverture d'un grand nombre de gènes, dont l'expression est modérée à forte, permet de faire des regroupements cellulaires et de compenser cette perte d'information (Pollen *et al.*, 2014).

Il est à noter que cette étude a été menée en utilisant un seul système de scRNAseq, or, la profondeur minimum nécessaire est dépendante de l'efficacité de capture des ARNm. Elle va donc varier d'une technologie à l'autre. De même, cette étude s'intéresse à la proportion des différents transcrits pour définir des sous-populations et non à la recherche de nouveaux variants d'épissage ou encore à la co-expression de certains gènes, ce qui nécessiterait alors une plus grande profondeur de séquençage (Haque, Engel, Teichmann, & Lönnberg, 2017). Dans ce cas, une profondeur de 1 million de *reads* par cellule est recommandée (Olsen & Baryawno, 2018).

e) Détermination de la sensibilité d'une technologie

La sensibilité d'une méthode se définit par sa limite de détection. Celle-ci va dépendre de l'efficacité de capture des ARNm et de leur RT en ADNc. Elle est également dépendante de la profondeur de séquençage, à savoir le nombre de reads par cellule disponible au séquençage, comme expliqué précédemment (section I.C.d)

L'efficacité des méthodes scRNAseq est calculée sur la base du nombre de molécules capturées par rapport au nombre de molécules d'ARN initialement présentes. La limite de détection est le seuil en dessous duquel un transcrit ne sera pas détecté, car présent en trop petit nombre de copies. Etant donné que le nombre de chaque transcrit dans une cellule n'est pas connu, ni fixe, les chercheurs ont notemment utilisé des molécules synthétiques de concentration connue pour déterminer l'efficacité de leur technologie et ainsi leur limite de détection. Le mix commercial ERCC Spike-in est préférentiellement utilisé dans ce but.

Ainsi, Macosko *et al.* ont évalué l'efficacité de leur technologie Drop-seq à 11% (Macosko *et al.*, 2015) tandis que Klein *et al.* ont estimé celle de la technologie inDrop à 7% (Klein *et al.*, 2015). Picelli *et al.*, quant à eux, ont évalué l'efficacité de leur technologie SmartSeq2 à 20% (Picelli *et al.*, 2014).

Ces efficacités sont certainement sous-évaluées car il a été démontré que la capture des ARN synthétiques du mix ERCC était moins efficace que celle des ARNm endogènes, du fait d'une queue polyA plus courte notamment. Ils constituent néanmoins un outil de comparaison très intéressant (Svensson *et al.*, 2017).

Il est également possible d'utiliser une lignée cellulaire homogène pour comparer l'efficacité de différentes méthodes. Bien qu'homogène, le bruit biologoque est toujours présent mais

cette stratégie permet de comparer l'efficacité générale de plusieurs méthode dans un contexte similaire aux futures études biologique (en terme d'échantillon biologique utilisé). L'efficacité est alors défini par en le nombre de gènes identifiés ainsi que le nombre de copie d'ADNc générés (nombre d'UMI). Cette stratégie est par exemple employée par Zhang *et al.* dans leur étude comparative des technologies de scRNAseq à très haut débit en gouttes sur une lignée lymphoblasoïde humaine (GM12891) (X. Zhang et al., 2019).

L'efficacité varie donc d'une méthode à l'autre mais également d'une cellule à l'autre, créant des biais techniques, qui peuvent être corrigés en utilisant des ARN synthétiques externes, présents en même quantité dans chaque compartiment ou en normalisant le nombre total d'UMI compté dans chaque cellule par le nombre total d'UMI moyen.

4. Comparaison de différentes technologies scRNAseq

De très nombreux protocoles scRNAseq existent et il convient de convenablement choisir celui étant le plus adapté à la question biologique posée.

La recherche de sous-populations rares nécessite des méthodes à très haut débit permettant de multiplexer de nombreuses cellules, et les technologies scRNAseq en gouttes sont alors particulièrement adaptées (Ziegenhain *et al.*, 2017). Alternativement, de nouvelles technologies, telles que la technologie SPLIT-seq, peuvent être employées. Elles offrent également un très haut débit, et n'utilisent pas d'équipements lourds. Ces technologies sont cependant très récentes, non commerciales et n'ont pas encore été adoptées par la communauté.

Au contraire, des études à plus bas débit mais pour lesquelles une détection plus fine est requise se tourneront vers des technologies plus sensibles et offrant la possibilité de séquencer les transcrits sur toute leur longueur.

Le Tableau 1 reprend les caractéristiques et performances des technologies scRNAseq appréciés des laboratoires de recherche.

Les 3 technologies en gouttes à ultra haut débit que sont inDrop (commercialisé par 1CellBio), Drop-seq et 10x Chromium ont récemment été comparées (X. Zhang *et al.*, 2019). Bien que similaires, ces 3 technologies présentent des différences expérimentales à chacun des niveaux décrits dans la section précédente I.A.3 (Figure 10). Cette étude a été menée pour les 3 technologies sur une même lignée cellulaire en plusieurs réplicas, et montre des résultats cohérents d'un réplica à l'autre pour une même technologie mais diffèrent d'une technologie à l'autre. Il y a donc une influence de la technologie sur les résultats obtenus.

La technologie 10x Chromium montre globalement de meilleures performances, avec des billes à la diversité effective plus proche de celle annoncée par rapport aux billes des autres concurrents, une plus grande sensibilité liée à une capture plus efficace (plus de transcrits et plus de gènes sont mesurés par cellule), une amplification moins biaisée car limitée mais c'est aussi la technologie la plus coûteuse. La machine 10x Chromium coûte environ $50 000, et le coût de fonctionnement par cellule, sans compter le séquençage, ni les frais de maintenance, est de $0.50. En comparaison, le coût par cellule est estimé à $0.25 avec la

technologie inDrop et à $0.10 avec la technologie Drop-seq, tandis que leurs instruments sont respectivement d'un coût de $50 000 et $30 000.

10x Chromium est la technologie commerciale la plus aboutie et optimisée, les résultats obtenus font donc sens. La technologie Drop-seq offre une performance acceptable, pour un coût moindre, et permet l'étude d'un grand nombre de cellules avec la plus grande combinaison de code-barres. En revanche, cette technologie souffre d'une double distribution de Poisson et ne sera pas à privilégier si l'échantillon à séquencer est rare et précieux. Il s'agira de la méthode de choix pour séquencer un grand échantillon à coût modéré. Finalement, la technologie inDrop est celle qui montre les moins bonnes performances, probablement lié à une suramplification lors de l'étape d'IVT et à l'utilisation de séquences aléatoires pour retro-transcrire les ARN lors de l'étape post IVT. Un meilleur contrôle qualité des codes-barres permettrait également de garder beaucoup de *reads* actuellement éliminés lors des étapes de filtration des données. En revanche, ce protocole est peu coûteux, ne subit pas une double distribution de Poisson des billes et cellules et est facilement adaptable. Cette technologie est à privilégier pour l'étude d'un échantillon rare ou si l'utilisateur désire faire du développement technologique.

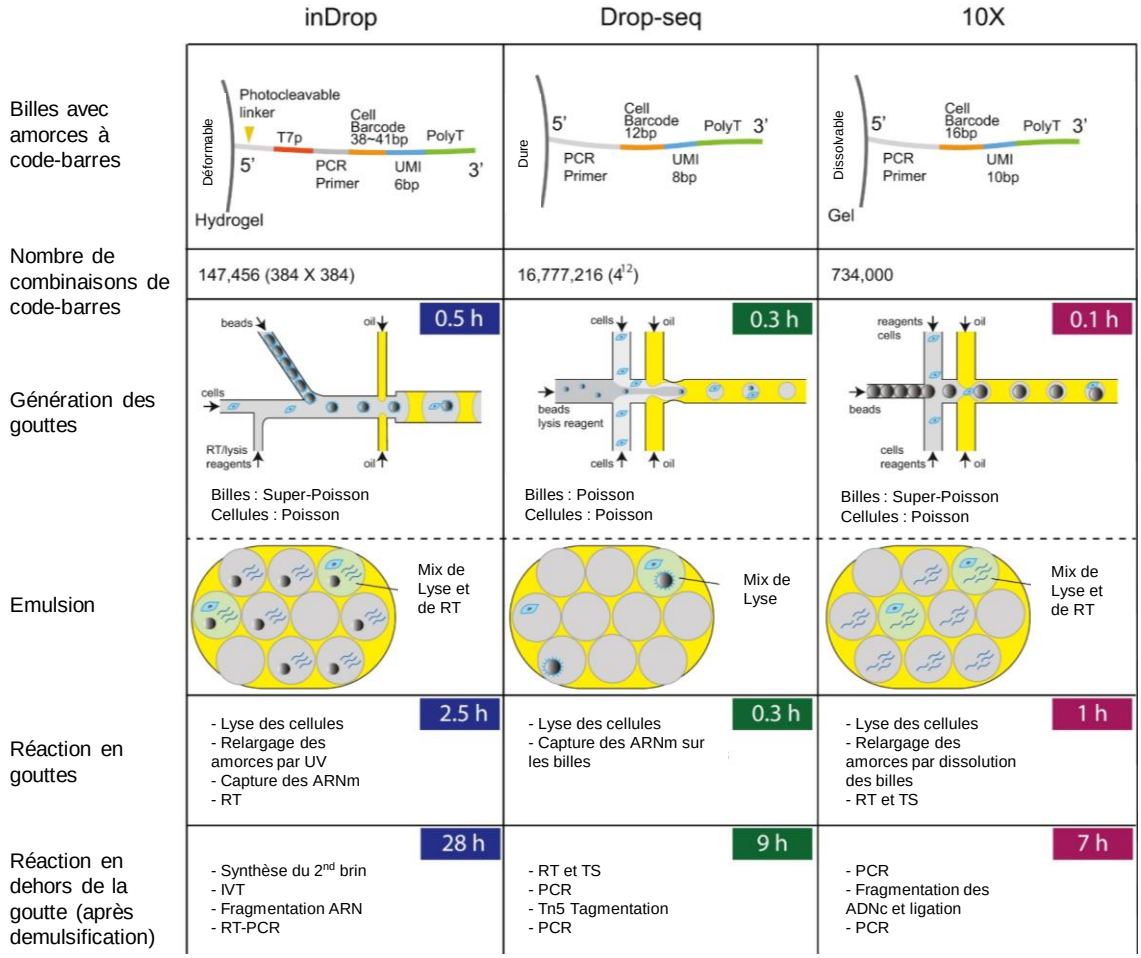


Figure 10 Schéma expérimental des 3 technologies scRNAseq en gouttes; inDrop, Drop-seq et 10x Chromium, Figure extraite et traduite à partir de (X. Zhang *et al.*, 2019)

D'autres études comparatives plus larges ont été faites sur les principales technologies scRNAseq afin d'évaluer leurs performances en utilisant un même modèle pour toutes les technologies. Ainsi, Ziegenhain *et al.* ont comparé 6 méthodes différentes en utilisant des cellules embryonnaires souches de souris (Ziegenhain *et al.*, 2017). Ils ont conclu que la technologie Drop-seq constitue la méthode la plus adaptée et rentable pour l'étude d'un échantillon large ne nécessitant pas une grande sensibilité, par exemple si le but est d'identifier des sous-populations de cellules. Si l'échantillon est précieux, la technologie MARS-seq sera privilégiée.

En ce qui concerne les performances en termes de sensibilité (ici défini par le nombre de gènes détecté) et de degré de couverture (la longueur des transcrits séquencés), le système SmartSeq2 a montré les meilleurs résultats. Il s'agit cependant de la méthode la plus coûteuse car nécessitant la plus grande profondeur de séquençage. Elle sera donc la méthode de choix si le but est l'annotation d'un transcriptome, la découverte de nouveaux transcrits ou variants d'épissage, sur un nombre limité de cellules. Enfin, les auteurs ont montré l'importance de l'utilisation des UMI. L'exactitude dans l'estimation de la concentration des différents transcrits dans les différentes cellules dépend de 2 facteurs ; le nombre de fois où les transcrits sont détectés dans les différentes cellules (soit la reproductibilité) et les biais d'amplification. Alors que la méthode SmartSeq2 est la meilleure concernant la reproductibilité dans la détection des transcrits, liée à une meilleure sensibilité, elle souffre de plus de biais d'amplification car elle n'utilise pas d'UMI. Un peu plus tard, Svensson *et al.* ont procédé à une étude à très grande échelle pour déterminer la sensibilité et la précision de nombreux protocoles scRNAseq, en ré analysant des données déjà publiées ou en procédant à des expériences à partir du mix ERCC Spike-in (Svensson *et al.*, 2017). Leurs conclusions sont similaires à celles de l'étude menée par Ziegenhain *et al.*

Tableau 1 Caractéristiques et performances de différentes technologies scRNAseq, inspirée de (Haque *et al.*, 2017) *IFC = Circuit Microfluidique Intégré, FACS = Fluorescent Activated Cell Sorting, TS = Template Switching, PCR = Polymérisation par Réactions en Chaine, IVT = In Vitro Transcription*

	C1 SmartSeq2	MARS-seq	inDrop	Drop-seq	10x Chromium	SPLIT-seq
Mode de capture des cellules	IFC (valves)	FACS + plaques	Microfluidiques en gouttes	Microfluidiques en gouttes	Microfluidiques en gouttes	Plaques + split-pool indexation
Amplification	TS + PCR	IVT	IVT	TS + PCR	TS + PCR	PCR
Couverture	Séquence totale	Biais 3'	Biais 3'	Biais 3'	Biais 3'	Biais 3'
Nombre de cellules	100 – 1000	100 – 1000	1000 – 10 000	1000 – 10 000	1000 – 10 000	1000 – 100 0000
Profondeur de séquençage par cellule recommandée	10^6	10^5	10^5	10^5	10^5	10^4
Volume réactionnel	Nanolitre	Microlitre	Nanolitre	Nanolitre	Nanolitre	Microlitre
Référence	(Pollen *et al.*, 2014)	(Jaiti &, Kenigsberg & Keren-Shaul, *et al.*, 2014)	(Klein *et al.*, 2015)	(Macosko *et al.*, 2015)	(Zheng *et al.*, 2017)	(Rosenberg *et al.*, 2018)

III. Etude de l'expression de gènes cibles à l'échelle de la cellule unique

A. Intérêt d'une étude transcriptomique ciblée

De nombreuses avancées ont vu le jour dans le domaine de la transcriptomique à l'échelle de la cellule unique. Les nouvelles technologies en scRNAseq permettent d'étudier de plus en plus de cellules en parallèle, avec toujours plus de profondeur. Cependant, elles sont limitées par les capacités des séquenceurs. Illumina a récemment mis sur le marché la technologie NovaSeq, permettant de séquencer jusqu'à 20 milliards de *reads* (sur une base d'un séquençage « single *read* » de 150 paires de bases). Cela peut sembler pharaonique mais en considérant qu'il faut 100 000 *reads* par cellule pour avoir une profondeur de séquençage suffisante pour analyser le transcriptome complet d'une cellule eucaryote, cela signifie qu'on peut au maximum analyser 200 000 cellules. S'il s'agit d'une expérience avec plusieurs conditions à comparer, ce chiffre peut être rapidement atteint.

Les technologies scRNAseq ont en commun l'utilisation d'amorces polyT pour la capture des ARN de la cellule, ce qui les limite à l'étude des ARN messagers porteurs d'une queue polyA en 3'. L'étude des ARN viraux, du transcriptome bactérien ou encore d'ARN non codants n'est ainsi pas compatible avec ces technologies. La technologie SUPeR-seq (X. Fan *et al.*, 2015) fait appel à des amorces possédant une séquence universelle, une séquence polyT et enfin une séquence aléatoire de 6 paires de base, afin de capturer à la fois des ARN polyadénylés ainsi que les ARN circulants non codants non polyadénylés chez des cellules uniques de souris embryonnaires. Cette méthode pourrait être employée afin de capturer les ARN non polyadénylés d'une cellule procaryote, voir même pour étudier les transcriptomes d'une cellule procaryote et de son hôte eucaryote et leurs interactions (Y. Zhang, Gao, Huang, & Wang, 2018). Cependant, les séquences aléatoires, contrairement à une amorce polyT pour les cellules eucaryotes, ne permettent pas de sélectionner les ARN messagers au détriment des ARN ribosomaux ou de transfert, qui représentent

pourtant une majorité des ARN contenus dans une cellule. L'utilisation d'amorces spécifiques pour cibler des gènes d'intérêt apparaît comme une solution alternative intéressante, même si de nombreux autres défis techniques restent à dépasser dans le cadre du séquençage de bactéries uniques ; la faible quantité de matériel allant de pair avec la courte durée de vie des transcrits bactériens ou encore la lyse de ces cellules à double couche membranaire.

Par ailleurs, la plupart de ces technologies, et notamment les technologies en gouttes de type inDrop, Drop-seq ou 10x Chromium, permettent d'obtenir l'information en 3' des ARN. Des segments situés beaucoup plus en amont (en 5') ne seront pas accessibles, comme c'est le cas des CDRs (Complementarity Determining Regions) des immunoglobulines (Saikia *et al.*, 2018) (Saikia *et al.*, 2018). Les technologies de type SmartSeq, qui permettent la lecture des ARN sur toutes leurs longueurs, demandent quant à elles beaucoup plus de profondeur de séquençage ce qui limite le nombre de cellules analysables en parallèle. Dans une revue sur l'étude des répertoires immunitaires des cellules B, différentes méthodes permettant le séquençage des segments codant pour les régions variables des chaines lourdes (VH, pour *Variable Heavy* = Variable Lourde) et légères (VL, pour *Variable Light* = Variable Légère) des anticorps sont présentées (Georgiou *et al.*, 2014). Ces techniques ont en commun de passer par une étape de ciblage des régions d'intérêt ; cela peut se faire dès l'étape de RT (Church, Vigneault, Laserson, & Bachelet, 2012), en utilisant des amorces ciblant les régions constantes des segments VH et VL, ou par capture des ARNm de cellules uniques sur billes magnétiques en gouttes, suivi par une étape d'émulsification des billes puis d'extension par chevauchement (*emulsified overlap extension RT PCR*) permettant de physiquement lier les segments VH et VL provenant de chaque bille (McDaniel, DeKosky, Tanno, Ellington, & Georgiou, 2016).

Enfin, les technologies scRNAseq actuelles ne permettent pas de récupérer l'information de transcrits peu abondants, qui peuvent pourtant s'avérer très informatifs, et on pourra leur préférer d'autres technologies telles que le smRNA FISH (Torre *et al.*, 2018). Le transcriptome d'une cellule a une gamme dynamique très large, de l'ordre de 5 ordres de grandeur, avec une grande majorité des transcrits étant codés

par une minorité de gènes, alors même que des gènes très faiblement exprimés peuvent avoir un fort impact fonctionnel, tels que certains facteurs de transcription. Cette contrainte est déjà évoquée dans le cadre du RNAseq populationnel, et des technologies RNAseq ciblées émergent pour assurer une meilleure couverture de gènes d'intérêt faiblement exprimés, selon plusieurs stratégies, par hybridation sur des amorces spécifiques de capture ou par amplification spécifique multiplexée. Une expérience visant à différencier des cellules de nature similaire, par exemple des cellules souches hématopoïétiques, nécessite l'utilisation de méthodes capables de détecter des transcrits faiblement exprimés, et des technologies ciblées s'avèrent plus adaptées (Andrews & Hemberg, 2018). Ainsi Mercer et al proposent un protocole détaillé de leur technologie CaptureSeq (Mercer *et al.*, 2014) pour sélectionner, à l'issu d'une préparation de librairie RNAseq classique à l'aide d'un kit commercial TruSeq de chez Illumina, des transcrits d'intérêts en utilisant des amorces de capture spécifiques à étiquette biotine et des billes magnétiques dont la surface est fonctionnalisée avec des streptavidines (Dynabeads M270 Streptavidin). Lors d'une étude comparative entre RNAseq classique, RT qPCR et CaptureSeq sur un mix de 92 ARN synthétiques commercial (ERCC Mix), cette dernière s'est avérée plus performante pour la détection des ARN synthétiques peu abondants (Clark *et al.*, 2015). Cette méthode a d'ailleurs déjà été employée pour la sélection de quelques 400 oncogènes d'intérêt par hybridation avec des amorces spécifiques à modification biotine lors d'une étude sur une lignée modèle de leucémie myéloïde (Levin *et al.*, 2009). En comparaison avec un RNAseq classique, le ciblage de régions d'intérêt a permis la détection de nouvelles mutations, de nouveaux isoformes d'épissage et de nouveaux produits de fusions entre transcrits tout en conservant une information quantitative sur les niveaux d'expression des cibles.

Un dernier avantage que peut offrir une étude ciblée par rapport à une étude du transcriptome complet est la possibilité de séquencer plus de cellules du fait d'un besoin moindre en profondeur de séquençage. Cela peut aider à identifier des populations très rares. De plus, de récentes technologies ont vu le jour permettant de coder les cellules provenant d'échantillons différents et de pouvoir ainsi toutes les regrouper pour réduire les coûts mais surtout les biais techniques entre échantillons préparés séparément (l'effet « batch »). Ainsi, Stoeckius et al modifient leur

technologie CITE-seq (Stoeckius *et al.*, 2017) pour marquer les cellules provenant de différents échantillons à l'aide d'anticorps, reconnaissant des protéines de surface ubiquitaires, et étant chacun étiqueté avec une séquence ADN (Stoeckius *et al.*, 2018). Cette séquence varie d'un échantillon à l'autre et sera lue en même temps que le transcriptome des cellules lors du séquençage. En plus de permettre de regrouper plusieurs échantillons ensemble et de réduire les biais techniques, cette solution offre également la possibilité de filtrer les données, en repérant les doublets de cellules ou les compartiments négatifs ne contenant que de l'ARN flottant. Une autre méthode propose d'intégrer ce code d'échantillon directement à l'intérieur des cellules par transfection (Shin, Lee, Lee, & Bang, 2019). En combinant ces méthodes avec une technologie ciblée de scRNAseq, un grand nombre d'échantillons pourrait être multiplexé, pour, par exemple, comparer l'impact de différentes drogues et doses sur la réponse immunitaire.

B. Exemples d'étude transcriptomique ciblée à l'échelle de la cellule unique

Au cours de cette section, 3 exemples de technologies en cellule unique permettant d'étudier le transcriptome de manière ciblée sont présentés.

1. DART-seq

Saikia et son équipe proposent une technologie, baptisée DART-seq pour *Droplet-Assisted RNA Targeting by single-cell sequencing* (Saikia *et al.*, 2018), permettant de surmonter plusieurs des limites des technologies scRNAseq faisant appel à des amorces polyT, énumérées précédemment. Ils détournent les billes à codes-barres de la technologie Drop-seq en liguant en 3' de l'amorce polyT, des séquences spécifiques pouvant cibler plusieurs régions d'intérêt. Les billes sont *in fine* porteuses d'amorces simple brin à code-barres uniques, d'une séquence polyT permettant de capturer les queues polyA des ARN messagers mais aussi (dans une proportion adaptable), de séquences spécifiques pouvant capturer des transcrits d'intérêt non polyadénylés (Figure 11 B). La capacité de capture des ARN messagers n'est pas affectée par cette modification, comme vérifié sur la capture du transcrit du gène de ménage *Gapdh*. Le protocole (Figure 11 A) est le même que celui de la technologie Drop-seq (Macosko *et al.*, 2015) ; une fois les transcrits polyadénylés et d'intérêt de chaque cellule unique capturés sur les billes, après encapsulation de billes et cellules en goutte, l'émulsion est cassée et la réaction de rétrotranscription se fait, en présence d'un oligonucléotide TSO permettant l'ajout d'une séquence connue en 3' des ADN complémentaires générés, selon la stratégie Smart Seq. Les librairies sont amplifiées avec des amorces complémentaires en 3' de la séquence TSO et en 5' de la séquence commune située en amont du code-barres et sont envoyées en séquençage.

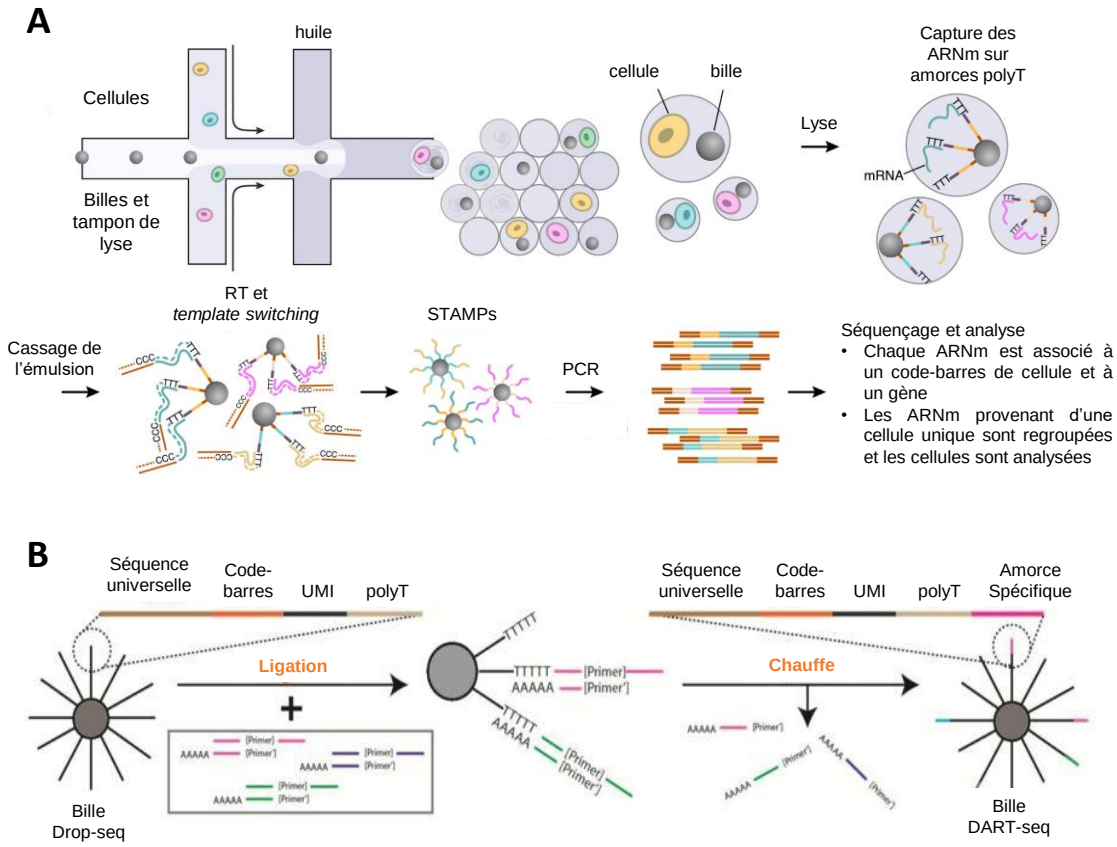


Figure 11 Stratégie DART-seq pour le séquençage de transcrits d'intérêt et du transcriptome complet de cellule unique en adaptant la technologie Drop-seq ; A= Protocole Drop-seq, extrait et traduit à partir de (Macosko *et al.*, 2015) *; B= Préparation des billes DART-seq, extrait et traduit à partir de* (Saikia *et al.*, 2018)

Ils ont dans un premier temps utilisé leur technologie pour étudier le transcriptome de cellules murines de type fibroblaste infectées par des orthoréovirus (souche Dearing de type 3) en parallèle du génotype des segments viraux (des ARN double brin non polyadénylés) qui ne sont pas détectables avec la technologie Drop-seq classique. Ils ont ainsi mis en évidence 4 sous-groupe de cellules infectées, dont l'un, associé à une surexpression de gènes immunitaires, n'est pas présent chez le contrôle négatif (cellules non infectées). Un autre sous-groupe, associé à une surexpression de gènes mitotiques, présente une quantité plus importante de segments viraux que dans les autres groupes. Ils ont également constaté que le génotype viral présente de nombreuses mutations ponctuelles, et notamment une forte transition G->A mais que celle-ci n'est que rarement observée dans les cellules infectées avec un fort taux de transcrits viraux, signe que ce type de mutation peut affecter l'efficacité de la transcription virale dans l'hôte. Il est à noter que seules les 200 paires de bases en

aval de la séquence cible sont exploitables. Il est donc nécessaire d'utiliser plusieurs sites spécifiques espacés d'environ 200 paires de bases pour obtenir la séquence complète du segment viral S2 auquel les auteurs s'intéressent ici.

Ils ont ensuite appliqué leur technologie pour une étude de l'immunité adaptée ; ils se sont intéressés au répertoire des cellules B. Les technologies permettant d'accéder aux paires VH-VL des IgG ont été évoquées dans le paragraphe précédent et sont un exemple de séquençage ciblé à l'échelle de la cellule unique. Les auteurs vont ici plus loin car ils nous donnent accès au transcriptome complet des cellules B en plus de l'information sur les régions variables des chaînes lourdes et légères. Ces séquences sont situées à quelques 800 paires de bases de l'extrémité 3' des ARN et ne sont donc pas accessibles avec un Drop-seq classique. En effet, la proportion des segments variables séquencés à l'aide de la technologie Drop-seq classique ne représente que 3% de tous les gènes identifiés, contre 30% à l'aide de la technologie DART-seq. Il est par ailleurs démontré au travers de ces expériences que l'ajout d'une amorce spécifique sur les billes n'affecte pas l'analyse du transcriptome complet de la cellule car autant de gènes et d'UMI sont identifiés à l'aide des 2 techniques (Drop-seq ou DART-seq). Il y a cependant une limite à leur technologie ; seules les cellules pour lesquels plus de 2000 UMIs sont comptabilisés ont une probabilité forte d'identifier la paire VH-VL, témoignant une nécessité de séquencer profondément pour obtenir l'information ciblée.

Cette technologie permet donc d'accéder à de nouveaux savoirs et ouvrent de nombreuses possibilités mais présentent cependant des limites. De même que la technologie Drop-seq, la technologie DART-seq n'est pas utilisable pour des échantillons rares du fait du faible taux de co-encapsulation de billes et de cellules dans une même goutte. Ensuite, étant donné que les billes de la technologie DART-seq capturent tous les ARN polyadénylés en plus des ARN d'intérêt, la profondeur de séquençage nécessaire par cellule est très élevée, limitant le nombre de cellules analysables en parallèle.

2. Single molecule RNA FISH

Une autre technique qui permet de s'intéresser à l'expression ciblée de certains transcrits et à l'échelle de la cellule unique, tout en ajoutant une dimension *in situ*, est la technologie smRNA FISH (Raj *et al.*, 2008) pour *single-molecule RNA Fluorescence in situ hybridization.* S'inspirant des techniques d'hybridation *in situ* conventionnelles, faisant appel à des sondes spécifiques liées à une enzyme catalysant une réaction colorimétrique (Raap *et al.*, 1995) ou à des sondes modifiées avec 5 fluorophores (Femino, Fay, Fogarty, & Singer, 1998), cette technologie consiste à venir hybrider, sur des cellules préalablement fixées, une multitude de sondes modifiées avec un fluorophore simple, de quelques 50 paires de bases, se fixant sur un transcrit d'intérêt. En hybridant des dizaines de sondes tout le long de la séquence d'un transcrit d'intérêt, la sensibilité est à son maximum, c'est-à-dire qu'une simple molécule est détectée. Du fait de limiter le marquage de chaque sonde à un seul fluorophore, les possibles effets d'extinction entre fluorophores physiquement proches, ou d'efficacité de marquage variable d'une sonde à l'autre, sont limités, évitant, respectivement, les risques de faux négatifs et de faux positifs. Enfin, contrairement à un marquage avec une réaction colorimétrique dont le produit fluorescent se diffuse, la fluorescence est maintenue en un point et l'information spatiale est conservée. Cette information peut se montrer primordiale en sachant que la cellule peut restreindre la localisation de ses ARNm pour assurer une réponse protéique ciblée, et cet aspect spatial se montre tout particulièrement intéressant lors d'une étude sur les cellules du développement. Le marquage de plusieurs transcrits en parallèle peut être analysé en utilisant des fluorophores aux spectres d'excitation distincts pour chaque transcrit ciblé, mais ce multiplexage est très limité.

Cette technologie a été utilisé pour l'identification de cellules tumorales résistantes (Shaffer *et al.*, 2017) ou encore pour l'étude de la dynamique d'expression de gène du métabolisme chez la levure face à un changement d'environnement (Schwabe & Bruggeman, 2014). Elle a aussi été utilisée comme outil de référence (ou golden standard) dans une étude comparative (Torre *et al.*, 2018) de 2 technologies de RNAseq à l'échelle de la cellule unique ; la technologie C1 mRNA-seq HT IFC de Fluidgm et la technologie Drop-seq.

3. CytoSeq

Développé en 2005, la technologie CytoSeq fait appel au système microfluidique des micro-puits (H. C. Fan *et al.*, 2015) pour charger dans les différents compartiments une cellule unique et une bille magnétique porteuses d'une amorce polyT à code-barres et possédant un UMI. La dimension des quelques 100 000 puits (30µm) assurent la présence d'une bille unique de 20µm par puits, tandis que les cellules sont chargées dans les puits par gravité, en suivant une loi de Poisson. La densité cellulaire de l'échantillon déposé est déterminée en fonction de cette loi statistique de sorte à éviter les doublets et environ 10% des puits sont chargés avec une cellule.

Les cellules une fois lysées et les ARN polyadénylés capturés sur les billes magnétiques, ces dernières peuvent être mises en commun dans un tube pour procéder à la synthèse des ADNc puis aux étapes d'amplification et de préparation au séquençage. Toutes les étapes de biologie moléculaire sont donc réalisées dans un tube unique, facilitant grandement la manipulation.

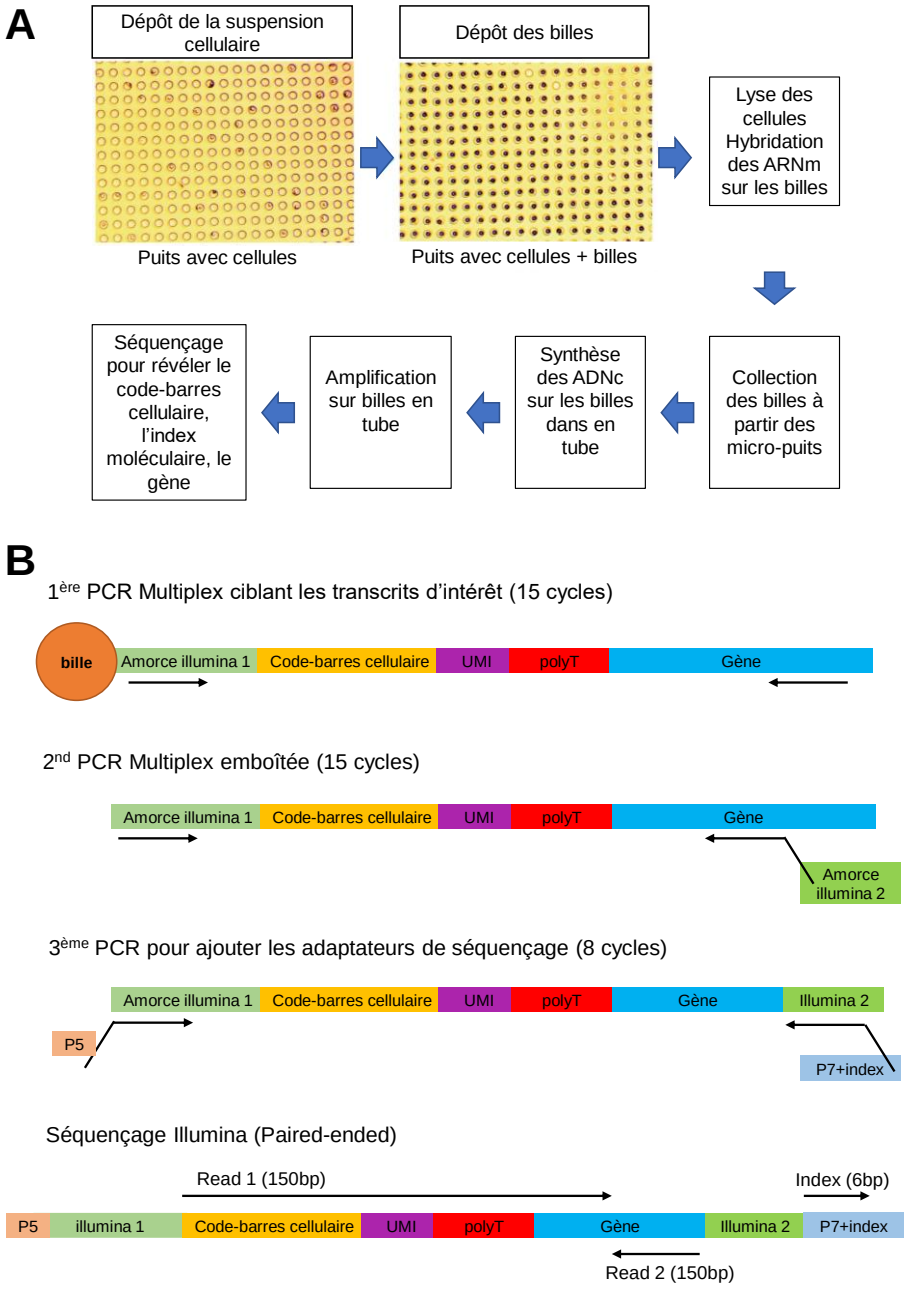


Figure 12 Description de la technologie CytoSeq A= Procédure expérimentale, B= Amplification des ADNc par PCR multiplex en amorces spécifiques en 3 étapes, Extraite et traduite à partir de (H. C. Fan *et al.*, 2015)

Dans la version initiale, les auteurs proposent de ne s'intéresser qu'à un panel de gènes prédéfinis, afin de diminuer la profondeur de séquençage par cellule. Comme mentionné précédemment, le séquençage optimal d'une cellule eucaryote unique, pour couvrir 10 fois le nombre total de transcrits polyadénylés (estimé à 200 000), nécessite jusqu'à 1 à 2 millions de *reads*, soit 2 milliards de *reads* pour analyser 1 000 cellules.

La préparation de l'échantillon se fait en procédant à une première PCR multiplexée afin de sélectionner les cibles (Figure 12) ; l'amorce antisens cible une séquence universelle située à l'avant du code-barres tandis que les amorces sens sont spécifiques des différents transcrits d'intérêt. Une seconde PCR multiplexée à l'aide d'amorces porteuses d'extension en 5' permet d'ajouter les adaptateurs nécessaires au séquençage. Les amorces sens utilisées pour cette deuxième étape vont venir se fixer à l'intérieur des amplicons générés lors de la première étape de PCR multiplexée, permettant d'augmenter la spécificité lors de la sélection selon un protocole de PCR emboitée (*nested PCR*).

En réduisant le nombre de transcrits ciblés à 100, la profondeur par cellule s'en voit grandement réduite et le séquençage de 1000 cellules ne nécessite plus que 10 millions de *reads*.

En appliquant leur méthode, ils ont été capables de distinguer des lignées cellulaires de nature différente. Ils ont également discriminé des cellules B cancéreuses mélangés à 1% avec des cellules B primaires saines, ce qui montrent la capacité d'identifier une sous population rare. Enfin, ils se sont intéressés à des échantillons primaires, pour discerner les différents sous-types au sein des cellules périphériques mononuclées du sang (PBMCs) ou encore pour montrer l'hétérogénéité de la réponse face à une stimulation de cellules T.

Leur système requiert moins d'équipements spécialisés que les circuits microfluidiques intégrés ou les microplaques automatisées. Il permet l'analyse d'un grand nombre de cellules, l'identification de populations rares et de cellules dans un état transcriptionnel transitoire. Comme toute méthode ciblée, tout le transcriptome n'est pas couvert, au risque de ne pas identifier certains sous-types, et une connaissance préalable de l'échantillon est nécessaire. Seule l'extrémité 3' des transcrits est considérée, du fait de la méthode de capture avec des amorces polyT, il y a donc un bais en 3'. Cela ajoute une contrainte pour la conception des amorces de PCR, déjà fastidieuse, rendant celle-ci difficile pour distinguer certains transcrits et complique la possibilité d'ajouter une jonction exon-exon au sein de l'amplicon d'intérêt, au risque d'avoir une contamination génomique.

Le système CytoSeq est aujourd'hui commercialisé sous le nom de BD Rhapsody (Shum, Walczak, Chang, & Christina Fan, 2019). Deux types d'étude sont aujourd'hui proposées en fonction du type de préparation des librairies choisi après synthèse des ADNc ; étudier le transcriptome entier en ajoutant par ligation un adaptateur universel à l'extrémité 3' des ADNc, qui servira de sites aux amorces de PCR ou étudier un panel de gènes d'intérêt en procédant à des amplifications avec des amorces spécifiques multiplexées.

Tandis que la première stratégie sera préférée dans le cadre d'une étude exploratoire sans connaissances préalables de l'échantillon à tester, la seconde offre l'avantage de se focaliser sur des transcrits d'intérêts, tels que des facteurs de transcriptions, parfois faiblement exprimés, et de ne pas récupérer de l'information inutile pour l'étude, telle que les gènes de référence, fortement exprimés par les cellules.

Dans leur revue sur ce système, Shum et son équipe propose une implémentation intéressante ; pouvoir mélanger plusieurs expériences en même temps et ainsi augmenter le débit d'analyse en réduisant les biais techniques entre échantillons. Ils proposent en effet d'utiliser la technologie des anticorps marqué avec un code-barre ADN pour multiplexer plusieurs échantillons. Cet anticorps reconnaissant un marqueur de surface universel est conjugué à un oligonucléotide (appelé TAG) contenant une queue polyA et un code-barre différent pour chaque expérience. La même stratégie peut être utilisée pour ajouter une dimension protéomique à l'étude transcriptomique, en utilisant plusieurs anticorps reconnaissants chacun un épitope différent et associé un code-barres unique à chaque type d'anticorps. De la même manière que les technologies CITEseq (Stoeckius *et al.*, 2017), REAPseq (Peterson *et al.*, 2017) ou Abseq (Shahi, Kim, Haliburton, Gartner, & Abate, 2017), cette méthode, baptisée BD Abseq permet d'associer une information protéomique, à savoir les marqueurs de surface d'une cellule, à une information transcriptomique, son expression génique. Utilisant ces différents outils, les chercheurs ont pu distinguer les différents sous-groupes au sein d'une population cellulaire complexe, les PBMCs, avec un degré de distinction accru quand les marqueurs de surface et l'expression de quelques 300 gènes d'intérêt étaient combinés. De même, en procédant à l'analyse des PBMCs de patients atteints d'arthrite rhumatoïde contre des patients sains, analyse multiplexée grâce à l'utilisation des anticorps avec un TAG d'échantillon, ils ont pu observer une diversité entre patient et contrôle dans l'une des sous-population, celle des monocytes.

Description du projet de thèse

Le projet s'est articulé autour de 2 applications différentes, avec comme élément central le développement d'une technologie RNAseq ciblé à l'échelle de la cellule unique en gouttes.

De nombreuses optimisations et validations ont été nécessaires dans le cadre de la mise au point de la technologie. Dans un premier temps, nous verrons les différentes expériences d'optimisation faites initialement en tube afin de déterminer les points clés à améliorer avant de passer en gouttes.

La mise au point de la technologie a continué à l'échelle de la cellule unique, en utilisant comme outil d'étude une lignée microgliale de souris. Les expériences menées nous ont permis d'établir un protocole de référence pour notre technologie et de tester différents paramètres, tels que l'efficacité de capture des ARN ou l'amplification des ADNc. Nous sommes ainsi parvenus à détecter des marqueurs de l'inflammation de manière significative et à l'échelle de la cellule unique.

Finalement, nous avons appliqué notre technologie à l'étude du réassortiment génétique chez le virus Influenza de type A (IAV). Les segments viraux de 2 souches IAV humaines ont pu être détectés à partir de notre technologie, et la souche d'origine a pu être identifiée. Des améliorations sont encore à prévoir mais les résultats sont encourageants pour la finalisation d'une première preuve de concept.

La dernière partie du manuscrit conclura quant aux résultats obtenus et aux perspectives futures.

Chapitre 2 : Matériel et Méthodes

I. Protocoles de biologie cellulaire

A. Culture cellulaire et activation de lignées BV2

Les lignées de cellules microgliales de souris dites BV2 m'ont été généreusement fournies par le Pr Pierre Gressens (UMR 1141, Hôpital Robert Debré). Elles sont cultivées dans des flasques T75 dans du Dulbecco's Modified Eagle's Medium (DMEM) pauvre en glucose et avec 1 g/L de pyruvate (Thermo scientifique #318851), supplémenté avec 10% de Sérum Fœtal Bovin inactivé à la chaleur (FBS, Thermo scientifique #16140071) et 1% d'antibiotiques Penicilin-Streptomycin (Thermo scientifique #15140122). Les cellules sont passées 2 à 3 fois par semaine. Pour cela, ces cellules adhérentes sont détachées du fond de la flasque par traitement avec 2mL de trypsine EDTA (Sigma #T4049) après que le milieu ait été retiré et que les cellules aient été rincées au DPBS 1x (Thermo Scientifique #141900941) pour retirer toute traces de Sérum. Le traitement est stoppé par ajout de 8mL de milieu complet et elles sont culotées en les centrifugeant à 200g pendant 4min. Le culot de cellules est suspendu dans 10mL de milieu frais et 200µL de cette suspension cellulaire sont ajoutés dans une flasque T-75 neuve pré remplie avec 10mL de milieu complet.

Pour une expérience d'activation, les cellules sont plaquées dans une plaque 6 puits pendant au moins 6h pour adhérence, avant ajout de la drogue. Pour une activation après 6 à 8h en plaque, il faut initialement ajouter 250 000 cellules par puits dans 2mL de milieu complet. Pour une activation après 18 à 24h, il faut préalablement ajouter 200 000 cellules par puits dans 2mL de milieu complet. Une fois l'adhérence stable et les cellules à moins de 70% de confluence, les cellules sont activées par ajout de 1µg/mL final de LPS (Sigma #L2880). Le contrôle négatif correspond à l'ajout d'un même volume de DPBS 1x. Les cellules sont incubées pendant 2h à 24h, les cellules sont prêtes à être analysées. Elles sont détachées par traitement trypsine, suspendues dans du DPBS 1x froid et comptage en chambre KOVA ou avec le Septum Millipore

II. Biologie moléculaire des ARN

A. Extraction de l'ARN total

L'ARN total est extrait en utilisant le kit RNeasy Mini kit (Qiagen #74106), sans traitement DNAse. Jusqu'à 5 millions de cellules sont culotées et sont suspendues dans 350µl de tampon RLT, supplémenté avec 3,5µL de beta Mercaptoéthanol 99%. 350µL d'éthanol 70% fraîchement préparé sont ajoutés et mélangés au lysat à l'aide d'une pipette P 1000. Le tout est déposé au fond de la colonne de purification RNeasy spin column et est centrifugé pendant 30 secondes à 8 000g. Le surnageant est jeté et 700µL de tampon RW1 est déposé au fond de la colonne qui est de nouveau soumise à une étape de centrifugation pendant 30 secondes à 11 000g. Le surnageant est éliminé et la colonne est lavée 2 fois avec 500µL de tampon RPE. Toute trace d'éthanol est éliminé en plaçant la colonne dans un tube de collection neuf et en la centrifugeant 2min à 11 000g. L'ARN est élué en ajoutant 40µL d'eau à la colonne et en la centrifugeant pendant 1 minute à 11 000g. Cette opération est répétée pour récupérer un maximum d'ARN. L'ARN total extrait est quantifié en utilisant le kit Qubit HS RNA (Thermo scientifique Q32852) et sa qualité est vérifiée en le soumettant à une tape station HS RNA (Agilent #5067) et en analysant le RIN (*RNA Integrity Number*), calculé à partir du rapport des intensités des bandes d'ARN ribosomaux 28s et 18s (Figure 13).

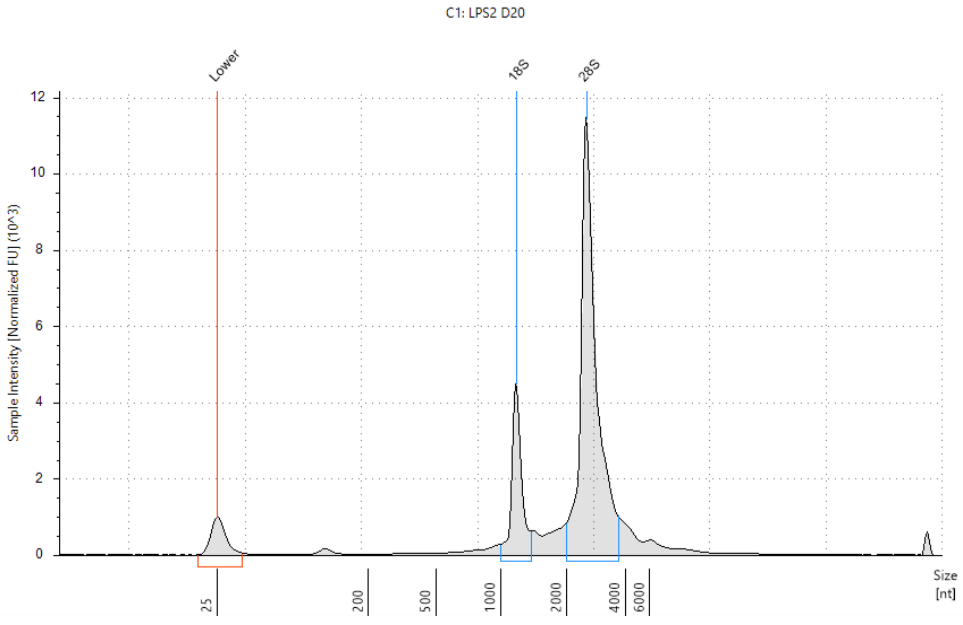


Figure 13 Migration d'ARN extrait à partir de cellules BV2 activées au LPS sur tape-station et détermination du RIN d'après le rapport des intensités des bandes ARN ribosomaux 28s et 18s pour évaluer la qualité de l'extrait

B. Production d'ARN synthétique

Des ARN synthétiques de séquences connues sont produits à partir d'ADN simple brin, qu'on nommera ultramère (IDT, séquences données en annexe). Cet ADN ultramère est dans un premier temps amplifié par PCR avec une amorce sens (gene_T7_forward) contenant la séquence du promoteur T7 en 5' et une amorce antisens contenant une séquence polyT en 5'. Pour cela, 0,2nM d'ultramère sont mélangés à 0,5µM de chaque amorce, 0,2mM de chaque dNTP, 1X de tampon Phu Hot Start Flex et 0,01 U/µL de Phusion Hot Start Flex (NEB #M0535S) dans un volume final de 100µL. La PCR suit un programme en 3 étapes classique, avec une température d'appariement de 56°C et 30 cycles d'amplification. L'amplicon généré est purifié sur billes magnétiques AMPure XP (Beckman Coulter #A63881), à un ratio volume de billes magnétiques : volume à purifier équivalent à 1,2. L'amplicon est élué dans 30µL d'eau dépourvu de DNAses et RNAses et est quantifié par qubit dsDNA HS (Thermo Scientifique Q32851). L'amplicon est ensuite transcrit en ARN en utilisant le kit HiScribe™T7 Quick High Yield RNA Synthesis Kit (NEB #E2050S) ; environ 3µg d'amplicon sont mélangés à du Tampon NTP buffer Mix 1X, 2µL de T7 RNA Polymerase Mix (NEB #E2050S) et 1U/µL de SUPERase IN RNAse Inhibitor (Thermo Scientifique #AM2694) dans un volume final de 30µL. Le mélange est incubé à 37°C pendant 15 heures avant d'être purifié avec des billes AMPure XP à un ratio 1,2x, puis élué dans 80µL d'eau et enfin soumis à un traitement DNAseI. Les 80µL sont ainsi supplémentés avec du tampon DNAseI à 1X et 0,1 U/µL de DNAseI (NEB #M0303S) et le mélange est incubé à 37°C pendant 50 minutes puis à 75°C pendant 5 minutes pour inactivation. Le tout est ensuite purifié en Gel PAGE 12%. Les bandes d'intérêt sont coupées et incubées dans 650µL de solution Sodium Acétate 0,3M sur la nuit à température ambiante. L'éluât est centrifugé à 13 000rpm, 4°C pendant 1 heure et le culot est lavé 2 fois à l'éthanol 70%. Toutes traces d'éthanol sont éliminées en passant les tubes au vacuum puis le culot est solubilisé dans 40µL d'eau avant quantification par Qubit HS RNA.

III. Préparation des outils microfluidiques

Les puces microfluidiques sont fabriquées par photolithographie. La résine photosensible SU-8 2025 est déposée sur la surface d'une plaquette de silicium propre, et est uniformément réparti en une couche d'une épaisseur de 40 µm par rotation (l'épaisseur désirée dépend de l'utilisateur et la vitesse de rotation appliquée de même que le type de résine influencent cette épaisseur). La résine, négative, est ensuite isolée avec des rayons UV qui passent à travers un photomasque (Services CAD / Art) qui a la propriété de stopper les rayons UV, sauf aux endroits où ont été imprimés les circuits microfluidiques, précédemment conçus sur le logiciel AutoCAD. Les zones exposées aux UV sont solidifiées par réticulation. La résine non réticulée est éliminée par lavage dans un tampon de développement. Ne reste donc que les zones qui ont été exposées à travers le photomasque. Le disque obtenu, avec les circuits microfluidiques en relief, sert de moule. Du polydiméthylsiloxane (PDMS) est mélangé à 10% (v/v) de Syligard 184, un agent de durcissement. Après le dégazage, le mélange est étalé sur le moule SU-8 et mis 2 heures à 70 ° C. La puce PDMS est extraite de la plaquette et des trous sont percés avec une pince à biopsie pour créer les entrées et les sorties des circuits. Le PDMS est ensuite lié à une lame de verre (côté design du PDMS faisant face à la lame de verre), après activation au plasma-oxygène des surfaces du verre et du PDMS. La puce est mise 10 min à 90 ° C. Pour finir, les bords des canaux sont traités avec du (1H, 1H, 2H, 2H-perfluorodécyltrichlorosilane (Sigma Aldrich, USA) dilués 10 fois dans de l'huile Novec 7100 (3M, France), pour les rendre lipophiles.

IV. Les billes d'hydrogel

A. Production de billes d'hydrogel (PEG DA ou PA)

1. Production de billes PEG DA

La solution aqueuse est préparée en mélangeant 10% (w/w) de Polyéthylène Glycol Di acrylate (PEG-DA)-6000 (Sigma #701963), 1% (w/w) de PEG-DA-700 (Sigma #455008), 400µM d'acrydite ADN double brin, dont le brin supérieur est porteur d'une modification 5'Acrydite et le brin inférieur d'une séquence cohésive de 4 bp en 5' et d'une modification 5'Phosphate, nommée AcRanA (IDT, séquence donnée en annexe), 1mM d'EDTA (Thermo-scientifique #15575-038) et 1% (v/v) de photo-initiateur (2-hydroxy-2-méthylpropiophenone, Sigma #405655), et 20µM de Fluorescéine isothiocyanate (FITC-Na), le tout suspendu dans un tampon contenant 75mM de Tris HCl Ph7,4 (Thermo scientifique #AM9851), 50mM de NaCl (Sigma #S3014). Le fluorophore et le photo-initiateur sont ajoutés à l'abri de la lumière. La phase aqueuse est poussée dans une puce microfluidique avec de l'huile fluorée supplémentée avec 2% de tensio-actif (di-krytox) à l'aide de pompes à un flux de 150µL/h et 500µL/h respectivement, permettant la production des gouttes de 9pL par concentration des flux. Le suivi de la taille des gouttes se fait via un système optique fait d'un laser à 488nm, de miroirs et de photomultiplicateurs, où le signal réémis par chaque goutte après excitation est réamplifié et analysé en temps réel via un logiciel conçu au laboratoire. Les gouttes sont collectées dans un tubing en PTFE de 0,3mm de diamètre, dont 3cm de longueur sont exposées à une irradiation UV (de 365nm de longueur d'onde et 360mW de puissance) pour assurer la polymérisation des gouttes en billes solides. La zone d'irradiation est confinée dans le noir. Les gouttes polymérisées sont finalement collectées dans un tube Eppendorf de 5 mL. Les billes passent alors par une étape de lavage ; après avoir retiré l'huile sous l'émulsion à l'aide d'une seringue, 4 mL d'hexane (Sigma #227064-1L) sont ajoutés, afin d'éliminer les traces d'huile et de tensio-actif. Les billes sont culotées par centrifugation pendant 10 secondes à 80g et l'hexane est retiré. Elles sont ensuite lavées 3 fois dans 4 mL de tampon de lavage BW (20 mM Tris-HCl pH 7.4, 50 mM NaCl, 0.1% of Tween 20) en procédant à des centrifugations à 3000g pendant 2 minutes. Enfin les billes sont filtrées à travers un filtre en Nylon de 20µm (Millipore, # SCNY00020) et stockées à 4°C dans du tampon de lavage BW supplémenté avec 1mM d'EDTA.

2. Production de billes PA

Les réactifs sont préparés en amont de la production, filtrés à travers membrane de 0,2µm et subdivisés dans plusieurs tubes stock. Une solution 4X AB est faite à partir de 14% (w/w) d'Acrylamide/bis-acrylamide (Sigma #A4058), 10% (w/w) de solution acrylamide (Sigma #A9926) dans de l'eau et est stockée à 4°C. Le tampon TBEST 1X consiste en 10mM de Tris HCl pH8 (Thermo scientifique # 15568025), 137mM de NaCl (Sigma #S3014), 2,7mM de KCl (Sigma # P9541) et 0,1% (v/v) de Triton X100 et est stocké à 4°C.

La solution aqueuse est préparée en mélangeant 1X de solution AB, 400µM d'acrydite ADN double brin, dont le brin supérieur est porteur d'une modification 5'Acrydite et le brin inférieur d'une séquence cohésive de 4 bp en 5' et d'une modification 5'Phosphate, nommée AcRanA (IDT, séquence donnée en annexe), 20µM de Fluorescéine isothiocyanate (FITC-Na), 0,3% (w/v) d'ammonium persulfate (APS) (Sigma # A9164) et 0,1X de tampon TBEST. La phase aqueuse est poussée dans une puce microfluidique avec de l'huile fluorée supplémentée avec 2% de tensio-actif (di-krytox) et 0,4% de tétraméthyléthylènediamine (TEMED) (Sigma #T7024) à l'aide de pompes, à un flux de 150µL/h et 500µL/h respectivement, permettant la production des gouttes de 10pL par concentration des flux. Des billes de 100pL peuvent également être produites à l'aide du design adapté et en poussant les flux aqueux et huileux respectivement à 200 et 400 µL/h. Le suivi de la taille des gouttes se fait via un système optique fait d'un laser à 488nm, de miroirs et de photomultiplicateurs, où le signal réémis par chaque goutte après excitation est réamplifié et analysé en temps réel via un logiciel conçu au laboratoire. Les gouttes sont collectées, via un tubing en PTFE de 0,3mm de diamètre, dans un tube Eppendorf de 2 mL prérempli avec 300µL d'huile minérale. Les billes sont incubées pendant une nuit à 65°C pour assurer leur polymérisation. 500µL de tampon TBEST 1X sont ajoutés au-dessus de l'émulsion qui est ensuite cassée par ajout de 1mL de perfluoro-octanol à 20% (dans de l'huile HFE 7500). Les billes sont culotées par centrifugation 30 secondes à 5000g et l'huile sous l'émulsion est retirée à l'aide d'une seringue. 1 mL d'hexane supplémenté avec 1% (v/v) de Span80 (Sigma #S6760) est ajouté, afin d'éliminer les traces d'huile. Les billes sont culotées par centrifugation pendant 30 secondes à 5000g et l'hexane est retiré. Cette opération est répétée une seconde fois. Finalement, les billes sont lavées 3 fois

dans 4 mL de tampon TBEST 1X en procédant à des centrifugations à 3000g pendant 3 minutes. Les billes sont stockées à 4°C.

B. Construction d'un code-barres sur bille d'hydrogel

Le protocole décrit ci-après correspond au protocole avant optimisation de la réaction de ligation, comme décrit au cours du chapitre 3. Le schéma expérimental reste identique après optimisation (seules les concentrations des ADN double brin à liguer changent)

250µL de billes culotées sont lavées 3 fois dans du tampon de lavage BW (20 mM Tris-HCl pH 7.4, 50 mM NaCl, 0.1% of Tween 20) par centrifugation à 4°C, 3000g pendant 2 minutes. A l'issu du dernier lavage, le surnageant est retiré et les billes sont soumises à une première étape de ligation. La réaction se fait dans 2 mL total avec du tampon ADN Ligase T7 1X, 12 µM de premier adaptateur double brin possédant deux extrémités cohésives différentes de 4 bp, la première étant complémentaire à l'extrémité cohésive de 4 bp sur les billes, et contenant un site de restriction BclI et la séquence d'adaptateur illumina *Read* 2 (Annexe, séquences), 30 U/µL d'ADN ligase T7 (NEB, # M0318L). Le mélange est incubé pendant 30 minutes sous agitation à 600 tours par minute. Les billes sont ensuite lavées 5 fois avec 4 ml de tampon de lavage BW et culotées après le dernier lavage. Pour la prochaine étape de ligation, un mix général de 1,6 mL est préparé à partir du culot de billes, du tampon ADN T7 ADN Ligase 1X, 30 U/µL d'ADN Ligase T7. Il est ensuite réparti dans les puits d'une plaque à 96 puits profonds (16 µL par puits), préremplie avec 4 µL de 20 µM d'index de code-barres (index A) (Annexe, séquences), où chaque puits contient une version d'index différente. Ces index sont pourvus de deux extrémités cohésives différentes de 4 bases, la première étant complémentaire de l'extrémité libre du premier adaptateur ligué à l'étape précédente (Annexe, séquences). La plaque est scellée et incubée à 25°C à 600 tours / min pendant 30 min. Ensuite, la ligase est inactivée par la chaleur à 65°C pendant 10 min et le contenu de tous les puits est regroupé en un tube Eppendorf 5mL maintenu sur glace, après addition de 200 µL de tampon de lavage BW froid supplémenté avec 1mM d'EDTA. Après regroupement, les billes sont lavées 7 fois avec 4 ml de tampon BW froid. Le processus complet de division et regroupement (ou split-and-pool) est répété pour lier les index de code-barres 2 (Index

B) et 3ème (Index C) (Annexe, séquences), générant une diversité totale de codes-barres de 8,8 x 10^5. Après la dernière étape de regroupement, les amorces de RT sont liguées. Elles sont partiellement double brin du côté 5', possédant une extrémité cohésive spécifique de l'extrémité libre de l'index D, et une séquence consensus de 15 bp. Elles sont simple brin en 3', avec un UMI de 5 bp et une amorce spécifique de RT (Annexe, séquences).

C. Contrôle qualité des billes en utilisant des sondes fluorescentes

Afin de déterminer la quantité relative d'une séquence d'intérêt sur les billes, celles-ci sont marquées à l'aide de sondes ADN fluorescentes spécifiques. 2µL de billes culotées sont tout d'abord dénaturées de sorte à avoir un code-barres simple brin et pouvoir hybrider une sonde. Pour cela, elles sont lavées à 3 reprises dans du tampon de dénaturation (150mM NaOH et 0,5% (w/w) Brij) par centrifugation 2 minutes à 3000g. Elles sont ensuite lavées 3 fois dans du tampon de neutralisation (100mM Tris-HCl pH8, 10mM EDTA, 0,1% Tween20 et 100mM NaCl), puis 3 fois dans du tampon d'hybridation (10mM Tris-HCl pH8, 0,1mM EDTA, 0,1% Tween20 et 330mM KCl). Après la dernière étape de centrifugation, le surnageant est retiré de sorte à laisser 40µL de culot (contenant les billes) et 10µM de sondes fluorescentes (FAM, Annexe, séquences) sont ajoutés puis le tout est placé à 70°C pendant 2 minutes avant de redescendre lentement à température ambiante pour hybridation (sous agitation à 600 tours / min). Les billes sont alors de nouveaux lavées 3 fois dans du tampon d'hybridation pour éliminer l'excédent de sondes puis observées en microscopie à épifluorescence.

D. Contrôle de la pureté des codes-barres par séquençage

Pour vérifier la diversité des amorces de code-barres sur les billes, des billes uniques ont été triées dans des puits contenant 4 µL d'eau dans une plaque à 96 puits à l'aide d'un trieur de cellules activé par fluorescence (FACS; ARIA III, BD Biosciences, San

Diego). à la plateforme de cytométrie et d'immunobiologie, CYBIO, de l'Institut Cochin à Paris. Ensuite, une réaction de transcription inverse, dans un volume total de 10 µL, par puits, par ajout de 0,1µM d'un rapporteur ARN synthétique à cheveux spécifique à chaque ligne (soit 8 ARN différents au total), 10 U/µL d'enzyme SSIII (SuperScript III Reverse Transcriptase, Thermo Scientifique #18080085), 0,43 U/µL d'enzyme BclI, dNTPs, DTT et 1X de tampon First Strand Synthesis. Cela génère des ADNc avec le code-barres de la bille couplée à la séquence du rapporteur ARN spécifique de la ligne. Après une heure d'incubation à 60°C, 2 µl de la réaction de RT de chaque puits sont transférés sur une nouvelle plaque et deux PCR séquentielles sont effectuées pour préparer les échantillons à séquencer. La première PCR est réalisée dans un volume de 20 µL en utilisant une amorce sens spécifique à la colonne et une amorce antisens commune (Annexe, séquences). Les produits de PCR sont purifiés séparément de chaque puits en utilisant 1,2 équivalent de billes magnétiques AMPure XP (Beckman Coulter #A63881) et remis en suspension dans 20 µL d'eau. 4 µL de produits purifiés sont utilisés pour la seconde PCR de 40 µL final, à partir d'amorces sens et antisens communes pour tous les puits. Après la deuxième PCR, les produits sont analysés sur des gels d'agarose à 1%. 71/96 puits ont montré une amplification. Ensuite, 1 µL de produits de PCR provenant des puits positifs est regroupé, purifié à l'aide de 1,2 équivalent de billes magnétiques AMPure XP et séquencé sur un système HiSeq 4000 (mode 2 * 150 High Output à BGI). Séquençage, Hong Kong, Chine). Les données sont analysées de sorte à identifier les codes-barres, les codes de lignes et colonnes ainsi que les UMI.

Aucun filtre basé sur le nombre de lectures par UMI n'a été appliqué. Les lectures normalisées UMI ont été associées à une bille à l'aide d'un code ligne-colonne. Pour chaque bille (n = 71), le pourcentage du code-barres le plus abondant a été calculé.

V. Séquençage du transcriptome total à partir d'ARN total

100ng d'ARN extraits sont mélangés avec 1X de tampon de First Strand Synthesis, 0,5mM de chaque dNTP, 5mM de DTT, 1 U/µL de SUPERase IN (Thermo-scientifique # AM2604), 10 U/µL de SSIII (Thermo-scientifique # 18080085) et une amorce contenant un promoteur T7, une séquence Illimina Rd2, une séquence consensus en enfin un site polyTVN, à 1µM. Le tout est incubé pendant 1 heure à 50°C puis 15 min à 70°C.

Les ADNc sont digérés par traitement à l'ExoI puis sont ensuite purifiés en utilisant 1,2 équivalent de billes magnétiques AMPure XP (Beckman Coulter #A63881). L'élution se fait dans 17µL d'eau, auquel sont ajoutés 1X de tampon Second Strand Synthesis et 1µL de mix enzymatique afin de générer le second brin complémentaire des ADNc. Le tout est placé à 16°C pendant 2,5 heures puis purifié avec 1,2 équivalent de billes magnétiques AMPure XP.

L'élution est faite dans 7µL de tampon d'élution ADN (10mM Tris-Cl pH 8.0, 0.1mM EDTA, filtré à travers une membrane de 0,2µm), auquel on ajoute 2µL de T7 RNA Polymérase (NEB #E2050S), 1 U/µL de SUPERase IN et 1X de tampon supplémenté avec un mix NTP. Pour amplifier le matériel par transcription *in vitro* à 37°C pendant 13 heures.

L'ARN résultant est purifié avec 1,3 équivalent de billes AMPure XP et les produits purifiés sont élués dans 16µL de tampon ARN d'élution (10mM Tris-Cl pH 7.5, 0.1mM EDTA, filtré à travers une membrane de 0,2µm). La moitié du produit élué est stockée à -80°C. On ajoute aux 7µL d'ARN purifié restants 0,5mM de chaque dNTP, 10µM d'amorce contenant une séquence adaptatrice illumina Rd1 en 5' et une séquence aléatoire de 8 bp en 3'. Le tout est incubé 3 minutes à 70°C puis placé 1 minute sur glace avant d'y ajouter 1X de tampon First Strand Synthesis, 5mM de DTT, 10 U/µL de SSIII Inhibitor (Thermo Scientifique #18080085) et 1 U/µL de SUPERase IN (Thermo Scientifique #AM2694). La seconde RT se fait par incubation 5 minutes à 25°C puis 1 heure à 50°C avant inactivation 15 minutes à 70°C.

Les ADNc néosynthétisés sont de nouveau purifiés avec 1,2 équivalent de billes AMPure XP puis élué dans 10µL d'eau. La dernière étape permet l'ajout des

adaptateurs illumina P5 et P7. Pour cela, les 10µL de produits sont mélangés avec 1X de tampon Phusion, 0,2mM de chaque dNTP, 0,02 U/µL de Phusion HS Flex (NEB #M0535S) et 0,5 µM de chaque amorce (P5 Rd1 et P7 Rd2). Le thermo cyclage se fait par incubation à 98°C pendant 30 secondes pour activation suivi par 15 cycles à 98°C pendant 15 secondes puis 72°C pendant 45 secondes. Une étape finale d'extension 3 minutes à 72°C achève cette amplification.

Les échantillons sont quantifiés pour séquençage à l'aide du kit de dosage Qubit dsHS DNA et su kit Kapa d'amplification des librairies illumina par qPCR (KapaBiosystems #KK4824)

VI. Séquençage ciblé de l'ARN en cellule unique

A. Encapsulation des billes et cellules en gouttes

Le protocole ci-dessous décrit l'encapsulation de billes de 10pL dans des gouttes de 130pL. L'encapsulation de billes de 100pL dans des gouttes de 2,5nL est similaire, avec les mêmes mix utilisés. Seuls les flux appliqués diffèrent, ils sont de 250µL/h pour les phases contenant le mix cellulaire et celles contenant les enzymes, de 50 à 100µL/h pour les billes, et de 500µL/h pour l'huile HFE 7500 supplémentée avec 2% de tensio-actif dikytox.

Les billes à codes-barres sont lavées 2 fois avec 1mL de tampon de lyse 3X (0,6% Triton X100, 150mM Tris HCl pH 7,4) en les culotant par centrifugation 1 minute à 3000g et en retirant le surnageant. Les opérations sont effectuées sous hotte à flux laminaire. Le culot de billes est ensuite incubé à température ambiante pendant 10 minutes dans du tampon RT/Lysis (10mM DTT, (0,6% Triton X100, 150mM Tris HCl pH 7,4, 2X de tampon First Strand Synthesis de chez Thermo scientifique). Les billes sont culotées 1 min à 3000g. Le surnageant est totalement retiré et les billes sont conservés à 4°C.

Le mix de RT est préparé à partir de 1X de tampon de First Strand Synthesis, 1,5mM de chaque dNTP, 5mM de DTT, 1µM de Dy647, 3 U/µL de SUPERase IN (Thermo-scientifique # AM2604), 30 U/µL de SSIII (Thermo-scientifique # 18080085) et 0,43 U/µL de BclI (NEB #R0160). Il est conservé à 4°C jusqu'à encapsulation.

La phase cellulaire est préparée en mélangeant des cellules dans du DPBS 1x à une densité de $2,5x10^6$ cellules/ml avec 1,7 mg/mL de méthylcellulose.

Trois seringues de 1mL, reliées via du tubing PTFE de 0,56mm de diamètre à un cône 200µL fermé par un embout PDMS, sont préremplies avec de l'huile minérale (Sigma #5904). Tout le circuit est rempli d'huile et l'air est entièrement chassé. Ces seringues, ainsi qu'une seringue remplie d'huile HFE 7500 supplémenté avec 2% de tensio-actif dikrytox et relié à un tubing en PTFE de 0,33mm de diamètre, sont installés sur des pompes NEMESYS. Les 3 phases aqueuses (mix cellulaire, mix de RT et billes), sont aspirées à un flux de 1000µL/h à l'intérieur des cône remplis d'huile minérale. Les 3 phases aqueuses et l'huile sont connectées à une puce microfluidique, et la sortie est connectée via du tubing en PTFE 0,33mm de diamètre à un tube Eppendorf de 1,5mL scellé avec un embout PDMS dont sorte un tubing d'entrée et un tubing de sortie, et

prérempli avec de l'huile fluorée supplémentée avec du tensio-actif à 2%. Ce système permet d'éviter tout contact de l'émulsion avec de l'air et ainsi limiter les phénomènes de coalescence entre gouttes. Le remplissage des entrées est amorcé en poussant chaque phase à 200µL/h. Une fois les différentes phases arrivées dans le circuit, les flux finaux sont appliqués à savoir 100µL/h pour les 3 phases aqueuses et 500 à 600µL/h pour l'huile. Le flux d'huile peut être ajusté (par palier de 5µL/h) de sorte à produire des gouttes de 130pL, avec un taux d'encapsulation de l'ordre de 80% de gouttes contenant une bille. La taille des gouttes est mesurée grâce au fluorophore et à un laser d'une longueur d'onde de 647nm. Le retour de fluorescence est envoyé sur des photomultiplicateurs et analysé à l'aide du logiciel µFluidique mis au point au laboratoire.

Les gouttes sont collectées pendant 10 minutes puis les émulsions sont placées avant 1 heure à 55°C puis 20 minutes à 70°C. Après cette étape d'inactivation, la couche d'huile minérale, située au-dessus de l'émulsion, ainsi que le couche l'huile HFE avec tensio-actif, située au-dessus de l'émulsion, sont retirées à l'aide d'une seringue et un équivalent volumique de solution démulsifiante (20% (v/v) perfluoro-octanoll dans de l'huile HFE 7500) est ajoutée. Les émulsions ainsi cassées sont stockées à -80°C ;

B. Préparation des librairies de séquençage

1. Elimination des amorces à code-barres contaminantes par traitement enzymatique

Les échantillons de RT sont décongelés sur glace puis mis à centrifuger 5 minutes à 19 000g à 4°C. La phase aqueuse est passée à travers un filtre de 0,45µm (Costar SpinX) par centrifugation à 4°C, 16 000g pendant 5 minutes et le surnageant est récolté. 20µL de solution de digestion, contenant 1 U/µL d'ExoI (NEB #M0293) et 3X de tampon enzymatique associé pour 40µL d'échantillon de RT sont de nouveau passés à travers la membrane et mis à centrifuger à 4°C, 16 000g pendant 5 minutes. Le surnageant est récupéré et ajouté à celui déjà collecté et le tout est placé à 37°C pendant 30 minutes.

2. PCR multiplex

Les ADNc digérés sont purifiés en utilisant 1,2 équivalent de billes magnétiques AMPure XP (Beckman Coulter #A63881). L'élution se fait dans 10µL d'eau, auquel sont ajoutés 0,2mM de chaque dNTP, du buffer HF à 1x, 0,02 U/µL de Phusion HS Flex (NEB), 0,5µM d'amorce anti-sens T7as, 1,6 mM de MgCl2 additionnel, et 0,05 µM de chaque amorce cible, possédant une extension 5'Rd1. Les produits sont amplifiés pendant 30 cycles à une température d'hybridation de 67°C. Les produits amplifiés sont purifiés en utilisant 1 équivalent de billes magnétiques AMPure XP. On peut passer à la PCR finale.

3. Transcription *in vitro*

Les ADNc digérés sont purifiés en utilisant 1,2 équivalent de billes magnétiques AMPure XP (Beckman Coulter #A63881). L'élution se fait dans 17µL d'eau, auquel sont ajoutés 1X de tampon Second Strand Synthesis et 1µL de mix enzymatique afin de générer le second brin complémentaire des ADNc. Le tout est placé à 16°C pendant 2,5 heures puis purifié avec 1,2 équivalent de billes magnétiques AMPure XP.

L'élution est faite dans 7µL de tampon d'élution ADN (10mM Tris-Cl pH 8.0, 0.1mM EDTA, filtré à travers une membrane de 0,2µm), auquel on ajoute 2µL de T7 RNA Polymérase (NEB #E2050S), 1 U/µL de SUPERase IN et 1X de tampon supplémenté avec un mix NTP. Pour amplifier le matériel par transcription *in vitro* à 37°C pendant 13 heures.

L'ARN résultant est purifié avec 1,3 équivalent de billes AMPure XP et les produits purifiés sont élués dans 16µL de tampon ARN d'élution (10mM Tris-Cl pH 7.5, 0.1mM EDTA, filtration à travers une membrane de 0,2µm). La moitié du produit élué est stockée à -80°C. On ajoute aux 7µL d'ARN purifié restant 0,5mM de chaque dNTP, 10µM d'amorce contenant une séquence adaptatrice illumina Rd2 en 5' et une séquence aléatoire de 8 bp en 3'. Le tout est incubé 3 minutes à 70°C puis placé 1 minute sur glace avant d'y ajouter 1X de tampon First Strand Synthesis, 5mM de DTT, 10 U/µL de SSIII Inhibitor (Thermo Scientifique #18080085) et 1 U/µL de SUPERase IN (Thermo Scientifique #AM2694). La seconde RT se fait par incubation 5 minutes à 25°C puis 1 heure à 50°C avant inactivation 15 minutes à 70°C. Les ADNc néosynthétisés sont de nouveau purifiés avec 1,2 équivalent de billes AMPure XP puis élués dans 10µL d'eau. On peut passer à l'étape finale.

4. PCR finale et quantification

La dernière étape permet l'ajout des adaptateurs illumina P5 et P7. Pour cela, les 10µL de produits à séquencer sont mélangés avec 1X de tampon Phusion, 0,2mM de chaque dNTP, 0,02 U/µL de Phusion HS Flex (NEB #M0535S) et 0,5 µM de chaque amorce (P5 Rd1 et P7 Rd2). Le thermo cyclage se fait par incubation à 98°C pendant 30 secondes pour activation suivi par 5 à 15 cycles à 98°C pendant 15 secondes puis 72°C pendant 45 secondes. Une étape finale d'extension 3 minutes à 72°C achève cette amplification.

Les échantillons sont quantifiés pour séquençage à l'aide du kit de dosage Qubit dsHS DNA et du kit KAPA d'amplification des librairies illumina par qPCR (KapaBiosystems #KK4824)

VII. Analyse des données de séquençage (cas du RNAseq ciblé en gouttes sur cellules BV2)

Les échantillons sont envoyés à la plateforme de l'ICM pour un séquençage en mode *Paire-Ended*, en lisant 100bp du côté où se situe le code-barres (Rd2) et 50 bp de l'autre côté, pour l'identification du gène. Les fichiers Bcl2 sont démultipliés en fonction de l'index Illumina reconnu et les données renvoyées par la plateforme consiste en un dossier contenant les différents fichiers fastq associés à chaque index et à chaque sens de lecture.

La première étape consiste à analyser la qualité du séquençage en utilisant l'outil Fastqc. En cas de mauvaise qualité, habituellement en fin de séquences, les séquences peuvent être tronquées. Les séquences obtenues sont ensuite alignées contre des génomes de référence afin d'identifier, pour chaque molécule séquencée, le code-barres, l'UMI et le gène transcrit associés.

Le démultiplexage des codes-barres se fait en alignant les fichiers fastq Rd2 contre des fichiers de références fasta construits à partir des 96 versions d'index B, C et D et en utilisant l'outil d'alignement Bowtie2 (Annexe 2, *Supplementary File 1, Droplet barcode, UMI, IGS, tag and hairpin RNA reporter identification from sequencing data*).

L’identification des transcrits se fait en alignant les fichiers fastq Rd1 contre des fichiers de références fasta construits à partir des séquences ciblées en utilisant l’outil d’alignement Bowtie2, et l’option --local.

Un fichier de type csv est créé à partir des informations obtenues suite aux alignements précédents et les séquences d’une longueur de 5 bases des UMI sont extraites, ainsi que l’identifiant de *read* auxquels ces UMI sont associés, à partir des fichiers Rd2, en localisant une séquence consensus situé en amont de la séquence UMI.

Les 3 informations, à savoir le code-barres, le gène et l’UMI, sont combinées en se basant sur l’identifiant de *read*.

Un premier filtre sur le nombre de *reads* par code-barres est défini après génération d’un histogramme de la fréquence du nombre de *reads* par code-barres. Ce graphe montre une distribution bimodale, où les codes-barres comptant peu de *reads* correspondent à des codes-barres contaminants devant être éliminés avant de procéder aux analyses.

Une matrice est finalement générée à partir du fichier csv contenant les informations code-barres, transcrit et UMI et en appliquant le filtre sur le nombre de *reads* par code-barres défini grâce à l’histogramme.

Des graphes peuvent alors être générés à partir de la matrice d’expression générée.

Chapitre 3 : OPTIMISATION

I. Introduction

Nous cherchons à développer une technologie de RNAseq ciblée à l'échelle de la cellule unique, en faisant appel à la microfluidique en gouttes. Ce développement débute par une phase d'optimisation que je décrirai au cours de ce chapitre.

Comme mentionné en introduction, les 3 principales technologies scRNAseq à ultra haut débit faisant appel à la microfluidique en gouttes sont les technologies inDrop, Drop-seq et 10X Chromium. Le principe est similaire pour les 3 ; des cellules uniques sont encapsulées dans des gouttes de quelques nanolitres, où elles sont ensuite lysées pour libérer leur ARNm. Ces ARNm sont alors capturés par des amorces à code-barres, délivrées par une bille. Cette bille permet d'apporter dans chaque goutte un code-barres différent ; en effet, chaque bille est fonctionnalisée avec 10^8 à 10^9 amorces, contenant une queue polyT, un UMI, et une séquence ADN code-barres, unique pour chaque bille, mais différente d'une bille à l'autre. Ce code-barres indique l'origine cellulaire de chaque ARN et permet de multiplexer de nombreuses cellules en parallèle. Une fois capturés, les ARNm sont rétro-transcrits en ADNc, qui sont porteurs d'un code-barres et de la séquence du transcrit. Ceux-ci peuvent être regroupés pour être amplifiés et préparés afin d'être séquencés.

Les protocoles diffèrent néanmoins à plusieurs niveaux :

- La nature chimique de la bille et son mode d'encapsulation :

Alors que les technologies inDrop et 10X Chromium utilisent des billes d'hydrogel déformables et ultra concentrées ce qui permet d'encapsuler des billes uniques une à une au-delà de la densité de Poisson (environ 80% des gouttes contiennent une bille), la technologie Drop-seq utilise des billes en résine dure, qui sont encapsulées suivant une loi de Poisson. Finalement, pour un même lambda de 0,1, 7% des gouttes contiennent une bille et une cellule dans le cas de 10X Chromium et de inDrop, contre seulement 1% pour Drop-seq.

- La structure des amorces et la génération des code-barres :

Les amorces inDrop sont dotées d'un site photo clivable permettant le relargage de celles-ci dans les gouttes sous l'action des UV, d'un promoteur T7 pour initier l'amplification par IVT, d'une séquence code-barres, puis enfin d'un UMI et d'une queue polyT. Le code-barres est composé de 2 séquences nommées index, existant chacune en 384 versions. Il y a donc environ 10^5 combinaisons théoriques de code-barres.

Les amorces Drop-seq sont constituées d'un code-barres, d'un UMI et d'une queue polyT. Le code-barres est construit en 12 étapes de synthèse ADN, où les billes sont tour à tour divisées aléatoirement entre 4 compartiments, correspondant chacun à la synthèse d'une des 4 paires de bases, puis regroupées, permettant de créer plus de 10^7 combinaisons différentes.

Les amorces des billes 10x Chromium sont constituées d'un code-barres qui existe en $8x10^5$ combinaisons, d'un UMI et d'une queue polyT.

- La réaction de RT :

Tandis que la rétrotranscription des ARNm en ADNc se fait à l'intérieur de la goutte pour les technologies inDrop et Drop-seq, celle-ci se fait après cassage des émulsions pour la technologie Drop-seq, sur les billes à code-barres qui ont capturé les ARN polyadénylés.

- Le mode d'amplification des ADNc :

Deux stratégies différentes sont employées selon la technologie ; une amplification linéaire par transcription *in vitro* pour inDrop, et une amplification exponentielle par PCR après ajout d'un oligonucléotide en 3' des ADNc par une réaction de *template switching* pour Drop-seq et 10x Chromium.

A l'aide de ces protocoles, nous avons mis au point notre technologie de séquençage ciblé de l'ARN ciblé à l'échelle de la cellule unique (Tableau 2).

Pour ce qui est de la nature chimique des billes, la structure du code-barres et la réaction de RT, nous nous sommes inspirés de la technologie inDrop. Nous souhaitons en effet mettre au point une technologie compatible avec l'analyse d'échantillons rares, et les propriétés de déformement des billes d'hydrogel et leur compaction offrent l'avantage d'encapsuler les billes à une densité allant au-delà de la loi de Poisson. Nous avons apporté quelques modifications, de sorte à cibler spécifiquement des transcrits d'intérêt. La principale différence entre notre technologie et celles décrites précédemment se situe au niveau du mode de capture des ARNm et de l'amorce de RT utilisée ; contrairement aux 3 technologies scRNAseq en gouttes qui capturent les ARNm à l'aide d'une amorce polyT, nous faisons appel à des amorces spécifiques de transcrits d'intérêt. Certaines technologies de RNAseq ciblées utilisent des amorces polyT et sélectionnent les transcrits uniquement lors de l'étape d'amplification (c'est par exemple le cas de la technologie CytoSeq). Cependant, l'utilisation d'une amorce spécifique offre plusieurs avantages. Tout d'abord, ce type d'amorces est compatible avec d'autres systèmes biologiques que les cellules eucaryotes telles que les bactéries ou les virus, dont l'ARNm n'est pas polyadénylé. Ensuite, il n'y a pas de biais en 3' et des transcrits situés en 5' des gènes peuvent être analysés, alors qu'ils seraient difficilement capturés avec une amorce polyT. Enfin, en ne capturant que les transcrits d'intérêt, on évite un échantillonnage aléatoire des transcrits capturés pour privilégier la capture de ceux qui nous intéressent. Nous espérons ainsi améliorer la sensibilité de la technologie, en détectant plus efficacement les transcrits ciblés, et ce, même s'ils sont faiblement exprimés.

Notre technologie de RNAseq ciblé en gouttes, schématisée dans la Figure 14, consiste à encapsuler dans des gouttes de quelques centaines à quelques milliers de picolitres un code-barres unique et une cellule unique. Le code-barres est apporté dans la goutte par une bille d'hydrogel. Il comprend un site de restriction, permettant la libération des amorces dans la goutte sous l'action d'une enzyme de restriction, un promoteur T7, un site universel illumina, une séquence ADN code-barres, un UMI et enfin une amorce spécifique. Chaque bille est porteuse de centaines de millions de molécules de code-barres, mais chacune possède une version unique de ce code-barres, qui est différent d'une bille à l'autre. Les amorces sont des séquences spécifiques ciblant des transcrits d'intérêt, libérés par la cellule dans la goutte. Les ADNc sont générés à partir des amorces à code-barres par une réaction de RT qui s'effectue dans la goutte. Ce n'est qu'une fois la molécule d'ADNc synthétisée que l'émulsion est cassée et que toutes les molécules sont mélangées. Elles sont alors amplifiées et on leur ajoute des extensions qui permettront leur séquençage en utilisant la technologie Illumina.

A l'issu du séquençage, une matrice d'expression est générée, après identification des codes-barres et des transcrits par alignement contre un génome de référence. Cette matrice donne accès au niveau d'expression des différents transcrits cibles dans chaque cellule. L'utilisation d'un UMI pour la quantification permet de s'affranchir des biais d'amplification. La matrice sert de base aux différentes analyses. Nous nous intéressons ainsi ;

- au nombre de transcrits exprimés par cellule pour chaque cible
- à l'expression relative des transcrits dans les différentes cellules et le regroupement de ces dernières en sous-groupes en fonction de leur profil d'expression
- à l'éclatement des cellules en sous-groupes par une projection en 2 dimensions en utilisant des algorithmes de réduction dimensionnelle.

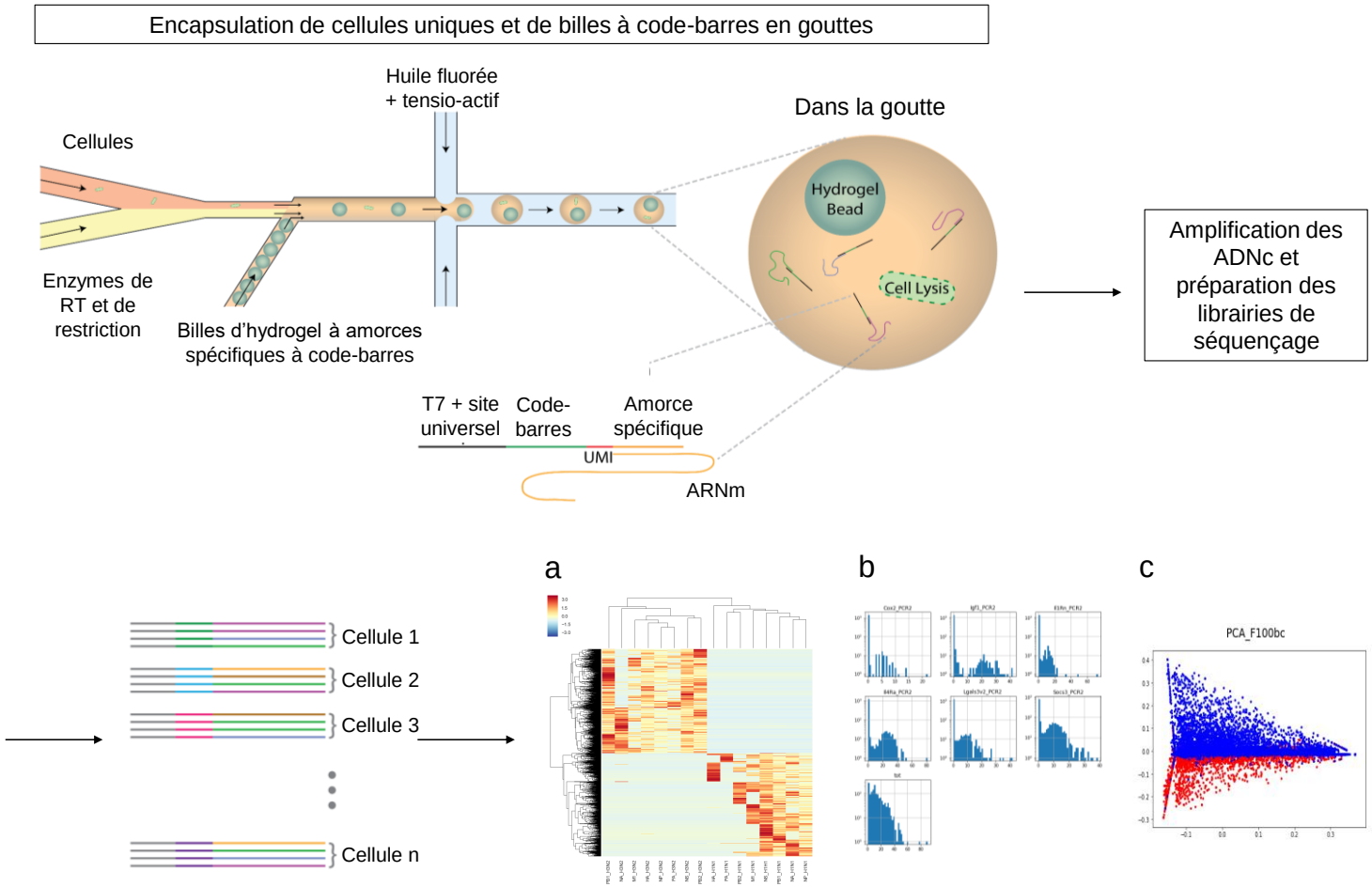


Figure 14 Schéma du déroulement du RNAseq ciblé sur cellule unique en gouttes

Technologie	inDrop	10x Chromium	Drop-seq	Ciblée
Nature des billes	Polyacrylamide 100pL (*Acrylamide bis Acrylamide* + *Acrydite DNA*)	Polyacrylamide avec ponts disulfites (*Acrylamide-bis(acryloyl)cystamine* + *acrydite-S-S-DNA*)	Microparticules en résine dure	2 chimies possibles • Polyacrylamide (*Acrylamide bis Acrylamide* + *Acrydite DNA*) • PEG DA (*PEG DA* + *Acrydite DNA*)
Amorces à code-barres	T7-BC-UMI-PolyT ~ 150 000 combinaisons	BC – UMI- polyT ~ 750 000 combinaisons	BC-UMI-polyT ~16 000 000 combinaisons	T7-BC-UMI-GSP ~800 000 combinaisons
Réaction dans la goutte	• Lyse des cellules • Relargage des amorces capture et des ARNm totaux • RT en ADNc	• Lyse des cellules • Relargage des amorces et capture des ARNm totaux • RT en ADNc • *Template switching*	• Lyse des cellules • Capture des ARNm polyadénylés totaux	• Lyse des cellules • Relargage des amorces et capture de transcrits d'intérêt • RT en ADNc
Amplification des ADNc	• IVT • RT avec amorces aléatoire • PCR finale	• PCR	• RT et *template switching* • PCR	2 options • PCR • IVT suivi de RT avec amorces spécifiques

Tableau 2 Principales caractéristiques des 3 technologies RNAseq cellule unique en gouttes de référence et de la technologie ciblée que nous développons.

Au cours de ce chapitre, les différentes étapes de mises au point de notre technologie seront détaillées. Des tests ont été réalisés de manière systématique afin d'optimiser chaque étape clé. Les différents paramètres investigués sont présentés dans les encadrés orange de la Figure 15.

L'un des points déterminants de l'efficacité de la technique est l'étape de capture des ARNm et leur rétro transcription en ADNc au sein de la goutte. Les billes d'hydrogel étant la source des amorces à code-barres, à partir desquelles est initiée la réaction de RT, constituent un point déterminant dans l'efficacité de la réaction. Il est donc crucial de se pencher sur leur caractérisation et leur optimisation.

Nous nous sommes également intéressés aux étapes suivant la réaction de RT et le cassage des émulsions. De nombreux protocoles de préparation de libraires de séquençages existent, avec leurs avantages et leurs inconvénients. Il a fallu vérifier lequel était le plus adapté à notre application.

Toutes ces optimisations ont été faites sur une lignée cellulaire de microglies de souris, la lignée BV2, la microglie étant mon application de départ, ou sur des ARN synthétiques. Celle-ci sera décrite plus en détails au cours du prochain chapitre. La plupart des expériences ont été faites dans un premier temps en tube et non en gouttes, pour des raisons évidentes de simplicité expérimentale.

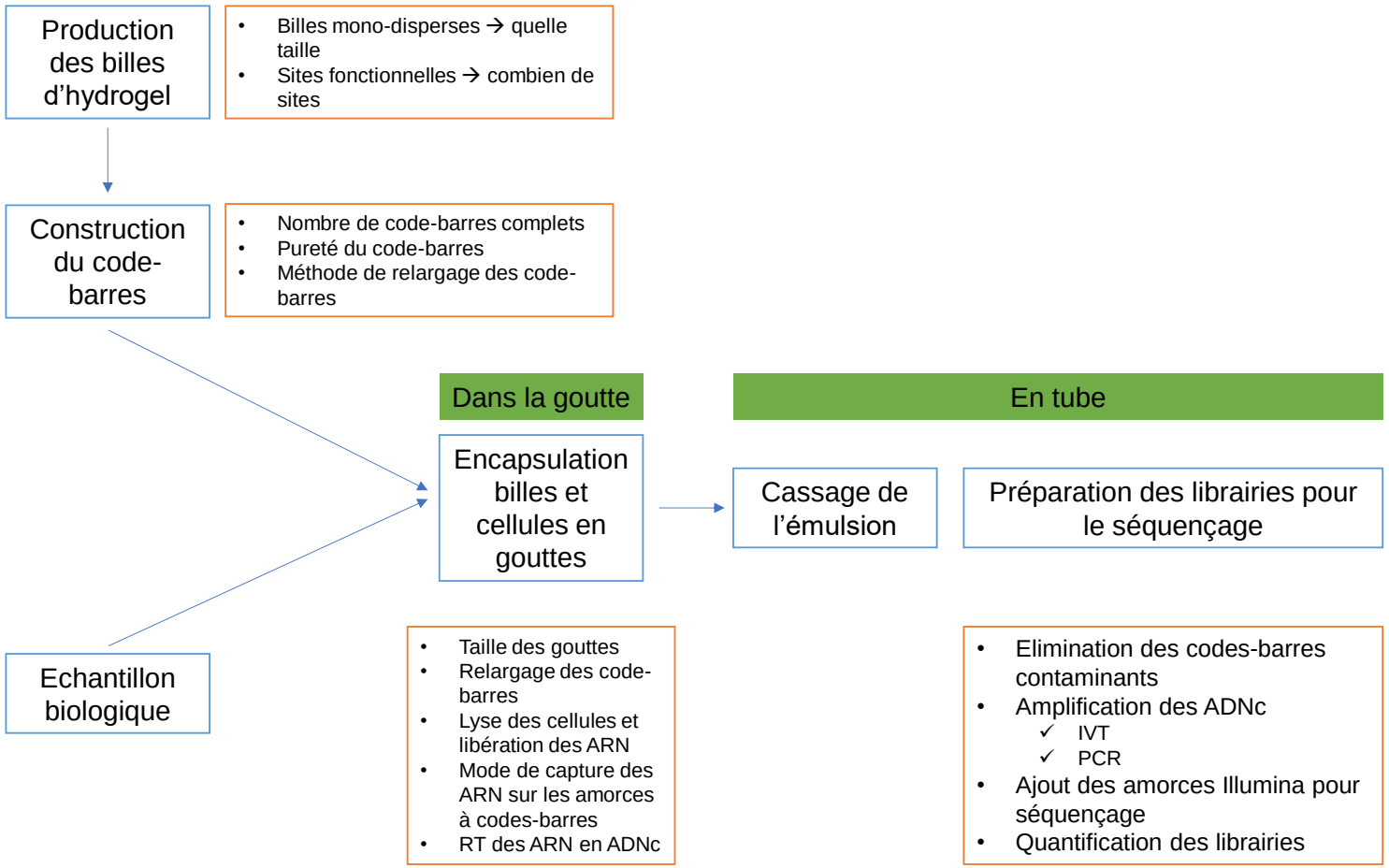


Figure 15 Les différentes étapes clés à optimiser pour la mise au point du séquençage de l'ARN ciblé en gouttes

II. Comment améliorer l'efficacité de la RT ?

A. Lyse des cellules en gouttes

L'influence des débris cellulaires sur l'efficacité de la RT a été testée en réalisant une RT en tubes sur ARN synthétiques en présence ou non de débris cellulaires. Les résultats ont montré que la présence de débris cellulaires n'entraine pas de diminution de l'efficacité réactionnelle. Le tampon de lyse optimale, assurant une RT directe sur cellule efficace, a ensuite été choisi en comparant différents protocoles sur nos cellules d'intérêt, les cellules BV2, une lignée cellulaire de microglie souris ; la lyse est assurée par ajout de Triton X-100 à 0,2% final.

Toutes les étapes réactionnelles, de la lyse des cellules pour libérer les ARN messagers à la synthèse des ADN complémentaires à partir des amorces à code-barres amenées par la bille d'hydrogel, se font au sein d'un même compartiment. On parle de RT directe sur cellules. Les protocoles de RNAseq classiques impliquent une étape préliminaire d'extraction et de purification de l'ARN, ce qui n'est pas applicable dans notre cas. Il va falloir établir le meilleur protocole permettant de libérer efficacement les ARN messagers sans inhiber pour autant la réaction de RT avec les débris cellulaires.

Afin de déterminer l'impact de la présence de débris cellulaires sur l'efficacité de la réaction de RT, une RT en amorces spécifiques a été effectuée sur des ARN synthétiques B et D, en présence de débris cellulaires ou non. Après préparation d'un mix de RT, comprenant enzymes de RT et inhibiteur de RNAses, dNTPs, tampon et l'agent de lyse Triton X-100 à 0,2% ainsi que les ARN synthétiques B et D, celui-ci a été équitablement divisé en 3 tubes. A chacun de ces tubes a été ajouté un tiers de milieu cellulaire avec cellules BV2 de souris, ou de milieu cellulaire dépourvu de cellules, ou simplement d'eau. Dans le cas du tube avec cellules, 200 000 cellules sont ajoutées dans un volume final de 25µL, soit une densité équivalente à une cellule par goutte de 125pL. Les amorces utilisées sont spécifiques des 2 ARN synthétiques B et D utilisés et ne reconnaissent pas le génome ou transcriptome de souris, comme vérifié à l'aide de l'outil Blast. La quantité d'ADNc synthétisés est mesurée par qPCR ; les résultats sont présentés dans la Figure 16.

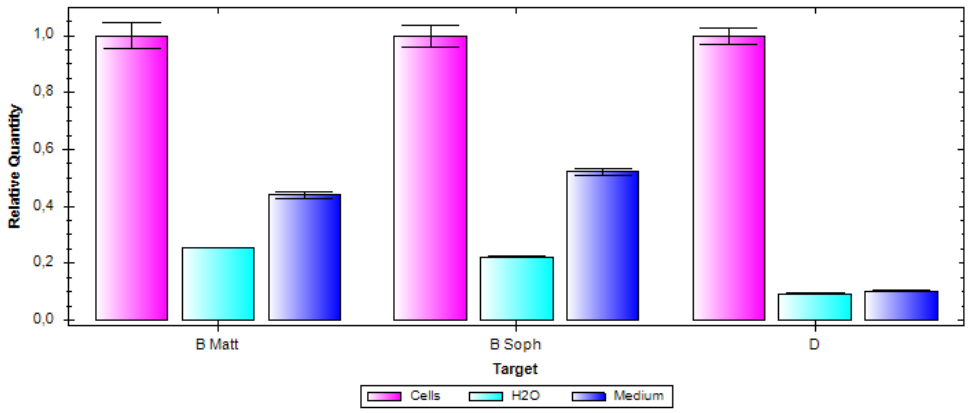


Figure 16 Effet de la présence de débris cellulaire sur l'efficacité de RT en amorces spécifiques sur des ARN synthétiques B et D. Mesure de la quantité relative d'ADNc synthétisés par qPCR à l'aide du logiciel CFX Manager

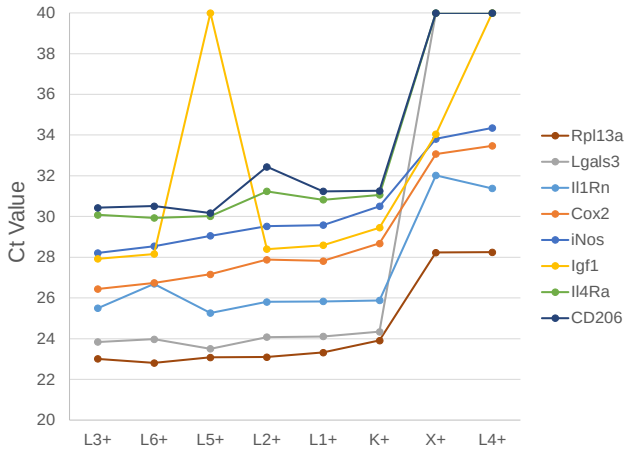
Le graphe présente les quantités relatives en ADNc ; elles ont été calculées par le logiciel de CFX Manager à partir du Ct et de l'efficacité de la réaction d'amplification. Le Ct est le cycle à partir duquel on détecte une fluorescence en qPCR. Plus celui-ci est bas, plus la quantité d'ADN de départ est grande. On remarque que non seulement la présence de débris cellulaire n'inhibe pas la synthèse d'ADNc mais qu'elle semble même en augmenter l'efficacité. On peut faire l'hypothèse que les débris cellulaires ont permis de tapisser le tube, réduisant les phénomènes d'adsorption des molécules sur les parois.

Différents protocoles de lyse cellulaire ont ensuite été comparés. Une réaction de RT directe en amorces spécifiques a été réalisée sur des cellules BV2, en présence de différents détergents, classiquement utilisés pour rompre les membranes cellulaires, ou simplement d'eau (contrôle négatif). Les différents ARN messagers ciblés sont transcrits à des taux différents, allant du gène faiblement exprimé (iNos, CD206) au gène fortement exprimé (Rpl13a). Chaque réaction se fait sur un même nombre de cellules. Après RT, le nombre d'ADNc synthétisés est quantifié par qPCR, afin de déterminer le protocole le plus efficace. Les résultats de qPCR sont présentés dans la Figure 17.

A

	L1	L2	L3	L4	L5	L6	K	X
Réactifs de lyse	Triton 0,2%	Triton 0,02%	Igepal 0,3%	BSA 0,1%	Igepal 0,3% BSA 0,1%	Igepal 0,5%	Protocole Klein Igepal 0,5%	H2O

B



C

Tampon de Lyse L	Delta Cq = Cq (L) – Cq (X)
L1	5,25
L2	5,1
L3	5,99
L4	-0,04
L5	5,13
L6	5,75
K	4,78

Figure 17 Mesure de l'efficacité de RT direct sur cellules en présence de différents agents de lyse; A= Description des réactifs de lyse comparés; B= Résultats de qPCR sur les différents gènes ciblés pour chacun des protocoles testés, une valeur de Ct de 40 a été attribuée aux cas où aucune fluorescence n'a été détecté ; C= Calcul de la différence de Ct moyen pour toutes les cibles entre un protocole de lyse et le contrôle négatif avec H2O

On n'observe que très peu de différences d'un protocole à l'autre et ce pour toutes les cibles testées. Le protocole K, inspiré du protocole inDrop de Klein et al n'est pas plus efficace que les autres, malgré le fait qu'il y ait plus d'enzymes de RT et de dNTPs.

Le protocole de RT directe en amorces spécifiques, avec du Triton X-100 à 0,2% final comme agent de lyse, a donc été validé.

B. Nombre d'amorces spécifiques

A ce stade, nous avons pu constater que la RT directe sur cellules avec amorces spécifiques fonctionne, et que le tampon de lyse utilisé, de même que la présence de débris cellulaire n'ont que peu d'influence sur l'efficacité de la réaction. La Figure 17 montre également qu'augmenter la concentration d'enzyme et de dNTPs n'a pas non plus d'impact. L'efficacité de RT en fonction du nombre d'amorces spécifiques sur des gènes faiblement, modérément ou fortement exprimés, a été déterminée par RT-qPCR sur cellules BV2 en variant la concentration initiale en amorces de RT. Cette expérience a montré une augmentation de l'efficacité de la RT avec l'augmentation de la concentration des amorces, et ce pour les 3 types de cibles, jusqu'à atteindre un plateau. Une concentration optimale doit donc être atteinte de sorte à réaliser une RT à l'efficacité optimale.

La Figure 18 présente les résultats d'une expérience de RT directe sur cellules BV2 en faisant varier la quantité d'amorces de RT, le nombre de cellules, et donc d'ARNm, étant constant d'un tube à l'autre. Après purification sur billes AMPure XP, le nombre d'ADNc générés pour chaque espèce cible est déterminé par qPCR. Les courbes de dissociation sont analysées pour vérifier que la bonne espèce a été amplifiée.

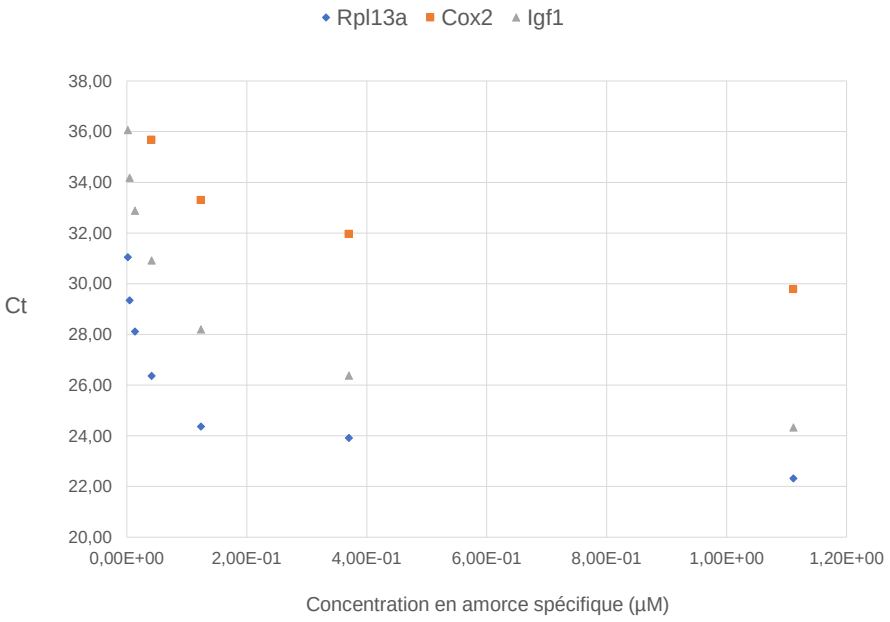


Figure 18 Quantification par qPCR du nombre d'ADNc synthétisés lors de la RT directe sur cellules BV2 en fonction de la concentration en amorces spécifiques de RT

Trois transcrits ont été ciblés par des amorces spécifiques, car ils correspondent à des gènes au niveau d'expression différent ; faible pour Cox2, intermédiaire pour Igf1 et fort pour Rpl13a. On constate que la quantité d'amorces a un réel impact sur la quantité d'ADNc produits. Plus celle-ci est importante, plus il y a de molécules d'ADNc synthétisés, pour un même nombre de départ d'ARN cible, et ce pour les 3 cibles. On atteint un plateau à 1µM d'amorce, seuil à partir duquel l'efficacité de RT n'évolue quasiment plus. Notre objectif va être d'atteindre cet optimum en gouttes de sorte à avoir la réaction la plus efficace possible.

C. Spécificité de la RT

La spécificité de la RT en amorces spécifiques a été testée en procédant à une RT sur 2 types d'ARN synthétiques avec une amorce spécifique de la cible d'intérêt ou non puis en mesurant la quantité d'ADNc produit par qPCR dans ces 2 cas. Les résultats ont montré que la RT fonctionne aussi bien avec une amorce spécifique qu'avec une amorce non spécifique mais est néanmoins beaucoup moins efficace, produisant de l'ordre de 300 fois moins d'ADNc, si l'amorce utilisée n'est pas spécifique de la cible. Ce résultat a été confirmé sur cellules en mesurant la RT sur le gène de ménage Rpl13 à partir de cellules BV2 en présence ou non d'une amorce spécifique de cette cible ; l'ADNc correspondant à ce transcrit n'est détecté en qPCR que dans le cas où une amorce spécifique à cette cible a été ajoutée au mix de RT, témoignant de la spécificité de la réaction de RT en amorces spécifiques.

Les amorces de RT utilisées ont été conçues en utilisant l'outil primer blast de NCBI. Les options ont été personnalisées de sorte à cibler des amplicons hautement spécifiques, d'une taille comprise entre 90 et 160bp, et intégrant une séquence intronique. L'amorce de RT correspond à l'amorce antisens et sa température de fusion (Tm) est de 60°C. La réaction de RT en goutte se fait à 55°C, soit 5°C en dessous de la température de fusion des amorces, assurant, en théorie, une spécificité optimale.

Afin de vérifier la spécificité de la réaction, une première expérience de RT a été réalisée sur des ARN synthétiques polyadénylés, dont les séquences correspondent à 2 amplicons d'intérêt sont données en Annexe (Igf1 et Rpl13a). Chacun de ces ARN synthétiques a été rétro transcrit en présence, soit d'une amorce spécifique de Igf1,

d’une amorce spécifique de Rpl13, ou d’une amorce polyT. Les ADNc synthétisés dans chaque réaction, et pour chacune des cibles, sont ensuite quantifiés par Qpcr. Les résultats sont présentés dans la Figure 19.

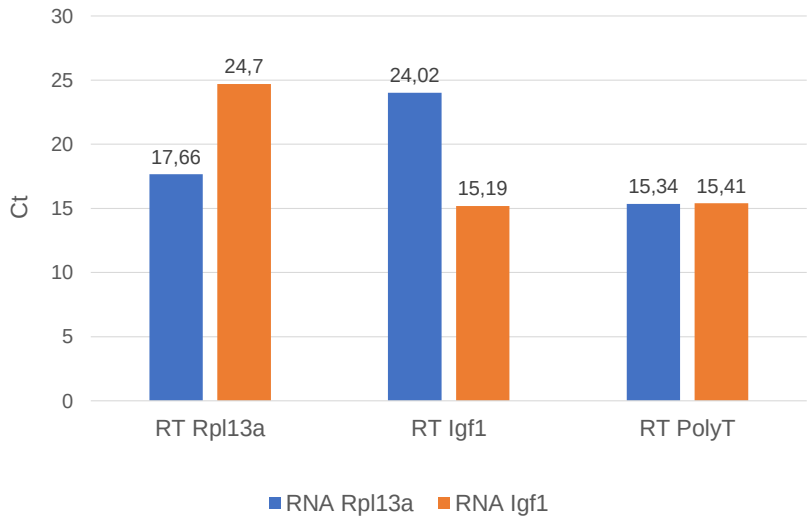


Figure 19 Mesure de la spécificité de la RT en amorces spécifiques sur ARN synthétiques. 6 réactions de RT ont été réalisées en parallèle ; pour chaque cible ARN Rpl13a ou Igf1, une RT en amorce spécifique de Rpl13a (RT Rpl13a), ou spécifique de Igf1 (RT Igf1) ou en amorces polyT (RT PolyT) a été faite. La quantité relative d’ADNc produits est mesurée par qPCR

On constate que la réaction de RT fonctionne dans les 3 cas, à savoir en utilisant une amorce polyT, une amorce spécifique de la cible mais aussi en présence d’une amorce non spécifique de la cible. La réaction en amorces spécifiques est néanmoins beaucoup moins efficace lorsque l’amorce n’est pas spécifique de la cible ; on observe en effet un écart de 7 à 9 Cq entre les 2 cas, soit environ 300 fois moins d’ADNc synthétisés lorsque l’amorce n’est pas spécifique de la cible. La réaction de RT en amorce polyT ou en amorce spécifique pour Igf1 est tout aussi efficace. En revanche, une RT Rpl13 est légèrement moins efficace en utilisant une amorce spécifique par rapport à une amorce polyT, montrant que l’efficacité de la réaction en amorce spécifique semble dépendre de la séquence de l’amorce. La même expérience a été répétée mais en augmentant la température d’incubation de la RT à 60°C, pour voir si cela pourrait augmenter la spécificité, mais aucune différence n’a été observé par

rapport à une RT faite à 55°C, comme recommandé par le manuel d'utilisation de l'enzyme.

Une seconde expérience a été faite afin de tester la spécificité de la RT sur ARN total extrait à partir de la lignée cellulaire BV2. Une RT en amorces spécifique sur plusieurs cibles a été réalisée, en présence ou non d'une amorce spécifique de Rpl13a, un gène de ménage fortement exprimé par les cellules. Les résultats sont présentés dans la Figure 20.

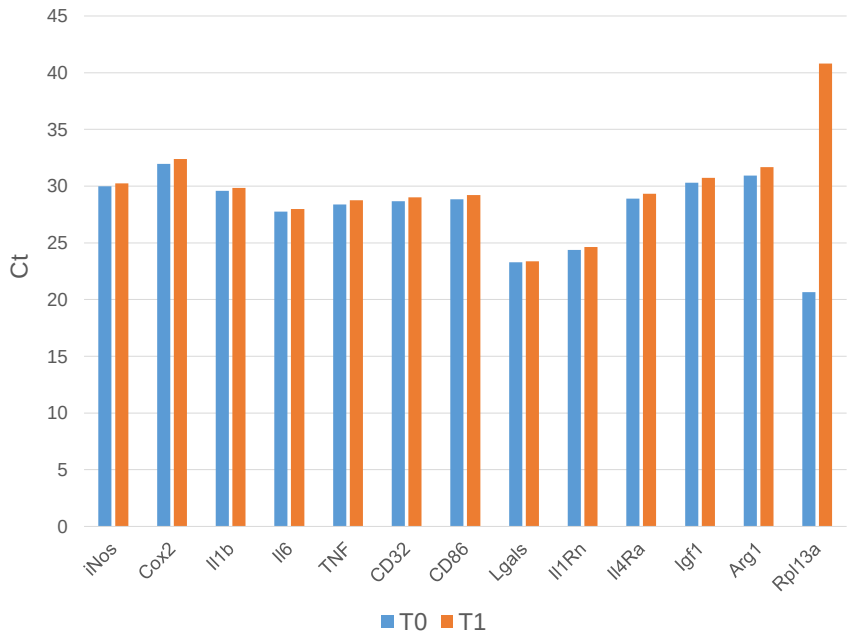


Figure 20 Mesure de la spécificité de la RT en amorces spécifiques sur ARN total extrait à partir de cellules BV2. La réaction a été réalisée en présence ou non d'une amorce spécifique du gène de ménage Rpl13a puis la quantité d'ADNc synthétisé à partir de chaque transcrit est mesurée par qPCR. L'échantillon T0 correspond à la réaction de RT en absence d'amorce Rpl13 et la réaction T1 celle en présence de l'amorce Rpl13

On peut observer que la même quantité d'ADNc a été synthétisée pour tous les transcrits cibles, hormis Rpl13a. Il y a un écart de plus de 10 Cq entre les 2 échantillons, montrant une grande spécificité de la réaction de RT.

D. Impact de la température de fusion des amorces de RT

Le nombre d'amorces a un réel impact sur l'efficacité de la RT. Le fait d'abaisser la température de fusion des amorces de 60° à 50°C devrait faciliter la fixation des transcrits cibles à leur amorce complémentaire à une température de 55°C. Afin de tester cette théorie, de nouvelles amorces plus courtes, à la température de fusion (Tm) abaissée de 5 degrés, ont été conçues pour 4 cibles d'intérêt. Une réaction de RT sur le même matériel ARN à l'aide des amorces original (Tm60) ou à l'aide des amorces au Tm réduit (Tm55) a été effectuée puis le nombre d'ADNc produit a été quantifié par qPCR. Les résultats de qPCR sont présentés dans Figure 21.

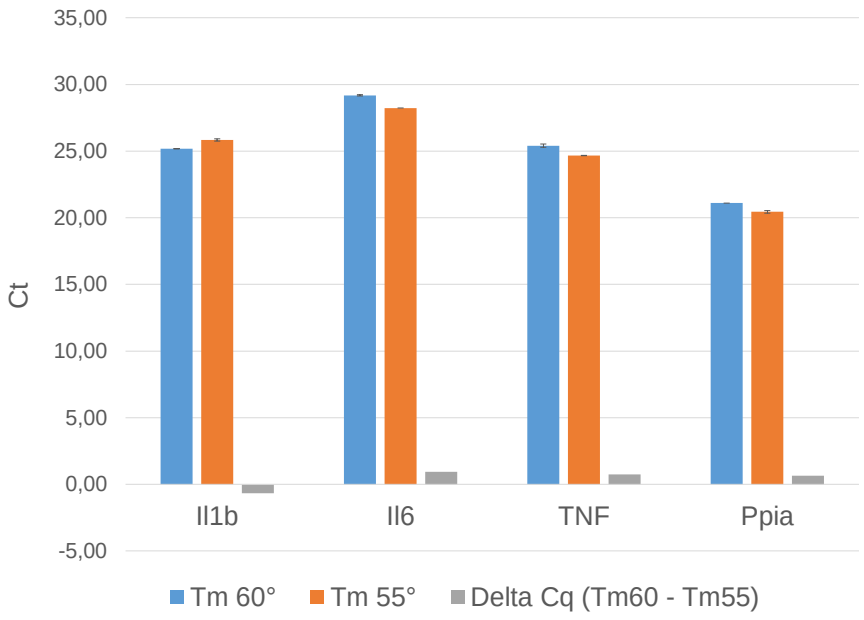


Figure 21 Comparaison de la quantité d'ADNc synthétisés à partir de la même quantité d'ARN de départ, en fonction du Tm de l'amorce de RT utilisée (Tm60 ou Tm55), mesurée en qPCR relative et pour 4 cibles différentes (Il1b, Il6, TNFa et Ppia)

Le nombre d'ADNc synthétisés avec les 2 types d'amorces est sensiblement le même dans les 2 cas, et ce pour les 4 cibles testés. Le changement de Tm n'a donc pas eu d'impact important sur l'efficacité de la réaction. Soucieux de maintenir une bonne spécificité, nous avons donc choisi de garder des amorces de RT avec un Tm de 60°C.

III. Les billes d'hydrogel porteuses des codes-barres cellule unique

Les expériences d'optimisation précédentes ont permis d'identifier un point essentiel dans l'efficacité de la réaction de RT ; le nombre d'amorces spécifiques de RT disponibles. Dans la goutte, cette quantité est directement liée à la capacité des billes d'hydrogel, qui délivrent les amorces spécifiques à code-barres.

Les billes d'hydrogel consistent en un réseau de polymère poreux et déformable. Leurs sites fonctionnels sont des molécules d'ADN double brin possédant une extrémité de 4 paires de bases en 3' et une modification acrydite en 5', via laquelle l'ADN est lié au réseau de manière covalente. Les billes sont produites en générant des gouttes à partir d'une phase aqueuse contenant les différents polymères, l'ADN acrydite et un inducteur de polymérisation. Elles sont ensuite solidifiées par une réaction de polymérisation (UV ou thermique selon la nature chimique des billes).

Un code-barres est alors construit sur ces billes, à partir de l'extrémité ADN, selon une stratégie dite de *split-and-pool* en 3 à 4 étapes (Figure 22). Environ 10 millions de billes sont regroupées. Un premier adaptateur, contenant un site de restriction, une séquence Illumina universelle, et éventuellement un promoteur T7, est ligué spécifiquement à la séquence ADN accroché au réseau, par son extrémité cohésive en 3' de 4 paires de bases. Cet adaptateur comprend également une extrémité cohésive en 3' de 4 paires de bases, mais de séquence différente. Les billes sont alors lavées pour éliminer les adaptateurs non ligués, puis elles sont mélangées à du tampon de ligation et de l'enzyme de ligation avant d'être aléatoirement réparties dans les 96 puits d'une plaque, pré remplie avec un index B. Cet index existe en 96 versions, une par puits. Il est pourvu de 2 extrémités cohésives constantes, complémentaires en 5' à l'extrémité 3' du premier adaptateur et en 3' à l'extrémité cohésive de l'index suivant. Ainsi, les billes déposées dans le puits 1 auront un index B1 tandis que les billes dans le puits voisin auront un index B2. Une fois la réaction de ligation terminée, l'enzyme est inactivée par chauffage et les billes provenant de tous les puits sont regroupées et lavées. Elles sont alors de nouveau réparties de manière aléatoire entre les 96 puits d'une seconde plaque pré remplie avec l'index C, existant en 96 versions également. Le puits C1 pourra contenir aussi bien des billes B1, ou B2, ou encore B96, et ainsi de suite pour les autres puits de la plaque. A l'issu de ces 2 étapes il existe

donc 96x96 combinaisons de séquences, soit 10 000 possibilités. Une bille donnée n'a suivi qu'un seul chemin, par exemple le puits B1 puis C90 et n'a donc qu'une seule combinaison de séquences à sa surface. Chaque extrémité cohésive entre les différentes séquences liguées est différente et spécifique, si bien qu'un seul ordre de ligation est possible. Cette étape de *split-and-pool* peut être renouvelée plusieurs fois de sorte à multiplier le nombre de combinaisons de code-barres possibles. En général, nous effectuons 3 étapes de ligation d'index, soit environ 800 000 combinaisons. Finalement, toutes les billes sont assemblées afin de procéder à la ligation d'une dernière séquence, comprenant une extrémité cohésive en 5', une séquence universelle, un UMI et une amorce spécifique. A cette étape, plusieurs amorces spécifiques sont ajoutées (en concentrations équimolaires), et chaque bille va liguer autant d'amorces différentes qu'il y en a dans le mélange.

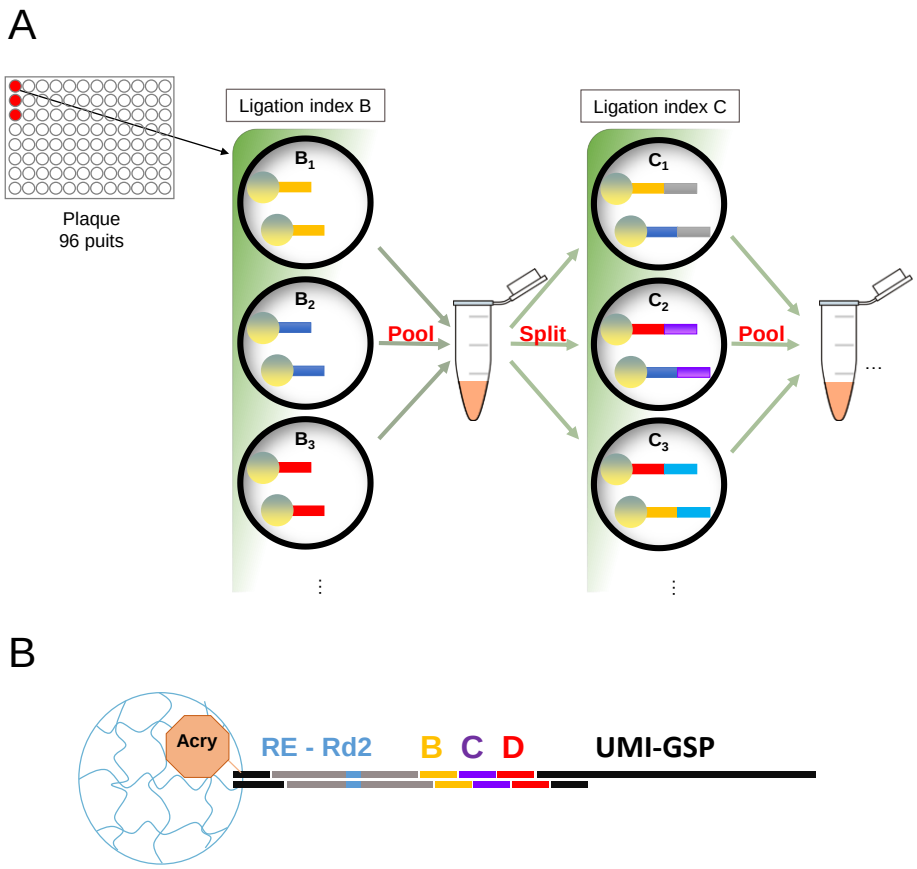
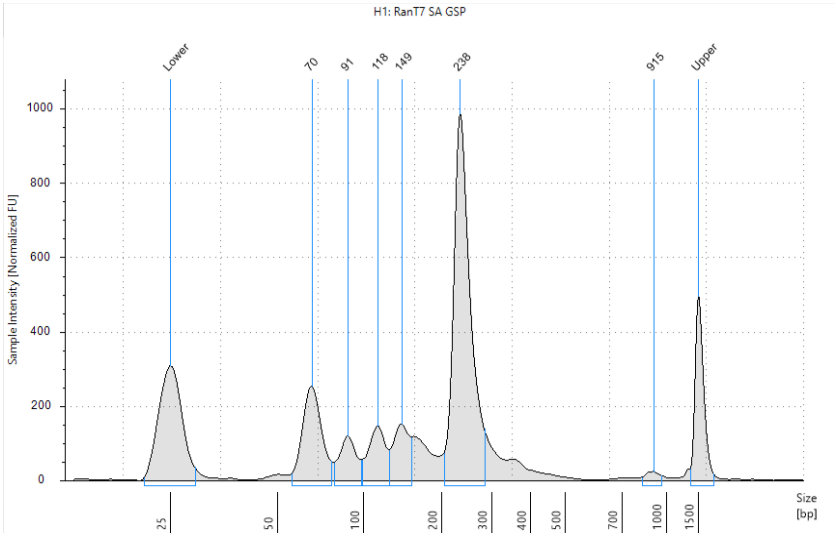


Figure 22 Schéma de la synthèse du code-barres selon une stratégie de ligation en split-and-pool en plusieurs étapes. A= schéma du split-and-pool, B= Structure du code-barres portés par les billes,

Acry = Acrydite, RE = Site de Restriction, GSP = Gene specific primer (pour amorce spécifique)

A. Quantification du nombre de code-barres libérés

Différentes méthodes ont été employées pour quantifier le nombre de molécules de code-barres libérées par chaque bille. La première consiste à digérer quelques dizaines de milliers de billes, quantifier le nombre de molécules d'ADN libérées dans le surnageant en utilisant le kit dsHS DNA Qubit, puis compter le nombre de billes ayant été digérées dans une chambre de comptage. Cette méthode, appliquée à des billes de type PEG DA de 10pL et avec 400 µM de sites fonctionnels par billes, a permis d'évaluer à $1{,}3.10^{8}$ le nombre de molécules relarguées par billes. Etant donné que toutes les molécules d'ADN double brin sont prises en compte par une telle mesure, la proportion de codes-barres complets a été estimée en faisant migrer le produit de digestion des billes sur tape-station. Le profil de migration est présenté dans la Figure 23.



Size [bp]	Conc. [pg/µl]	Peak Molarity [pmol/l]	% Integrated Area
25	421	25900	-
70	279	6130	16.94
91	113	1910	6.84
118	143	1850	8.66
149	145	1510	8.83
238	**950**	**6150**	**57.67**
915	17.6	29.5	1.07
1500	250 (250)	256	-

Figure 23 Migration en électrophorèse capillaire (TapeStation) des produits de digestion d'une bille à la fin de la synthèse du code-barres, et mesure de la proportion relative de chaque pic

Pour les billes quantifiées, 60% de l'ADN ayant migré correspond au code-barres complet ; soit une estimation de $0{,}8.10^8$ codes-barres complets par bille.

Une quantification par qPCR de ces mêmes billes, en utilisant comme étalon un oligonucléotide de taille et structure similaire aux amorces à code-barres, et flanqué des mêmes séquences, a permis d'évaluer le nombre de codes-barres complets par billes à 1 à $3x10^7$ par bille. Dans un premier temps, les billes à code-barres complets sont triées par FACS dans une plaque 96 puits préremplie avec l'enzyme de restriction et son tampon. Après digestion, le surnageant est dilué et le nombre de molécules dans chaque puits est quantifié par qPCR. Cette méthode est bien plus précise, permettant de ne quantifier que les codes-barres complets, possédant les 2 sites complémentaires aux amorces de qPCR et également de mesurer les différences entre billes.

Les 2 méthodes donnent des résultats légèrement différents mais cohérents. L'optimum à atteindre, d'après les différentes expériences décrites précédemment, est de 1µM par amorce cible, soit 10^8 molécules dans une goutte de 100 pL pour chaque amorce. Nos billes d'hydrogel sont produites à 9 pL avec 400 µM de sites fonctionnels. Le nombre de sites théoriques est donc de 2.10^9 sites, de quoi ajouter 20 amorces différentes par billes. Nous sommes bien en dessous de ce chiffre. La chimie des billes, la construction du code-barres ainsi que le relargage des amorces dans la goutte constituent tous des points d'amélioration potentiels, dans le but d'amener dans la goutte suffisamment d'amorces pour assurer une réaction de RT efficace.

B. Chimie des billes

Les premières billes d'hydrogel produites avaient une chimie PEG DA et acrydite ADN double brin (baptisées billes PEG DA) et l'étape de polymérisation se faisait sous l'action des UV, comme décrit dans le chapitre 2. Ces billes offraient plusieurs avantages, des pores larges permettant un maximum de sites accessibles dans tout le réseau, une simplicité de production et une polymérisation en direct, rapide et efficace. Nous avons cependant dû faire face à 2 problèmes majeurs qui nous ont contraints à changer de chimie de billes. L'un des principaux composants de ces billes PEG DA est le PEG DA 6000, et le nouveau lot fabriqué par notre fournisseur était légèrement différent, ne permettant pas de produire des billes polymérisées stables. De plus, lorsque nous avons voulu produire des billes PEG DA de 100pL au lieu des

10pL habituels, nous avons dû faire face à de nombreux problèmes lors de la production. La polymérisation se fait en continu, en sortie de puces de production. Des billes de 100pL, de grande taille, avaient tendance à boucher rapidement le tube de polymérisation.

Nous nous sommes tournés vers la même chimie que les billes d'hydrogel utilisées par la technologie inDrop et dont le protocole de production a été publié. Il s'agit d'une chimie en polyacrylamide et acrydite ADN double brin (baptisées billes PA), avec une polymérisation qui se fait en post production, par ajout de TEMED et d'APS et incubation à 65°. Ces billes offrent l'avantage d'une polymérisation plus simple, d'une plus grande flexibilité dans la taille de production des billes, la phase aqueuse hydrogel étant moins visqueuse que celle des billes PEG DA et la polymérisation se faisant en post production, évitant tout risque de créer de la résistance en sortie de puce de production.

Afin de vérifier que les nouvelles billes en chimie polyacrylamide sont comparables aux billes PEG DA, j'ai effectué des contrôles qualité sur ces 2 types de billes. Les billes PEG DA utilisées pour cette comparaison ont été produites avec l'ancien lot conforme de PEG DA 7000. La Figure 24 présente les résultats de ce contrôle qualité.

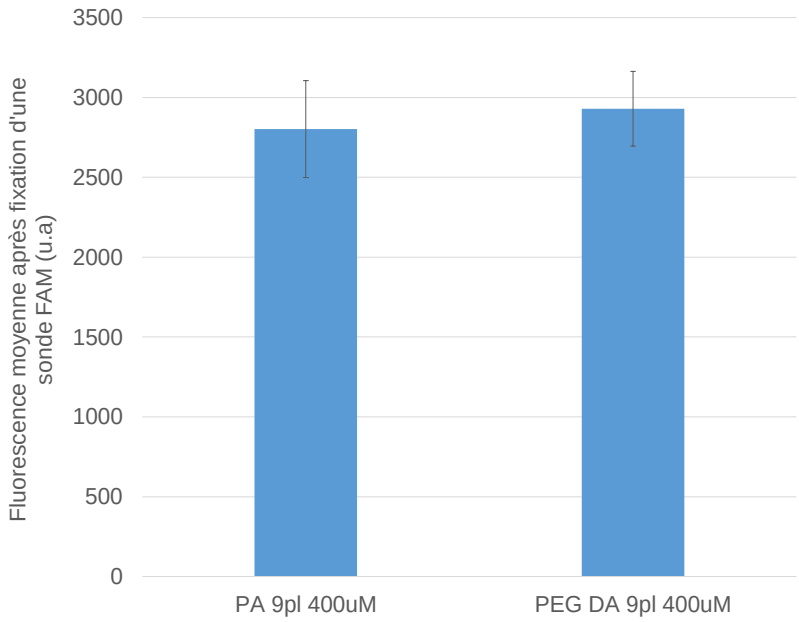


Figure 24 Analyse au microscope à épifluorescence de billes PEG-DA ou PA de 10pL, fonctionnalisée avec 400µM d'Acrydite-DNA et auxquelles a été ligué un premier adaptateur. Les billes sont marquées par une sonde fluorescente FAM spécifique du premier adaptateur

On constate que la capacité fonctionnelle des 2 billes est similaire. De plus, l'encapsulation en goutte de ces billes a été testée pour s'assurer qu'elles possédaient les mêmes propriétés de déformation, permettant de battre la loi de Poisson et d'obtenir un taux d'encapsulation d'une bille par goutte de l'ordre de 80%. Les résultats, non présentés ici, ont montré un taux d'encapsulation supérieur à 80%.des billes PA.

C. Construction du code-barres par ligation

Le code-barres se construit en plusieurs étapes de ligation successives. Nous devons nous assurer d'avoir une efficacité de réaction optimale de sorte à avoir un maximum de code-barres complets par goutte, et ainsi une rétrotranscription efficace elle aussi.

D'après les premiers protocoles, une étape de ligation consistait à ajouter à 250µL de culot de billes d'hydrogel fonctionnalisées à 400µM, 60kU de T7 DNA Ligase, du tampon de T7 DNA Ligase à 1x final et 480µL d'index à liguer à 50µM.

On constate dans ce protocole qu'il y a 4 fois plus de sites théoriques que d'index à liguer. Nous avons donc dans un premier temps voulu mesurer l'efficacité de ligation en fonction du rapport O/B (pour nombre d'oligonucléotides à liguer/ nombre de sites théoriques apportés par les billes d'hydrogel).

Des quantités variables d'index ont donc été liguées à différents types de billes d'hydrogel PA. Les billes sont fonctionnalisées avec de l'ADN double brin, ds-AcRanA, possédant une modification Acrydite à l'extrémité 5' du brin supérieur, une séquence RanA sur laquelle peut se fixer une sonde fluorescente FAM-RanAas, et une extension de 4 paires de bases en 5' du brin inférieur ; ds-AcRanA. L'index à liguer, dsBclRd2iBb, quant à lui, consiste en une séquence double brin avec modification phosphate en 5' du brin supérieur, une extension de 4 paires de bases en 5'complémentaires à celle du brin RanA ainsi qu'une séquence Rd2, que vient fixer une sonde fluorescente FAM-Rd2as. Cette expérience a été réalisée en parallèle sur 2 productions de billes 10pL ou 100pL avec 400µM d'Ac-RanA, soit 4 billes différentes. Les rapports O/B testés sont les suivants ;

- O/B = 0,25 (premier protocole)
- O/B = 1,5
- O/B = 3
- O/B = 6

Les résultats de cette expérience sont donnés dans la Figure 25

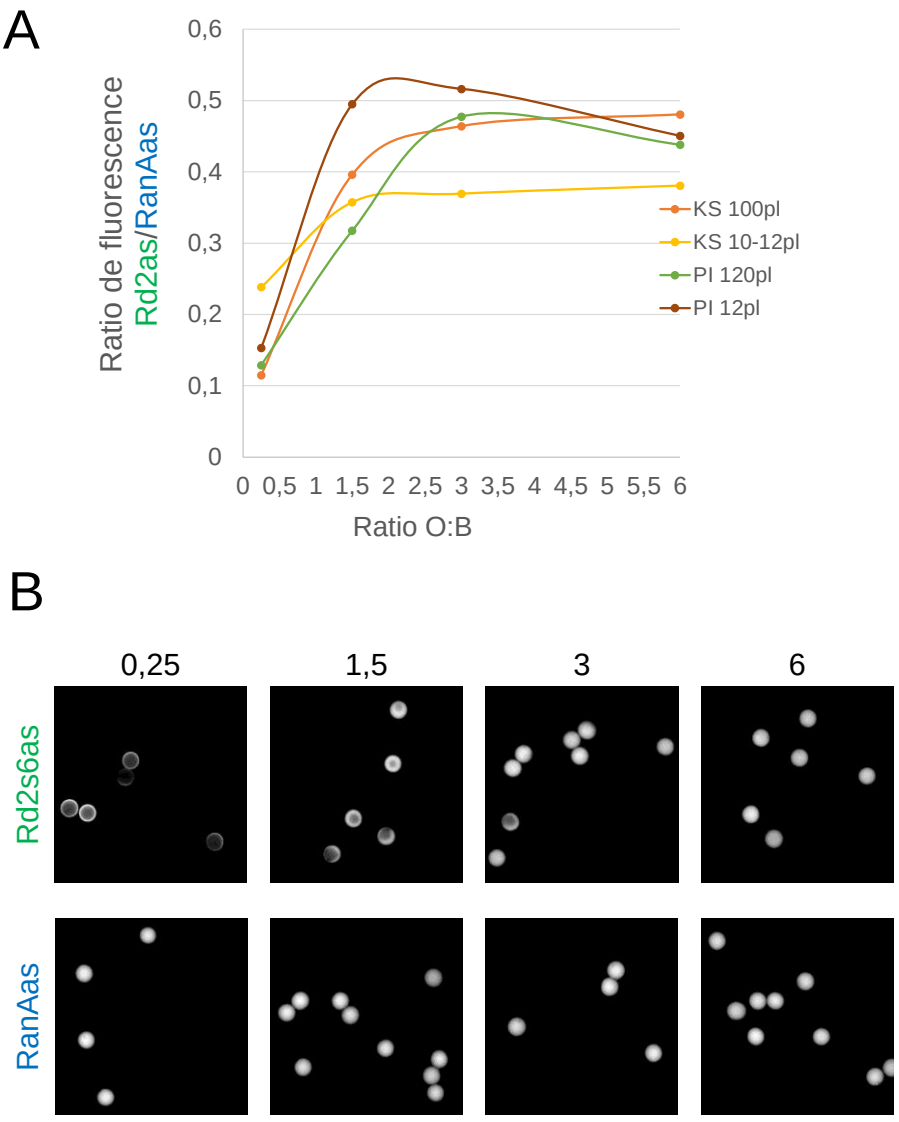


Figure 25 Influence du ratio O/B (nombre d'oligonucléotides à liguer/ nombre de sites théoriques apportés par les billes d'hydrogel) sur l'efficacité de ligation ; A= Ratio des fluorescence moyenne des sondes RanAas/Rd2as sur les différentes billes, dans les différentes conditions de ligation ; B= Image prise au microscope à Epifluorescence en FITC des billes PA 100pL (KS), fixées avec une sonde FAM RanAas ou Rd2as et pour les différentes conditions de ligation

On observe que le ratio de fluorescence Rd2as/RanAas augmente avec le rapport O/B jusqu'à atteindre un plateau à partir du rapport O/B =3, et ce pour tous les types de billes. L'efficacité de ligation n'est que de 10% dans les conditions du premier protocole contre 50% pour un protocole de ligation avec un excès par 3 d'index à liguer. Cette amélioration peut être également facilement constatée sur les images prises au microscope à épifluorescence (Figure 25 B), où l'on observe une fluorescence non homogène au sein d'une bille ainsi que d'une bille à l'autre pour les rapports O/B inférieurs à 3.

Nous avons ensuite voulu comparer différentes enzymes de ligation, différentes quantités d'enzyme, différents temps et température d'incubation ainsi que différents tampons de ligation. Je vais ici me contenter de présenter l'expérience ayant permis de valider la quantité d'enzyme à utiliser ainsi que le meilleur tampon de ligation. La Figure 26 présente les résultats de ce test.

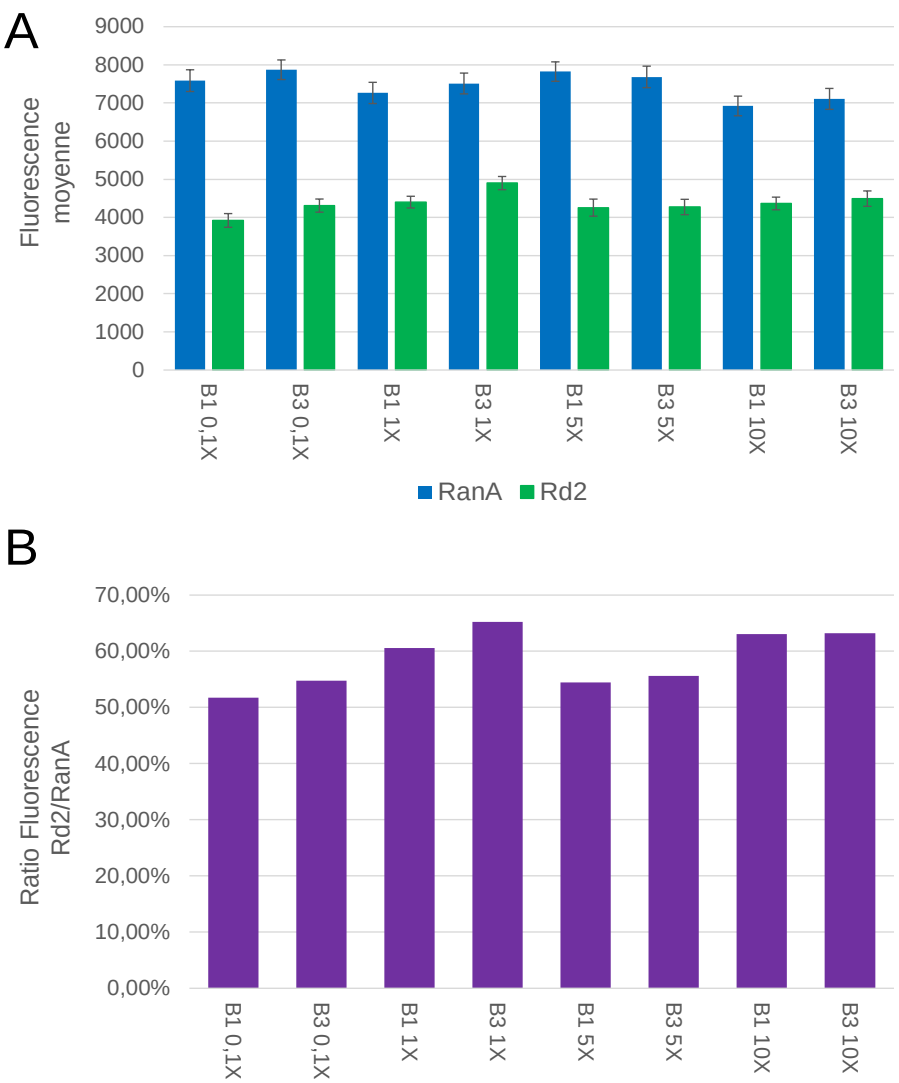


Figure 26 Comparaison de l'efficacité de ligation en fonction de la quantité d'enzyme T7 et du buffer de ligation. A= Fluorescence moyenne mesurée pour la sonde RanAas ou la sonde Rd2as, B= ration des fluorescence moyenne Rd2as/RanAas. B1 = Tampon T7 DNA Ligase ; B3 = Tampon NEB Quick Ligation Buffer ; 0,1X = 3U/µL d'enzyme T7 ; 1X = 30U/µL d'enzyme T7 ; 5X = 150U/µL d'enzyme T7 ; 10X = 300U/µL d'enzyme T7

D'après le graphe B, on n'observe que peu de variations sur l'efficacité de ligation dans les différentes conditions testées.

Le nouveau protocole de ligation à utiliser, en fonction des résultats de ces expérience est donc le suivant ; Ajouter à 250µL de culot de billes d'hydrogel fonctionnalisées à 400µM, 60kU de T7 DNA Ligase, du tampon Quick Ligation Buffer à 1x final et 600µL d'index à liguer à 500µM.

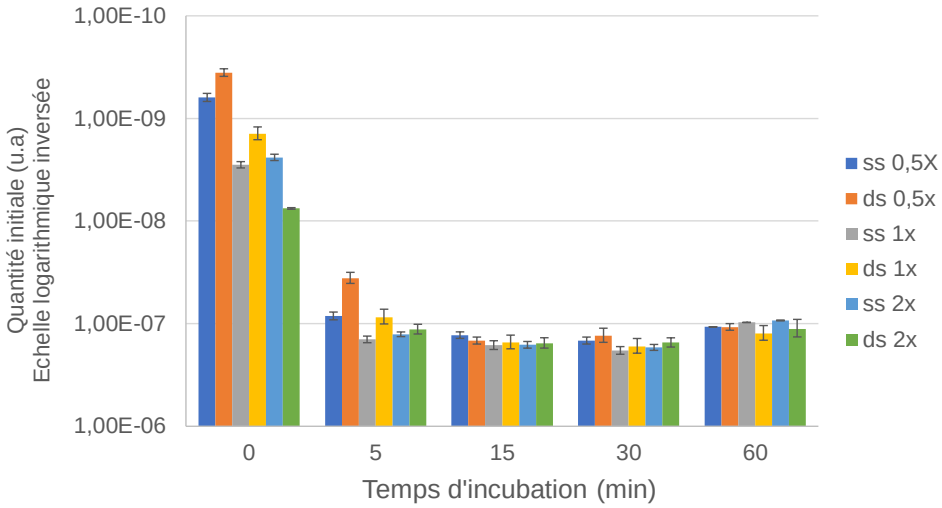
D. Relargage du code-barres en gouttes

S'il est essentiel d'optimiser la construction du code-barres afin d'augmenter les capacités de nos billes, il est tout aussi important de faire en sorte que ces code-barres soient efficacement libérés dans la goutte, pour aller rencontrer leur ARN cible.

Plusieurs méthodes de relargage sont possibles. Nous avons choisi de comparer 2 d'entre elles ; un relargage enzymatique en utilisant une enzyme de restriction ou par traitement UV grâce à la présence d'un domaine photo clivable. Pour ce qui est du traitement enzymatique, l'enzyme choisie est BclI, car elle fonctionne à température compatible avec la réaction de RT, est compatible avec les tampons de RT et a peu de sites de coupure dans le transcriptome et d'ailleurs, il n'y en a pas au niveau des cibles d'intérêt choisies dans le cadre du projet microglie, décrit dans le chapitre suivant.

La Figure 27 présente les résultats des expériences menées afin de comparer les deux méthodes de relargage, de déterminer la cinétique réactionnelle dans le cas de la méthode enzymatique et enfin de vérifier si les paramètres choisis sont optimaux.

A



B

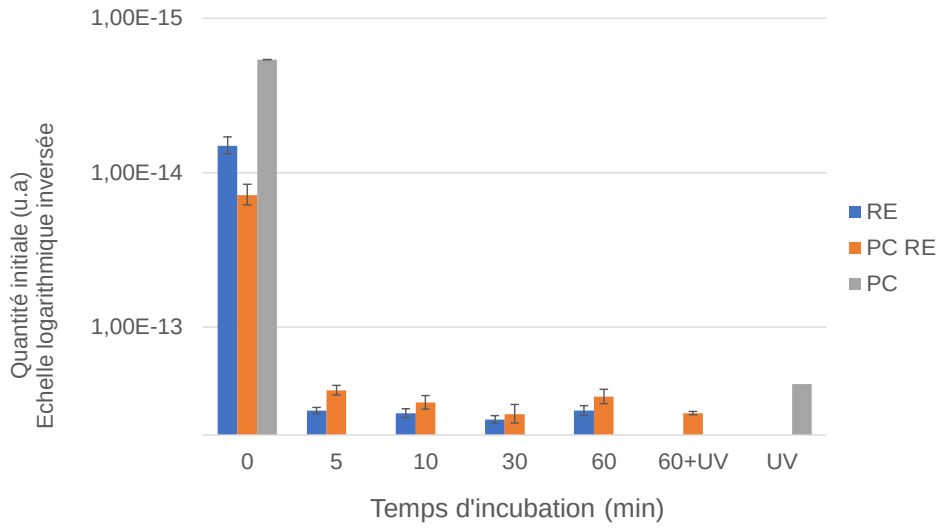


Figure 27 Quantification par qPCR du nombre de code-barres complets relargués des billes d'hydrogel par traitement UV ou enzymatique ; A= Cinétique de coupure par l'enzyme BclI à différentes concentration sur code-barres double brin (ds) ou partiellement simple brin (ss), 0,5x=0,2U/µL, 1x = 0,4U/µL, 2x = 0,6U/µL final ; B= Cinétique de relargage du code-barres par traitement avec BclI et/ou UV, sur des billes possédant un site de restriction (RE), ou un site de restriction + un domaine photo clivable (PC RE), ou un domaine photo clivable uniquement (PC)

Le premier graphe montre les résultats d'une expérience sur billes possédant un site de restriction, incubées dans le milieu de RT, à 55°C, en présence de différentes quantités d'enzyme de restriction BclI (0,5x, 1x ou 2X avec 1x=0,43U/µL finale) Des aliquotes ont été prélevés à différents temps et la quantité de code-barres relargués a été quantifiée par qPCR. Un autre paramètre a été testé ; le code-barres doit subir un traitement dénaturant en fin de construction afin de le rendre simple brin et de pouvoir ainsi éliminer les codes-barres non utilisés par traitement ExoI à l'issu de la RT. Le site de restriction situé à l'avant du code-barres devant être double brin pour qu'il y ait coupure, seule cette partie est rendue double brin à l'issu de l'étape de dénaturation. Les billes « ds » sont donc des billes n'ayant subi aucun traitement post-production du code-barres tandis que les billes « ss » ont été dénaturées, puis réhybridées au niveau du site de restriction. On constate ici que le relargage est effectif dès 5 minutes d'incubation. Il est même maximal pour une quantité 1x (et 2x) d'enzyme BclI. Considérant une quantité 1x d'enzyme de restriction et 5 minutes d'incubation, on ne constate pas de différence significative dans le nombre de code-barres libérés par des billes « ss » ou « ds ». Le traitement de dénaturation puis ré-hybridation ne diminue donc pas la capacité de relargage par BclI des codes-barres.

La seconde expérience a pour but de comparer les 2 méthodes de relargage, UV ou enzymatique. La technologie inDrop fait appel à la première méthode, mais cela implique de grandes précautions lors de la manipulation des billes qui ne doivent pas être exposées aux UV avant encapsulation en gouttes avec les cellules. Trois types de billes ont été ici comparées ; des billes possédant un site de restriction BclI (RE), des billes possédant un domaine photo clivable (PC), et des billes possédant les 2 (PC RE). La cinétique de relargage par BclI a de nouveau été faite. Là encore, des aliquotes ont été extraites à différents temps. Les codes-barres relargués dans le surnageant ont été récupérés et quantifiés par qPCR. Cette expérience permet de confirmer les résultats précédents quant à la cinétique enzymatique. De plus, on constate que les 2 méthodes de relargage sont tout aussi efficaces. Le traitement UV après 60min de traitement enzymatique ne permet d'ailleurs pas de libérer plus de code-barres.

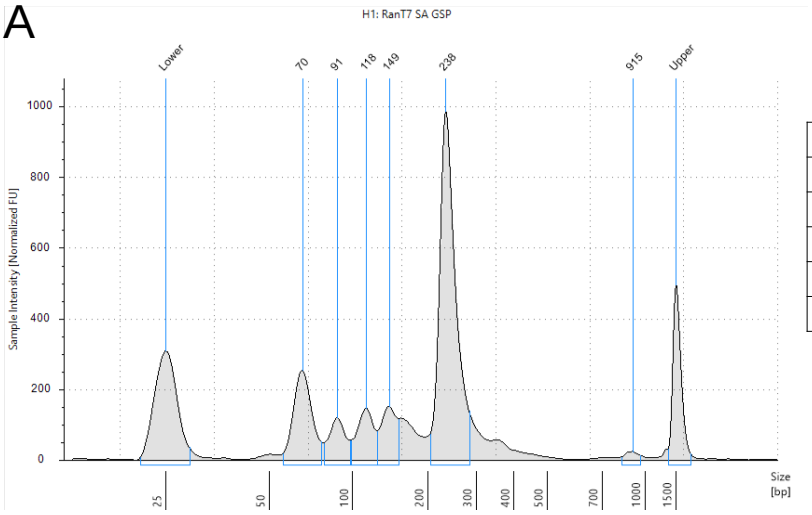
Pour des questions évidentes de facilité de manipulation, et pour limiter la diffusion aléatoire de codes-barres libres, qui peuvent être libérés par exposition à la lumière

avant encapsulation, j'ai choisi la méthode de relargage enzymatique et validé l'enzyme BclI à une concentration finale de 0,43U/µL.

E. Séquences des oligonucléotides constituant le code-barres

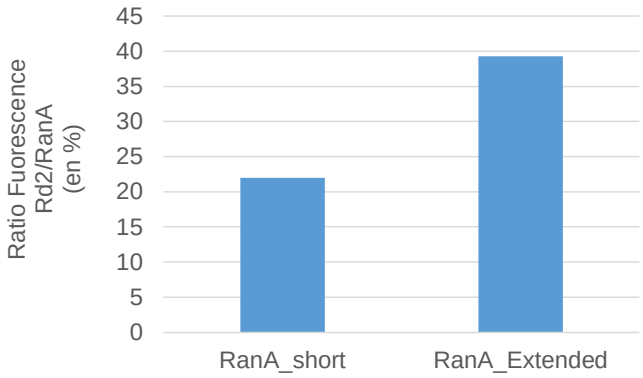
En s'intéressant à la taille des produits relargués par une bille d'hydrogel après digestion enzymatique, (Figure 28, A), on remarque la présence de codes-barres incomplets, correspondant à des réactions qui ne sont pas efficaces à 100% à chaque étape de ligation. Ceux-ci ne représentent que 5 à 10% des produits totaux pour ce qui est des étapes de ligation des index C et D, soit 90 à 95% d'efficacité. En revanche, les produits incomplets n'ayant pas réussi à liguer le premier index représentent 15% du total, soit une efficacité de ligation de 75% de l'index B sur le premier adaptateur. On ne peut pas quantifier, d'après ce résultat, le taux de ligation du premier adaptateur sur l'ADN acrydite, le site de restriction étant situé au niveau du premier adaptateur. Si on se fie aux mesures en microscopie de l'efficacité de ligation par ajout de sondes fluorescentes spécifiques présentées dans les parties précédentes, on estime l'efficacité de cette première étape de ligation à 50% environ. Il semblerait donc que les premières étapes de ligation soient les plus critiques. Cela pourrait être dû à un manque d'accessibilité des sites à liguer lorsque le code-barres est encore de petite taille et donc enfoui dans le réseau d'hydrogel. Nous avons donc essayé d'augmenter la taille de l'ADN acrydite lié de manière covalente au réseau d'hydrogel de sorte à le rendre plus accessible pour les premières étapes de ligation.

A

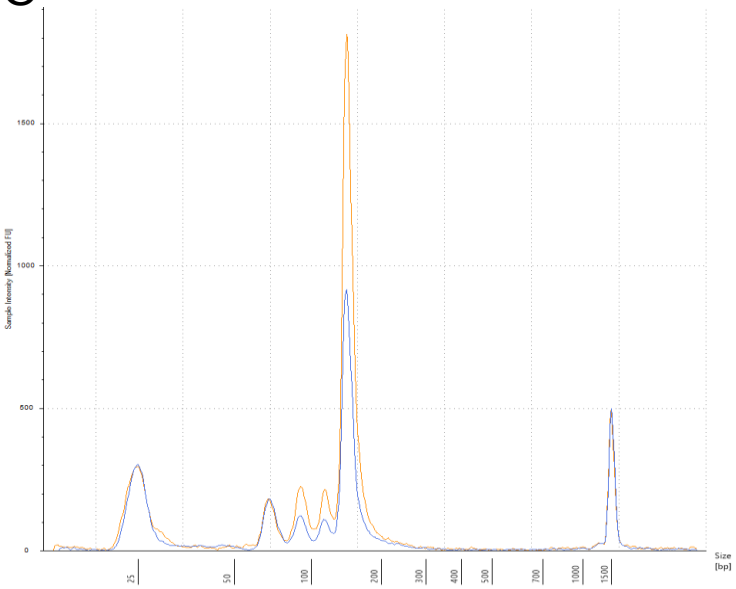


pic (bp)	
70	16,90%
91	6,80%
118	8,60%
149	8,80%
238	57,70%

B



C



pic (bp)	S1	S2
73	16,3%	8,9%
93	10,5%	10,7%
118	8,3%	8,5%
151	64,9%	71,9%

S2

S1

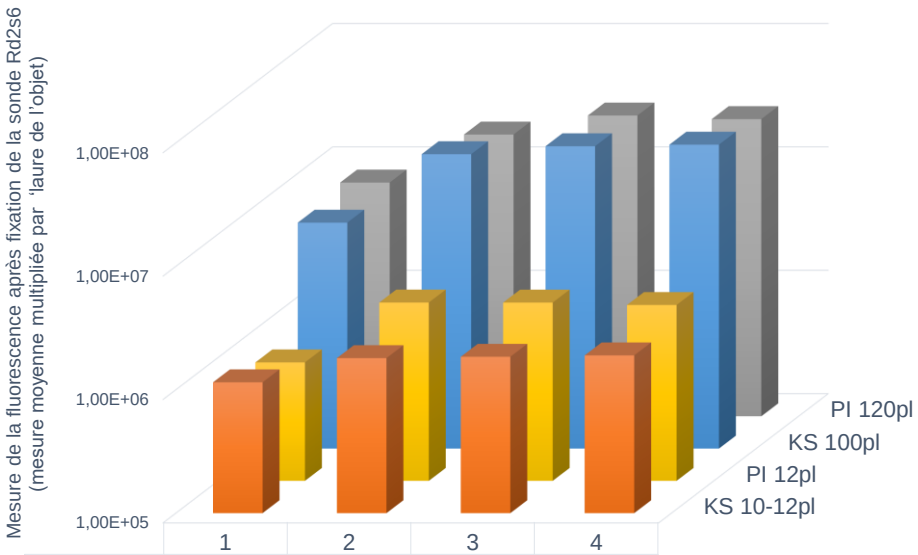
Figure 28 Impact de la longueur de l'oligonucléotide à liguer sur l'efficacité de réaction, A= migration des produits de digestion d'une bille avec code-barres en électrophorèse capillaire et mesure de la proportion relative de l'aire de chaque pic, B= Mesure du l'efficacité de ligation du premier adaptateur sur des billes produites avec une séquence Acrydite ADN courte de 15bp (RanA_short) ou longue de 45bp (RanA_Extended), C= Comparaison des produits de digestion après construction d'un code-barres, en liguant le premier adaptateur en 2 étapes, avec 2 séquences de 40bp (bille S2), ou en un seule étape, avec un adaptateur de 80bp (bille S1)

Les résultats sont présentés dans la Figure 28, graphe B. Il semblerait donc que l'utilisation d'une séquence de départ plus longue facilite la première ligation.

On peut également tenter de faciliter la rencontre entre les sites de ligation sur bille et les séquences à liguer en réduisant la taille de ces dernières. Nous avons comparé l'efficacité de ligation d'une séquence de 80 paires de base en une seule étape ou en 2 étapes. La ligation en 2 étapes correspond à 2 ligations successives de 2 séquences de 40 paires de bases, correspondant aux deux moitiés de la séquence de 80 paires de bases. On observe, d'après la Figure 28 C, qu'une stratégie de ligation en 2 temps de 2 séquences de moindre taille est légèrement plus efficace qu'une stratégie de ligation d'une longue séquence en une seule étape, avec respectivement 72 et 65% de code-barres complets. Le fait d'ajouter une étape de ligation supplémentaire est compensée par le fait que la ligation d'une séquence plus courte est plus efficace.

F. Taille des billes

En s'intéressant de plus près à l'expérience décrite dans la partie 0, on peut réinterpréter les données et comparer les billes 10pL et 100pL, dans les conditions de ligation optimales (ratio O/B = 3).



	1	2	3	4
KS 10-12pl	1,15E+06	1,80E+06	1,85E+06	1,90E+06
PI 12pl	9,11E+05	2,79E+06	2,78E+06	2,66E+06
KS 100pl	6,77E+06	2,43E+07	2,82E+07	2,90E+07
PI 120pl	7,83E+06	1,92E+07	2,74E+07	2,56E+07

Figure 29 Mesure de la fluorescence de billes de 10pL (2 réplicas, KS 10pL et PI 12pL) ou 100pL (2 réplicas KS 100pL et PI 120pL) après fixation d'une sonde spécifique Rd2s6. La fluorescence moyenne de chaque objet est multipliée par l'aire de l'objet

D’après la Figure 29, on observe qu’en augmentant le volume de la bille par 10, on augmente également le nombre de code-barres ligués, avec une même efficacité de ligation dans les 2 cas dans des conditions optimales. Un moyen d’apporter suffisamment d’amorces avec code-barres par goutte et donc tout simplement d’augmenter la taille de la bille. Cela aura quelques contreparties, et notamment la nécessité d’encapsuler ces billes dans des gouttes bien plus grande, diminuant la fréquence de production de gouttes avec billes et cellules. La comparaison de l’efficacité de la RT ciblée avec billes de 10pL encapsulées en gouttes de 100pL versus billes de 100pL encapsulées dans des gouttes de 1nL sera présentée dans le chapitre suivant.

G. Pureté du code-barres

Nous avons vérifié que chaque bille d'hydrogel était porteuse d'un code-barres unique. Nous avons donc séquencé quelques centaines de billes et avons quantifié les différents code-barres identifiés pour chacune d'elle (Ameta & Arsène *et al.*, en préparation, Annexe 2).

Pour cela, des billes à code-barres porteuses d'une amorce spécifique ont été triées à l'aide d'un FACS dans une plaque 96 puits. Des ARN synthétiques, reconnus par l'amorce spécifique sur les billes, sont ajoutés, avec les réactifs et l'enzyme nécessaires pour libérer les codes-barres et effectuer une réaction de RT. Les ARN possèdent une séquence variable reportrice, qui existe en 8 versions différentes, permettant de coder chaque ligne de la plaque. A la fin de la réaction, 2µL de chaque puits sont transférés dans une nouvelle plaque pour amplifier les ADNc synthétisés par PCR. L'amorce sens utilisée est porteuse d'une séquence reportrice de colonne, existant en 12 versions. A l'issu de ces 2 réactions successives, les ADNc provenant de chaque puits sont porteurs des codes-barres initialement présents dans le puits et également d'un code de ligne et de colonne. A l'issu du séquençage, chaque bille est identifiée grâce aux codes de ligne et de colonne et le pourcentage du premier et du second code-barres les plus abondants dans chaque puits sont calculés.

Les résultats sont présentés dans la Figure 30.

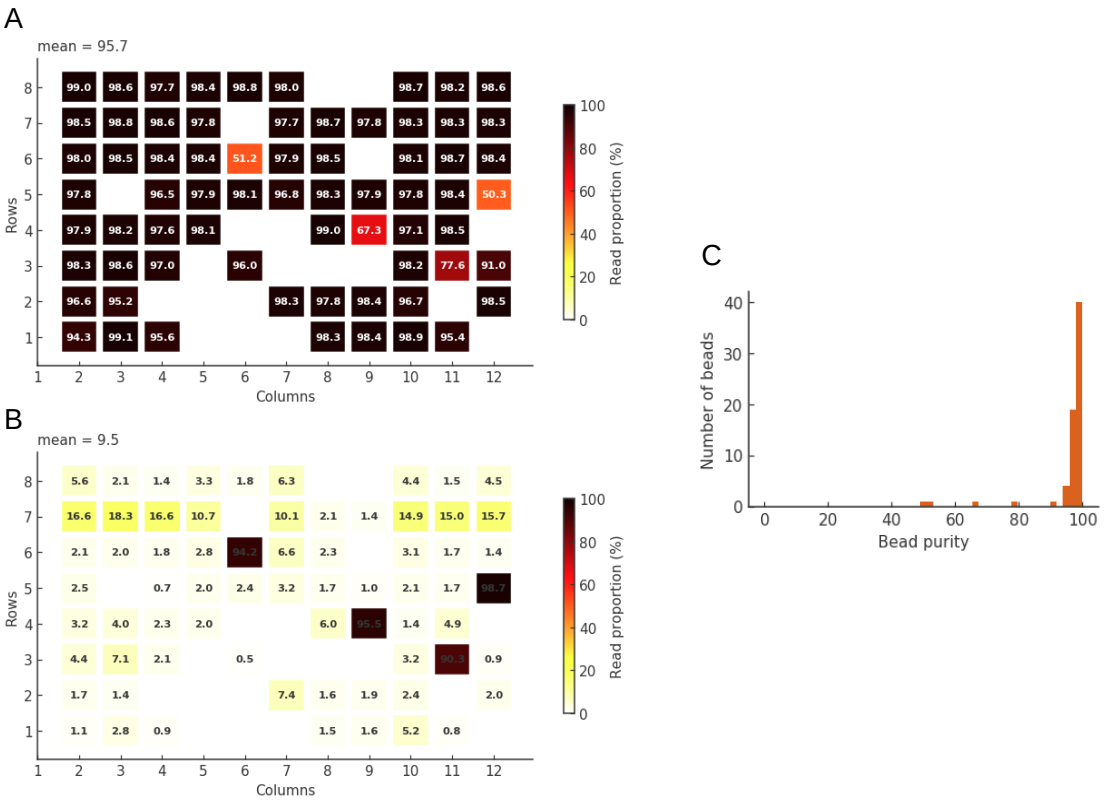


Figure 30 Mesure de la pureté du code-barres par séquençage de billes à code-barres complets, ayant été triées par FACS en plaque 96 puits, avant d'ajouter un code de puits aux amorces à code-barres protées par chaque bille, A= proportion du code-barres le plus abondant dans chaque puits où une bille a été triée, B= proportion du second code-barres le plus abondant, C= histogramme représentant la proportion de billes avec un certain degré de pureté (défini par le pourcentage du plus abondant code-barres)

Le pourcentage moyen de la proportion du premier code-barres le plus abondant est de 95%. Seules 3 des 71 billes analysées montrent une valeur plus faible, de l'ordre de 50%. Le second code-barres le plus abondants dans ces 3 puits représentent 50% environ des codes-barres présents, signifiant que ces puits contenaient vraisemblablement 2 billes.

Les billes sont donc bien porteuses d'un code-barres unique avec une pureté de plus de 95%.

IV. Amplification des ADNc et préparation des librairies de séquençage

Les technologies RNAseq cellules uniques emploient principalement deux stratégies pour l'amplification des ADNc : la PCR ou la transcription *in vitro*. La première permet une amplification efficace et fidèle, grâce à l'utilisation d'enzyme de PCR *high fidelity* (« haute fidélité » : de l'ordre de moins de 1 substitution par 10^6 nucléotides polymérisés) à partir de très faible quantité de matériel. En revanche, cette amplification est exponentielle et va avoir une efficacité variable en fonction de la taille de la séquence amplifiée ou encore de sa teneur en GC, générant des biais d'amplification. Ceux-ci peuvent néanmoins être corrigés en utilisant des UMI, qui renvoient directement au nombre de transcrits initiaux. Une PCR fait appel à deux amorces spécifiques et les technologies de séquençage de l'ARN total doivent donc inclure une étape permettant d'ajouter un site universel en 3' des ADNc néo synthétisés, généralement selon une stratégie de *template switching*.

La seconde méthode offre l'avantage de permettre une amplification linéaire, et donc de réduire les biais. En revanche, une quantité initiale minimale de l'ordre de quelques centaines de picogramme d'ARN total est nécessaire pour que la réaction fonctionne. De sorte à atteindre cette limite, le matériel provenant de plusieurs cellules est regroupé, après avoir ajouté un code-barres permettant d'identifier l'origine cellulaire des différentes cibles

A. PCR

Intéressons-nous dans un premier temps à la stratégie d'amplification par PCR. La technologie de séquençage ciblé de l'ARN se focalise sur des transcrits d'intérêt de séquence connue. On peut donc dessiner des couples d'amorces spécifiques des cibles. L'amorce anti-sens sert d'amorce de capture lors de l'étape de RT tandis que l'amorce sens est utilisée pour amplifier les ADNc. Il y a donc autant de couples d'amorces qu'il y a de cibles et l'amplification se fait selon une stratégie de PCR multiplex. Cette stratégie d'amplification multiplexée a d'ailleurs été employée par la technologie CytoSeq pour sélectionner quelques centaines de cibles en parallèle. Cette stratégie offre l'avantage d'être très spécifique et aussi très sensible, capable d'amplifier de très faible quantité de matériel.

Pour rappel, les amorces anti-sens ont été conçues en utilisant l'outil Primer-Blast de NCBI avec les contraintes suivantes :

- l'amplicon fait une longueur de 90 à 160 paires de bases, compatible avec les capacités des séquenceurs Illumina
- un intron est présent, permettant de différencier ARN et ADN génomique
- les amorces sont hautement specifiques
- les amorces sont d'une longueur de 20bp environ et ont un Tm de l'ordre de 60°c pour l'amorce antisens et de 65° pour l'amorce sens

La mise au point de l'amplification par PCR dans le cadre de notre technologie ciblée doit répondre aux besoins suivants ;

- assurer une PCR spécifique,
- amplifier toutes les espèces cibles en parallèle,
- limiter au maximum le nombre de cycle d'amplification pour réduire les biais au séquençage.

Pour ce qui est de la spécificité, des expériences préliminaires d'optimisation de la PCR, non présentées ici, ont démontré qu'augmenter le Tm des amorces sens afin d'effectuer une PCR à haute température d'hybridation permettait de réduire les dimères d'amorces. Les amorces Illumina Rd1 et Rd2 qui sont ajoutées aux extrémités de chaque molécule, car nécessaires au séquençage, ont en commun 15bp et forment donc des dimères à basse température. Malgré des traitements enzymatiques pour éliminer les codes-barres contaminants, ceux-ci sont en tels excès qu'ils sont toujours présents en grande quantité, et étant porteurs d'un site illumina Rd2, ils perturbent la PCR. L'augmentation de température permet de limiter la formation de ces dimères. Un autre moyen de gagner en spécificité est de procéder à une PCR en 2 étapes avec 2 séries d'amorce sens selon une stratégie de PCR imbriquée (*nested PCR*). La seconde amorce sens utilisée se fixe en 5' de la première séquence. Les séquences complètes et correctes sont donc préférentiellement amplifiées lors de cette seconde étape. Cette stratégie a montré de bons résultats mais elle nécessite de concevoir 2 amorces sens par cible, ce qui peut s'avérer complexe étant donné les contraintes appliquées. De plus, 3 étapes de PCR sont nécessaires et autant d'étapes de

purification, pour un total de 40 cycles d'amplification totale (ce qui augmente la fréquence de mutations dues à l'amplification).

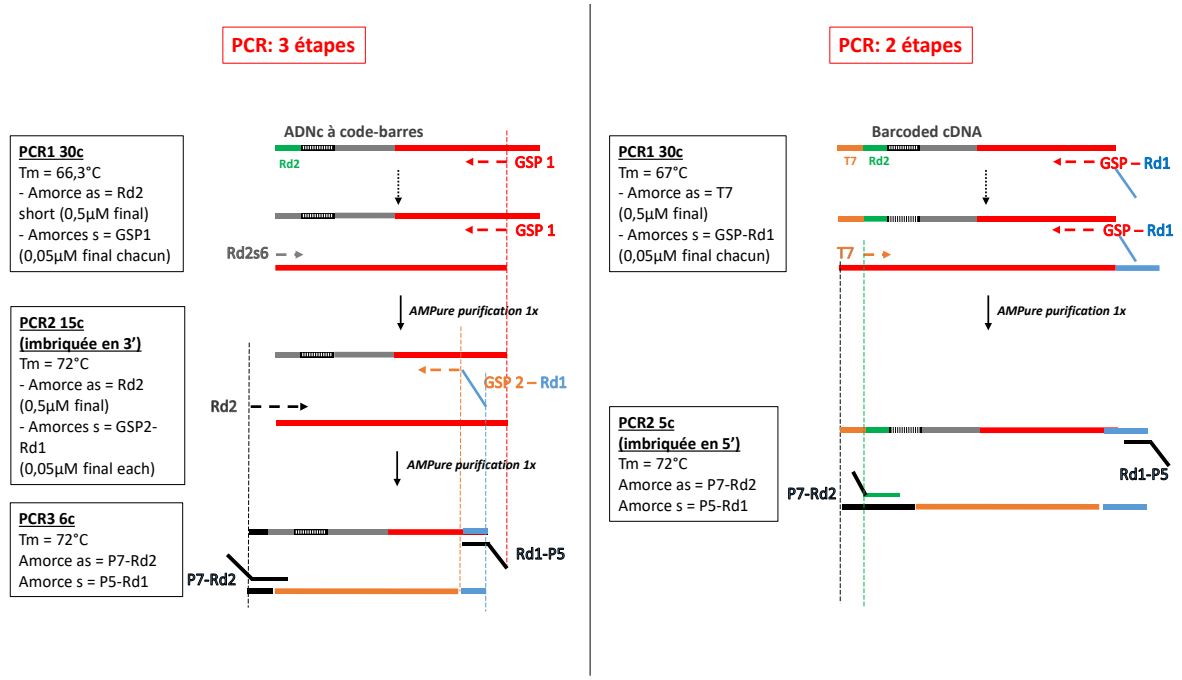


Figure 31 Schéma du protocole d'amplification par PCR multiplex en 2 ou 3 étapes

Après le problème de la spécificité, il faut s'assurer que toutes les espèces sont bien amplifiées et que les amorces sens ne forment pas de dimères lorsqu'elles sont toutes mélangées. La PCR pour chaque amplicon en simplex a dans un premier temps été optimisée, afin de déterminer le meilleur couple d'amorces à utiliser pour chaque cible, en mesurant l'efficacité d'amplification par qPCR. La meilleure température d'hybridation a ensuite été déterminée pour chaque espèce en procédant à des PCR avec gradients de température d'appariement. Finalement, l'impact de l'ajout de $MgCl_2$ (molécule impliquée dans la processivité des polymérases) a été testé. Nous avons ainsi pu définir un protocole optimal, compatible avec tous les amplicons d'intérêt : la température d'hybridation optimale est de 67°C et la PCR se fait en présence de 1,6mM de $MgCl_2$ additionnel (soit une concentration finale de 3,1mM ; Figure 32, B, C).

Afin d'assurer l'amplification de toutes les espèces en parallèle, et ce malgré des différences importantes de quantité de départ, nous avons réduit la concentration des amorces sens à une concentration limitante. Ainsi, lorsqu'une espèce très abondante atteint son plateau d'amplification, toutes ses amorces ont été consommées. L'enzyme

et les réactifs de PCR sont à présent disponibles pour amplifier les espèces moins abondantes ou à l'efficacité d'amplification moins bonne. Nous avons déterminé la concentration minimale d'amorces nécessaire à la réaction d'amplification en diminuant progressivement la concentration en amorces pour chaque espèce, une à une. La concentration minimmale d'amorces permettant une amplification est de 50nM (Figure 32 A). Cette diminution permet également de diminuer les amplifications aspécifiques. Cette concentration est de plus cohérente avec le protocole d'amplification multiplex de la technologie CytoSeq, où chaque amorce sens est ajoutée à une concentration de 50nM également. A ce stade, le protocole d'amplification par PCR multiplex correspond à une amplification de 30 cycles après RT en amorces spécifiques sur 12 cibles ARN, en ajoutant 50nM d'amorces sens, 500nM d'amorces anti sens (commune à tous les ADNc), 3,1mM de $MgCl_2$ final, et en programmant une température d'hybridation à 67°C.

Nous avons comparé l'utilisation d'amorces sens pourvues d'une extension 5' Rd1 ou d'une séquence M13 (non reconnue par le génome de souris ni par le code-barres) ou sans extension. De sorte à éviter la formation de dimères d'amorces Rd1-Rd2, une amorce anti sens T7 a été utilisée. Elle se fixe en amont du site universel Rd2 des ADNc (Figure 31). Ce test a permis de vérifier si la présence de l'extension 5' Rd1 perturbait les premiers cycles de PCR en créant de nombreux dimères d'amorces. Les produits d'amplification ont été déposés sur gel d'agarose pour migration et on constate l'absence de dimères. Le produit amplifié est à la taille attendue dans les 3 cas (Figure 32 E). La stratégie PCR en 2 étapes, avec utilisation d'amorces sens à extension 5' Rd1 lors de la première amplification multiplex a donc été validée. La seconde et dernière étape permet l'ajout des adaptateurs Illumina à l'aide d'amorces P5-Rd1 et P7-Rd2, et constitue donc une PCR imbriquée du côté 5'.

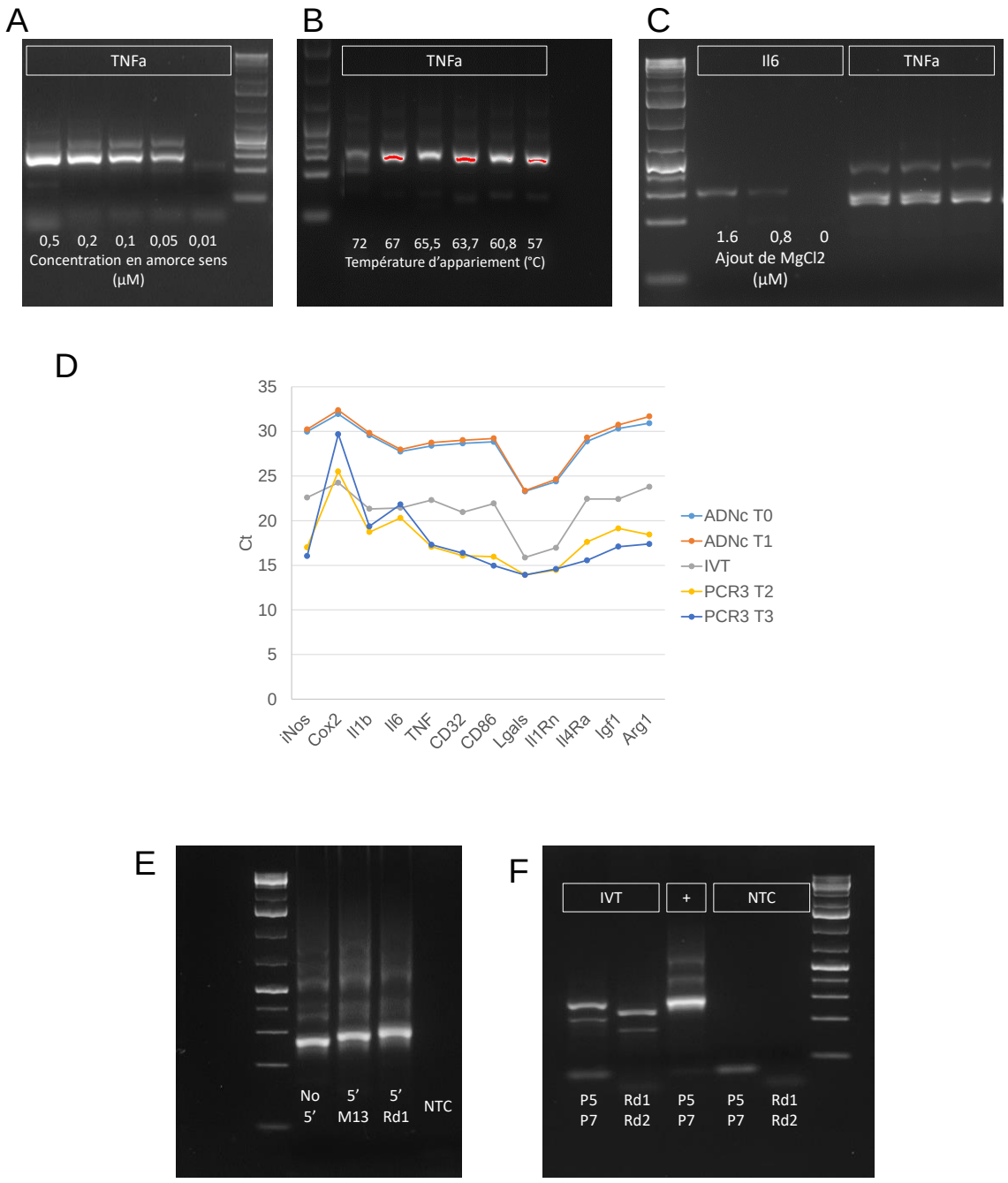


Figure 32 Optimisation de l'amplification des ADNc pour la préparation des librairies de séquençage, A, B, C = Migration sur gel d'agarose des produits d'amplification de PCR multiplex en faisant varier (A) la concentration d'amorce spécifique, (B) la température d'hybridation, (C) la quantité de $MgCl_2$ supplémentaire, D = Mesure de la quantité relative de 12 cibles ADN par qPCR, à l'issue d'une réaction de RT sur ARN total extrait en amorces spécifiques (ADNc T0 et T1), ou après RT et amplification (et dilution 200 fois) par IVT (IVT) ou PCR multiplex de 30 cycles (PCR3 T2 et T3), E = Migration sur gel d'agarose du produit d'amplification d'une PCR multiplex sur 30 cycles avec des amorces sens porteuses d'une extension 5' Rd1, ou 5' M13 ou sans extension, F = Produit d'amplification final après une stratégie d'amplification par IVT ou par PCR (puits +), NTC = Contrôle négatif

Enfin, nous avons vérifié que chaque espèce avait bien été amplifiée en procédant à une qPCR cible par cible du produit d'amplification dilué 200 fois. Un contrôle positif, correspondant à des ADNc non amplifiés et non dilués a été inclus (Figure 32, D). On constate que l'amplification fonctionne pour toutes les cibles, bien que moins efficace pour la cible Cox2. On constate un biais d'amplification. Le profil d'expression des 12 cibles n'est pas le même avant ou après amplification. Cette stratégie de PCR multiplex en concentration limitante d'amorces sens permet d'amplifier toutes les cibles, en tendant à atteindre un plateau d'amplification pour chacune d'elles et les amener ainsi à une quantité finale similaire. La mauvaise amplification de la cible 2, qui correspond d'ailleurs au transcrit le plus faiblement exprimé, pourrait être améliorée en augmentant le nombre de cycles de la PCR multiplex.

B. Transcription *in vitro*

L'amplification par transcription *in vitro* est intéressante de par son aspect linéaire. Les résultats précédents ont permis de mettre en évidence la spécificité de la réaction de RT ; l'étape de capture est donc à priori suffisante pour sélectionner les cibles d'intérêt.

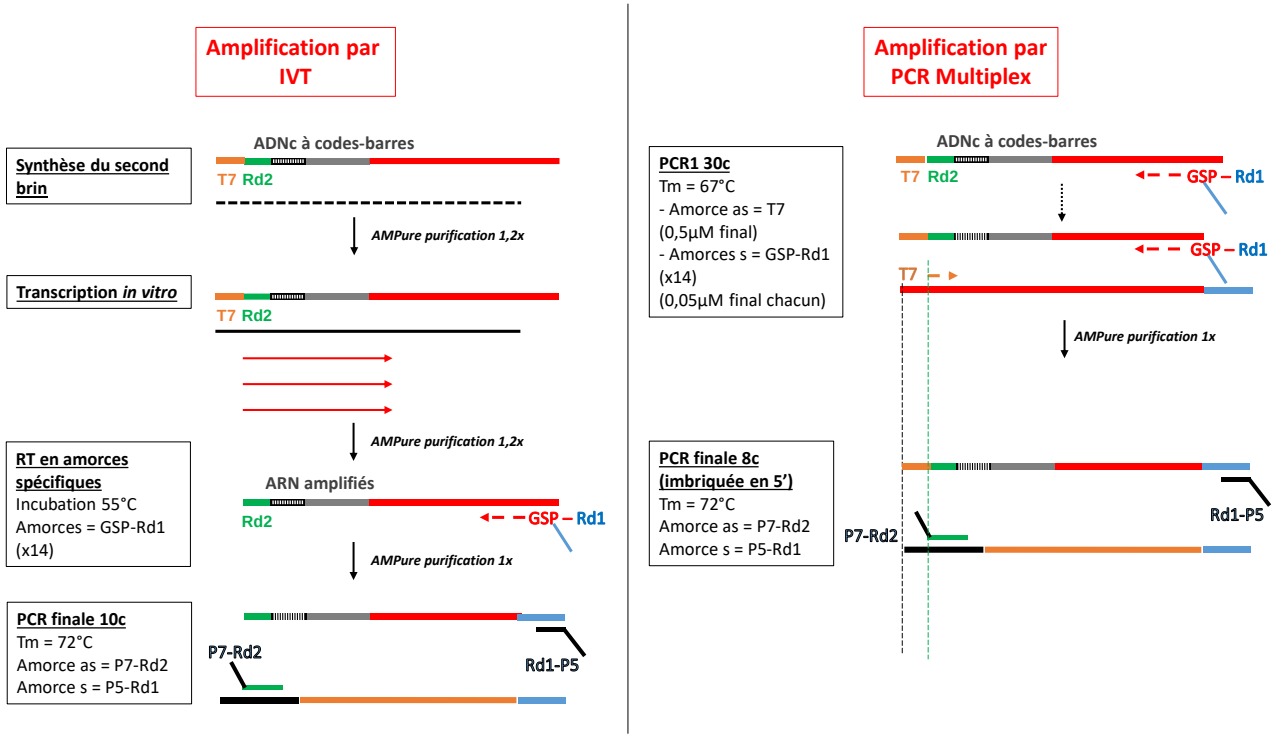


Figure 33 Schéma d'amplification selon une stratégie par PCR multiplex (Droite) ou par IVT (Gauche)

La stratégie inDrop a été adaptée à nos besoins et est présentée sur la partie droite de la Figure 33. Les ADNc à code-barres sont porteurs d'un promoteur T7. Le second brin est synthétisé suite à l'action d'une RNAse H, qui génère des amorces à partir du brin ARN, d'une ADN polymérase I et d'une ligase. Cet ADN double brin sert de matrice pour une transcription *in vitro* conduisant à l'amplification de cette matrice sous forme d'ARN par une T7 RNA polymérase. Afin de générer des amplicons de 100 à 150 paires de bases et d'augmenter la spécificité de sélection des cibles, cet ARN amplifié est rétro transcrit en ADNc en utilisant les amorces sens précédemment décrites, pourvues en 5' d'un site universel illumina Rd1. Une étape ultime de PCR permet l'ajout des adaptateurs Illumina P5 et P7.

Une RT sur ARN total extrait de cellules BV2 a été effectuée à l'aide de billes d'hydrogel à code-barres et amorces spécifiques de 12 cibles. Le code-barres est porteur d'un site de restriction, d'un promoteur T7, d'un site universel illumina Rd2 et de 3 index B, C et D. A l'issue de l'étape de synthèse des ADNc et après élimination des codes-barres contaminants par traitement ExoI, les ADNc ont été amplifiés selon une stratégie PCR ou IVT. A l'issu de l'étape d'IVT, une purification est effectuée, avant de passer à l'étape de RT avec les amorces sens à extension Rd1. Les ADNc générés peuvent alors être amplifiés par PCR pour l'ajout des adaptateurs Illumina P5 et P7. 1µL du produit de RT ont été dilué 200 fois puis soumis à un contrôle par qPCR de l'amplification des différentes cibles. Les résultats de la qPCR, présentés dans la Figure 32D, montrent que le profil d'expression des différentes cibles est le même avant amplification ou après amplification par IVT, alors que celui à la suite d'une amplification par PCR offre un profil différent. L'amplification par IVT est donc bien linéaire et permet d'amplifier toutes les cibles ayant été rétro-transcrites.

La dernière amplification a été faite à l'aide d'amorces P5-Rd1 et P7-Rd2 ou simplement Rd1 et Rd2, sur 10 cycles. Le produit d'amplification est déposé pour migration sur gel d'agarose (Figure 32, F), avec le produit d'amplification finale par PCR, servant de contrôle positif.

En conclusion, les 2 stratégies d'amplification permettent de générer un produit final de taille attendue. La stratégie par IVT, bien que plus laborieuse, ne nécessite que 10 cycles d'amplification exponentielle, contre 40 cycles pour la stratégie par PCR. Cette option de préparation des librairies de séquençage par IVT apparaît donc comme une solution intéressante pour limiter les biais d'amplification.

V. Conclusion et perspective

Nous avons développé une technologie de RNAseq ciblé en gouttes à l'échelle de la cellule unique, en nous inspirant de différentes technologies de scRNAseq, et notamment de la technologie inDrop, une technologie à ultra haut débit permettant d'encapsuler dans des gouttes de quelques nanolitres cellules uniques et billes d'hydrogel, porteuses de code-barres de goutte. Ce code-barres est le même pour une bille donnée mais diffère d'une bille à l'autre. Tous les transcrits provenant d'une même goutte sont donc associés à un seul et même code-barres. Dans cette technologie, la capture des ARNm se fait sur des amorces polyT, point d'initiation de la RT qui se déroule à l'intérieur des gouttes. Les ADNc néo synthétisés sont libérés par cassage de l'émulsion et sont amplifiés et préparés au séquençage.

Une étape déterminante de ce protocole sur la qualité des résultats obtenus est la capture des ARNm sur les billes d'hydrogel et leur RT en ADNc. Nous avons dans un premier temps cherché à améliorer cette étape. Des expériences préliminaires en tube nous ont permis d'identifier un point clé à améliorer pour optimiser notre technologie, à savoir la quantité d'amorces spécifiques délivrées en gouttes, l'efficacité de la RT augmentant avec la concentration en amorces spécifique, jusqu'à atteindre un plateau. Cette quantité est directement liée à la capacité fonctionnelle des billes d'hydrogel, qui apportent dans chaque goutte des amorces à code-barres unique. Nous avons donc optimisé les billes d'hydrogel, en comparant différentes chimies, en améliorant le protocole de construction du code-barres qui passe par plusieurs étapes de ligation et enfin en variant la taille de la bille.

Nous nous sommes ensuite intéressés à une autre étape clé du protocole ; l'amplification des ADNc. Plusieurs méthodes ont été décrites pour les différentes technologies scRNAseq, et notamment une amplification par PCR multiplex, technique utilisée par la technologie ciblée CytoSeq. Une amplification par PCR offre l'avantage de pouvoir amplifier de très faibles quantités de matériel et d'être très spécifique. En revanche, le nombre répété de cycle d'amplification peut engendrer l'apparition de mutations ponctuelles et des biais d'amplification, en favorisant l'amplification de certaines espèces au détriment d'autres. Afin de pallier ces biais, nous avons incorporé à nos amorces à code-barres une séquence UMI, pour normaliser les données de séquençage. Nous avons optimisé le protocole de PCR multiplex, en diminuant la concentration des amorces sens à une concentration limitante, en ajoutant du $MgCl_2$,

ou encore en procédant à des étapes d'amplification emboitées ou imbriquées (*nested PCR)*.

En parallèle, nous avons également mis au point un protocole d'amplification linéaire par IVT en nous inspirant de celui de la technologie inDrop, mais en faisant appel à des amorces spécifiques pour procéder à la seconde étape de RT, post-IVT, permettant la sélection spécifique des cibles d'intérêt et l'ajout d'une séquence universelle Illumina à l'extrémité de l'amplicon de 100 à 150 paires de bases sélectionné. Cette méthode d'amplification montre peu de biais d'amplification. Le profil d'expression de plusieurs cibles s'est avéré similaire à celui obtenu avant amplification, c'est-à-dire qu'aucune cible n'a été plus favorablement amplifiée qu'une autre, contrairement à ce qui est observé pour la PCR ;

Il est à noter que ce protocole a été validé en tube et que le passage en gouttes pourrait réduire la quantité de matériel de départ et jouer sur l'efficacité de la réaction d'IVT, moins sensible de la PCR.

Finalement, les points optimisés vont devoir être validés en gouttes à l'échelle de la cellule unique.

Chapitre 4 : La microglie

I. Introduction

A. Origine et fonctions

Le cerveau est un organe à part, cloisonné par la barrière hémato encéphalique. Il a son propre système immunitaire ; la microglie. Les cellules microgliales sont semblables aux macrophages périphériques, qui sont des cellules phagocytaires appartenant au système immunitaire inné. Les cellules de la microglie et les macrophages se distinguent cependant en plusieurs points tels que leur origine, leur durée de vie et l'éventail de leur fonction, de même que leur profil transcriptomique (Butovsky *et al.*, 2014)(Ginhoux *et al.*, 2010). Les cellules microgliales proviennent du sac vitellin, alors que les macrophages ont pour origine la moelle osseuse. Les cellules microgliales migrent du sac vitellin à un stade précoce du développement pour aller dans le système nerveux central (SNC), à 8-9 jours du stade embryonnaire chez la souris A l'âge adulte, les cellules microgliales de longue durée de vie se maintiennent par renouvellement à l'intérieur du SNC et non par l'apport de nouvelles cellules progénitrices périphériques via la circulation (Ginhoux & Prinz, 2015).

Les cellules microgliales représentent 10% des cellules du SNC chez l'adulte (Kettenmann, Hanisch, Noda, & Verkhratsky, 2011). Elles sont responsables de nombreuses fonctions très variées, en plus de leur rôle inflammatoire (Salter & Stevens, 2017). Ainsi, elles participent à l'architecture des circuits neuronaux, via l'élagage des synapses en excès juste après la naissance (Paolicelli *et al.*, 2011), et également à la plasticité synaptique et neuronale en lien avec la mémoire et l'apprentissage chez l'adulte (Salter & Beggs, 2014). A l'état de repos, ces cellules hautement ramifiées scannent en permanence la totalité du SNC afin de maintenir l'homéostasie du milieu et de répondre rapidement à tout signal. Elles éliminent les débris cellulaires et, en cas d'infection, sont activées et démarre un programme inflammatoire (Tremblay *et al.*, 2011).

Ces cellules sont de plus en plus étudiées, tant pour leurs nombreuses fonctions dans le cerveau normal, que pour l'étude de la neuro-inflammation mais aussi leur implication dans de nombreuses pathologies (Salter & Stevens, 2017).

B. La microglie à l'échelle de la cellule unique

Les cellules microgliales, longtemps considérées comme les macrophages du SNC, ont depuis révélé leurs fonctions variées et leurs nombreuses implications, que ce soit dans le SNC sain ou malade. Il est donc naturel que ce système complexe soit étudié à l'échelle de la cellule unique, au même titre que les cellules du système immunitaire périphérique (SNP) (Gertig & Hanisch, 2014).

De nombreuses publications mettant en avant l'étude de la microglie à l'échelle de la cellule unique ont vu le jour ces dernières années.

Ainsi, la diversité microgliale au cours du développement a été analysée en utilisant la technologie MARS-seq (Jaitin & Kenigsberg & Keren-Shaul, Elefant *et al.*, 2014). Un programme de développement en plusieurs étapes a été mis en évidence, où les cellules microgliales évoluent de manière coordonnée, et synchronisée avec le développement neuronal (Matcovitch-Natan & Winter, *et al.*, 2016).

Plus récemment, Hammond et son équipe ont mené une étude cellule unique à très haut débit, en utilisant la technologie 10X Chromium (Zheng *et al.*, 2017) sur près de 80 000 cellules microgliales issues de souris, au cours du développement, à un âge avancé, ou après une lésion au cerveau. Ils ont montré l'hétérogénéité de la microglie dans le temps, et en fonction de l'état du cerveau. Le nombre de groupes est plus important pendant le développement, signe d'une plus grande diversité, et certains de ces sous-groupes présentent des traits communs avec des sous-groupes identifiés en cas de pathologie (Hammond *et al.*, 2018).

La microglie joue un rôle important dans les neuropathologies, tantôt protecteur, tantôt délétère. Ainsi, l'implication de la microglie dans la maladie d'Alzheimer a été finement étudiée à l'échelle de la cellule unique à l'aide de la technologie MARS-seq et a permis de mettre en évidence un sous-type de cellules microgliales associées à la pathologie, chez des cerveaux de souris transgénique mimant la maladie d'Alzheimer. Cette population cellulaire, baptisée DAM (*Disease-Associated Microglia*), semble avoir un effet neuroprotecteur. Des études précédentes, à l'échelle populationnelle, allaient au contraire plutôt dans le sens d'un effet délétère de la microglie, pointant du doigt la présence de marqueurs pro-inflammatoires engendrant une accélération de l'évolution de la maladie (Heppner, Ransohoff, & Becher, 2015). Cette découverte pourrait ouvrir

de nouvelles stratégies thérapeutiques (Keren-Shaul *et al.*, 2017) (Deczkowska *et al.*, 2018).

Les technologies cellule unique ont également été utilisées dans le cadre de l'étude de l'inflammation. Sousa et son équipe font appel à la technologie Drop-seq (Macosko *et al.*, 2015) pour analyser la réponse microgliale suite à une injection de LPS (Lipopolysaccharide), une endotoxine d'origine bactérienne Gram négative, qui permet de mimer une infection et d'induire une réponse inflammatoire (Moreillon & Majcherczyk, 2003)(Sousa *et al.*, 2018). L'analyse des données de scRNAseq, sur cellules issues de souris perfusées avec du LPS ou avec une solution saline (contrôle), a montré la présence de 2 groupes principaux distincts, correspondant aux 2 traitements appliqués. Les auteurs ont ensuite comparé le profil d'expression du groupe de cellules qui apparaît suite à une activation au LPS à celui des cellules DAM décrites dans le cadre de l'étude de la microglie et de la pathologie d'Alzheimer. Ils ont montré que les cellules activées au LPS présentent une réponse inflammatoire élevée, avec une surexpression de gènes pro-inflammatoires et une sous-expression de gènes homéostatiques, tandis que les cellules DAM (Keren-Shaul *et al.*, 2017) montrent une signature phagocytaire/lysosomale, signe que la microglie est capable de s'adapter finement à son environnement en fonction des signaux reçus ; une injection de LPS ou l'accumulation de plaques amyloïdes. Finalement, les auteurs ont observé que les cellules activées au LPS se divisent en fait en 2 sous-groupes, un groupe principal associé à une réponse inflammatoire forte, et un sous-groupe de cellules correspondant à un état inflammatoire intermédiaire, moins activé. Tous ces résultats combinés sont le signe de l'hétérogénéité de la réponse microgliale *in vivo* face à un signal donné mais également en fonction du signal.

C. Inflammation au cours du développement

La fonction première de la microglie, et celle qui nous intéresse dans le cadre de notre étude, est son rôle inflammatoire. La microglie surveille en permanence l'environnement du SNC, et en cas d'invasion par un corps étranger, va déclencher une réponse immunitaire rapide (Das *et al.*, 2015). Cette réponse va être adaptée au signal perçu.

De la même manière que les macrophages auxquels elles sont souvent comparées, les cellules microgliales sont polarisées et présentent 2 états activés principaux, qui vont être induits par différents stimulus. Un état M1, correspondant à un phénotype pro-inflammatoire, est initié par une induction au LPS, et un état M2, correspondant à un état d'activation alternatif, par une induction à l'IL-4 par exemple (Popiolek-Barczyk & Mika, 2016). Chacun de ces phénotypes est associé à des marqueurs de surface spécifiques et à un profil de sécrétion protéique différent.

Il a été démontré que la persistance du phénotype pro-inflammatoire, lors d'une inflammation chronique par exemple, peut entraîner une dégradation conduisant à une lésion cérébrale. En effet, une analyse d'un cerveau prématuré, présentant une lésion de la substance blanche, a mis en évidence un phénotype activé cytotoxique M1 chez la microglie. D'autres études ont montré que les cytokines pro-inflammatoires sécrétées par les cellules microgliales de type M1 altèrent la prolifération et la maturation des oligodendrocytes (Kadhim, Tabarki, De Prez, Rona, & Sebire, 2002). Cependant, lors de l'exposition à des cytokines anti-inflammatoires, telles que la cytokine IL-4, induisant un phénotype activé de type M2, la différenciation des oligodendrocytes est stimulée et une remyélinisation des axones est observée (Miron *et al.*, 2013).

10% des naissances dans le monde sont prématurées (Liu *et al.*, 2012). Malgré une diminution du taux de mortalité depuis des décennies, le taux de morbidité reste lui, assez élevé, avec la recrudescence de problèmes neurologiques associés à la prématurité, tels que des déficits moteurs fins ou des troubles cognitifs et d'apprentissage. L'amélioration du taux de morbidité des enfants prématurés constitue un défi mondial. Le SNC peut être exposé à une inflammation en réponse à une infection virale ou bactérienne, mais également après des atteintes stériles. Normalement, les processus anti-inflammatoires et de réparation permettent le retour

à la normale. Néanmoins, une inflammation chronique du SNC peut entraîner des lésions de la substance blanche et des conséquences neurologiques (Hagberg, Gressens, & Mallard, 2012)(Hagberg *et al.*, 2015).

Une stratégie prometteuse pour réduire les lésions de la substance blanche pourrait consister à supprimer le phénotype M1 pro-inflammatoire et à promouvoir le phénotype M2 régénérateur de la microglie. Afin d'élaborer des traitements permettant de promouvoir cette transition, il est nécessaire de bien caractériser ces phénotypes et de comprendre les voies intracellulaires impliquées dans les différents mécanismes d'activation (Loane & Kumar, 2015). Une étude à l'échelle de la cellule unique offre une profondeur d'analyse plus grande pour améliorer la compréhension des mécanismes entrant en jeu dans cette transition. De plus, une étude à très haut débit permettra de paralléliser de nombreuses expériences afin de tester une multitude de cible thérapeutique en observant si la transition phénotypique a bien eu lieu.

Ce projet a été mené en collaboration avec l'équipe de Pierre Gressens de l'hôpital Robert Debré, spécialiste de la microglie et de la prématurité. Le but est de mettre au point une technologie à l'échelle de la cellule unique et à très haut débit, capable d'identifier les marqueurs de l'inflammation. *In fine*, cette technologie sera utilisée pour élucider les mécanismes d'actions entrant en jeu lors de la réponse inflammatoire lors des étapes précoces du développement. En effet, le nombre de publications concernant l'étude de la microglie à l'échelle de la cellule unique est en plein essor. Ces études se sont essentiellement intéressées à la microglie dans des cerveaux sains à différents stades de développement, en montrant des grandes disparités entre les phases précoces et adultes. D'autres études quat à elles se sont portées sur des cas pathologiques d'inflammation chez l'adulte, telles que la maladie d'Alzheimer, afin de rechercher d'éventuels sous-types associés à la maladie.

Nous voulons nous focaliser sur l'inflammation à un stade précoce et plus particulièrement sur l'expression de gènes impliqués dans le processus inflammatoire afin de comprendre la spécificité de la réponse inflammatoire à un stade très différent du stade adulte.

Le second objectif est d'utiliser cette technologie pour identifier l'état d'activation de la microglie après application à l'échelle de la cellule d'une banque combinatoire de médicaments approuvés par la FDA ; les médicaments ont plusieurs cibles et la

combinaison de deux médicaments peut avoir un effet synergique imprédictible (Bollenbach, Quan, Chait, & Kishony, 2009), pouvant permettre une modulation phénotypique. La dimension microfluidique permet de tester un très grand nombre de combinaisons en parallèle grâce à la haute fréquence de production des gouttes et le résultat de chaque combinaison sur la modulation phénotypique pourra être mesuré en utilisant notre technologie de RNAseq ciblé en gouttes.

II. Séquençage ciblé de l'ARN pour l'étude de l'inflammation chez la microglie

Etant un système complexe et hétérogène, la microglie constitue un sujet de choix pour une étude à l'échelle de la cellule unique. Nous souhaitons nous focaliser sur l'étude de l'inflammation de la microglie lors des premières étapes du développement. Dans un premier temps, nous avons utilisé une lignée modèle de microglie de souris, les cellules BV2 (Henn *et al.*, 2009), et en induit un état inflammatoire par ajout de LPS (Stansley, Post, & Hensley, 2012)(Das *et al.*, 2015). Le LPS utilisé est extrait à partir de la souche E.Coli O55 :B5. Cette molécule, naturellement présente dans la paroi cellulaire externe des bactéries Gram négatives, consiste en une partie lipidique A liée à un O-polysaccharide antigénique. Elle stimule le système immunitaire inné, et notamment les macrophages et la microglie, via le récepteur TLR4.

Les expériences présentées au cours de ce chapitre consistent en la mise au point et l'optimisation de notre technologie de RNAseq ciblé. L'objectif est d'être capable de détecter des marqueurs de l'inflammation et de quantifier la différence d'expression, à l'échelle de la cellule unique, en cas d'inflammation, chez une lignée cellulaire modèle.

Des expériences préliminaires de séquençage ciblé de l'ARN sur des cellules BV2 ont montré qu'il était possible de quantifier l'expression de transcrits cibles à l'échelle de la cellule unique, mais aussi que certains transcrits faiblement exprimés n'étaient que peu ou pas détectés. Nous avons donc entrepris d'optimiser notre technologie ciblée en gouttes, en nous basant sur les conclusions faites à l'aide des expériences décrites au cours du chapitre précédent, et à partir la lignée cellulaire BV2.

Afin de mimer le processus d'inflammation microgliale, du LPS bactérien est ajouté à des cellules BV2 déposées et adhérant au fond des puits d'une plaque 6 puits, permettant ainsi de mimer une infection bactérienne chez des cellules en culture. Le LPS étant dissout dans du PBS, pour chaque expérience d'activation, un contrôle évaluant l'impact du PBS dans l'activation des cellules BV2 est réalisé par ajout de la même quantité de PBS qu'il a été mis de LPS dans les puits de la plaque.

A. Expériences *in vitro* et choix des transcrits cibles

La première étape a consisté à choisir les meilleurs transcrits cibles, en procédant à des expériences de RT en amorces spécifiques *in vitro* suivi d'une quantification par qPCR. Notre but est d'identifier des transcrits suffisamment exprimés, parmi une liste de marqueurs de l'inflammation présentant une expression différentielle en cas d'inflammation, et en nous rapportant à l'expression d'un gène de référence dont la détection en gouttes est très bonne.

Les cibles à tester ont été préalablement sélectionnées en collaboration avec notre partenaire sur la base de leurs données d'expériences faites en puces à ADN Affymetrix et de RT-qPCR sur des cellules de microglie primaire, issues de nouveau-nés souris, en état d'inflammation systémique ou non (Krishnan *et al.*, 2017)(Chhor *et al.*, 2013). Les transcrits à cibler ont ainsi été choisis sur la base de leur expression différentielle entre l'état d'inflammation ou le contrôle. Il est à noter que les expériences sur puces à ADN ont été faites sur des cellules de microglie primaires alors que nos expériences sont faites sur une lignée cellulaire modèle ; il peut donc y avoir des différences dans l'expression de certains transcrits.

Avant de débuter les expériences en gouttes, nous avons validé le choix des transcrits ainsi que l'efficacité de leurs amorces spécifiques en procédant à une expérience de RT en tube sur des ARNs extraits à partir de cellules BV2 activées (LPS dans du PBS) ou non (PBS seul). Les niveaux d'expression relatifs de chaque transcrit dans les 2 échantillons (activés ou non) ont ensuite été évalués par qPCR. La quantité initiale d'ADNc, néosynthétisé lors de l'étape de RT, présents dans l'échantillon, est estimée grâce au cycle à partir duquel de la fluorescence est détectée lors de la qPCR, nommé Ct (*cycle threshold* ou cycle seuil). Ainsi, une espèce ayant été plus fortement exprimée aura un Ct plus petit. Après avoir étudié les courbes de dissociation, qui permettent de savoir combien d'espèces ont été générées à l'issu des cycles d'amplification à partir d'un couple d'amorces donné, les amorces non spécifiques, c'est-à-dire entraînant la formation d'une ou plusieurs espèces contaminantes, sont éliminées des analyses. Elles ne seront pas utilisées pour les expériences en gouttes. Plusieurs couples d'amorces ont ainsi été testés pour chaque cible potentielle. L'efficacité de chaque couple est déterminée en calculant le facteur d'amplification

entre différents points de dilution. Une efficacité de 100% correspond à une multiplication par 2 de la quantité d'amplicons à chaque cycle d'amplification. Sont conservées les amorces dont l'efficacité est comprise entre 90 et 110%. Finalement, nous avons sélectionné les transcrits dont l'expression est suffisamment grande pour être plus facilement détectable lors de l'expérience en gouttes. La Figure 34 présente les résultats de la RT qPCR en amorces spécifiques.

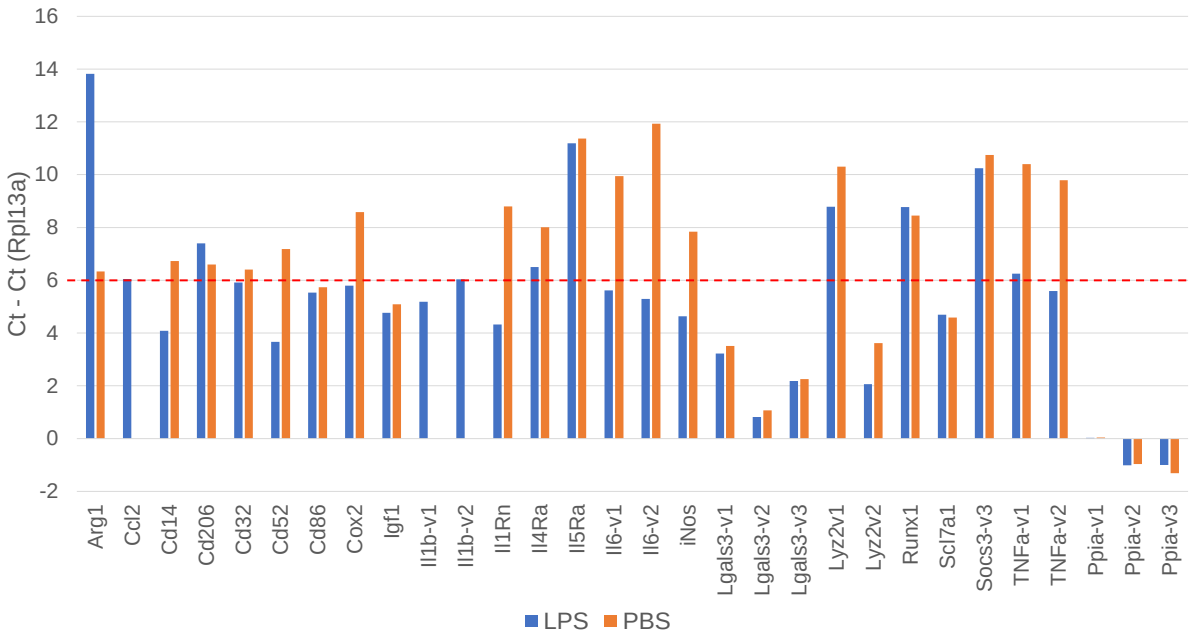


Figure 34 Quantification du niveau d'expression relatif de cibles d'intérêt par RT-qPCR en amorces spécifiques sur ARN extrait à partir de cellules BV2, inflammés par ajout de LPS, ou non (contrôle PBS). Plusieurs cibles ont été testées, et la différence d'expression par rapport à un gène de référence fortement exprimé (Rpl13a) a été calculée par soustraction des Ct obtenus. La courbe de dissociation des espèces amplifiées a été préalablement analysée afin de ne conserver que les amplifications spécifiques.

La différence de Ct par rapport à celui du gène de référence fortement exprimé Rpl13a est calculée pour chaque transcrit. Le gène Ppia représente également un gène de référence. Les transcrits montrant un écart de 6 Ct (dans l'une ou l'autre condition) par rapport à la référence Rpl13a (Ct-Ct(Rpl13a)) sont considérés comme étant exprimés à un niveau de base suffisant. Sur la base de ces résultats, une quinzaine de transcrits ont été présélectionnés en tant que candidats marqueurs de l'inflammation pour nos expériences en gouttes.

La dernière étape fut de valider le bon fonctionnement de ces couples d'amorces et leur compatibilité en PCR multiplex. 14 transcrits cibles ont été conservés à l'issu de

cette étape. Ils sont présentés dans le Tableau 3, avec leur code de référence RefSeq, la fonction de la protéine qu'ils encodent chez la microglie dans le cadre de l'inflammation ainsi que le phénotype d'inflammation auquel ils sont associés.

Cible	RefSeq Code	Nom complet	Phénotype	Fonction protéique chez la microglie dans le cadre de l'inflammation
iNos	NM_010927.3	*Inducible nitric oxide synthase*	M1	Enzyme participant à la production du radical neurotoxique NO (Nitric Oxyde) en réponse à un des médiateurs de l'inflammation de type LPS, médiateur de la pro-inflammation
Il1b	NM_008361.4	*Interleukine 1 beta*	M1	Cytokine médiatrice de la pro-inflammation, impliquée dans la prolifération cellulaire, la différenciation et l'apoptose
TNFa	NM_013693	*Tumor Necrosis Factor alpha*	M1	Cytokine médiatrice de la pro-inflammation et inductrice de l'apoptose neuronale
Cd32	NM_001077189.1	*Cluster of Differentiation 32* *IgG Fc receptor II-a*	M1	Co-récepteur des cellules B, liant la région fonctionnelle constante (Fc) des IgG. Impliqué dans l'induction de la phagocytose
Cd86	NM_019388.3	*Cluster of Differentiation 86*	M1	Co-récepteur impliqué dans l'activation des cellules T Modulateur phénotypique de la microglie
Cd14	NM_009841	*Monocyte differentiation antigen CD14*	M1	Co-récepteur pour la fixation du Lipopolysaccharide bactérien, médiateur de l'immunité inné et de l'inflammation
Lyz2	NM_017372	*Lysozyme C2*	M1	Fonction bactériolytique via une activité d'hydrolyse et de Trans glycosylation
Ccl2	NM_011333	*C-C motif chemokine 2*	M1	Chimiokine permettant le recrutement et la prolifération de cellules microgliales au site de sécrétion pendant l'inflammation
Lgals3	NM_010705.3	*Galectin 3*	M1 / M2	Lectine Galactose spécifique fixant les IgE. Impliqué dans la chimio-attraction, l'apoptose, l'inflammation innée. Régulateur via son implication dans l'épissage d'ARNm dans le noyau
Il1Rn	NM_031167.5	*Interleukine 1 Receptor antagonist*	M2	Récepteur fixant les cytokines IL1b et IL1a pour inhiber leur effet pro-inflammatoire. Effet neuroprotecteur
Il4Ra	NM_001008700	*Interleukine 4 Receptor subunit alpha*	M2	Récepteur de la cytokine Il4 sécrétée par les neurones, induisant une transition vers une phénotype M2, via une boucle de régulation positive Effet neuroprotecteur
Arg1	NM_007482.3	*Arginase 1*	M2	Cicatrisation et régénération des tissus via la production de polyamine (L-ornithine) par dégradation d'arginine
Igf1	NM_010512.5	*Insulin-like growth factor I*	M2	Médiateur de la transition de la microglie vers un phénotype de type M2
Scl7a1 =Cat1	NM_007513	*Solute Carrier family 7, member 1, High affinity cationic amino acid transporter 1*	M0	Transport de l'Arginine dans des conditions basales

Tableau 3 Marqueurs de l'inflammation sélectionnés pour le séquençage ciblé de l'ARN en cellules uniques de lignée BV2, activées par ajout de LPS

B. Comparaison de différents protocoles de RNAseq ciblé en gouttes

1. Résumé des résultats obtenus

Nous avons effectué une expérience de séquençage ciblé de l'ARN sur des cellules BV2 (activées ou non) afin d'optimiser différentes étapes du protocole. L'efficacité de la capture des ARNm sur les billes à amorces spécifiques à code-barres puis leur RT en ADNc en gouttes ont ainsi été comparées pour 2 volumes de gouttes différents. Le protocole d'amplification des ADNc a également été étudié, en comparant une stratégie d'amplification linéaire par IVT contre une amplification exponentielle par PCR multiplex. La Figure 35 présente le schéma expérimental.

Au cours de ce paragraphe, nous allons ainsi répondre aux questions suivantes :

- L'analyse moyennée des données en RNAseq ciblé est-elle similaire à celle obtenue à partir de données de RT-qPCR en tube à partir de mêmes échantillons biologiques ?
- Quel est le protocole de RT en gouttes optimal ? Les expériences d'optimisation en tube ont montré qu'un point crucial affectant l'efficacité de la réaction de RT est le nombre d'amorces disponibles. Ces mêmes paramètres ont été testés en gouttes sur des cellules BV2 ?
- Quel est le protocole de préparation des libraires optimal ? Les expériences d'optimisation en tube ont également permis de mettre au point deux protocoles d'amplification suivant une RT en amorces spécifiques. La comparaison de ces 2 protocoles a été faite par RT-qPCR. Nous allons maintenant comparer ces 2 méthodes sur des données de séquençage à l'échelle de la cellule unique

Les réponses à ces différentes questions sont résumées ci-dessous :

Une première analyse moyennée a été effectuée à partir des données obtenues à l'issue des expériences en gouttes ou d'expériences contrôles effectuées en tube sur ARN extraits à partir des mêmes cellules que celles utilisées en gouttes. Nous avons constaté que les résultats en gouttes étaient similaires à ceux obtenus par RT-qPCR en tube, signifiant que la RT en gouttes sur amorces spécifiques est fonctionnelle et permet d'aboutir aux mêmes conclusions biologiques qu'en tube donc sans biais expérimental.

Les expériences en gouttes de volumes différents nous ont permis de conclure que la RT en gouttes fonctionne aussi bien dans des gouttes de 100pL que dans des gouttes de 2nL, avec une efficacité meilleure quant à la capture et la RT des ARNm faiblement exprimés dans des gouttes plus volumineuses en combinaison avec des billes à la capacité fonctionnelle plus grande (nombre d'amorces plus grand).

Le choix du meilleur protocole d'amplification est plus discutable. Une stratégie d'amplification par PCR semble plus sensible, permettant de collecter bien plus d'UMI par gouttes. Cependant, des biais techniques peuvent artificiellement surévaluer ce nombre, via la création de faux UMI par mutation. Pour pouvoir déterminer définitivement le protocole le plus compatible avec notre modèle expérimental, des investigations plus poussées devront être effectuées, notamment via la mise au point d'outils bio-informatiques de filtrations des UMI.

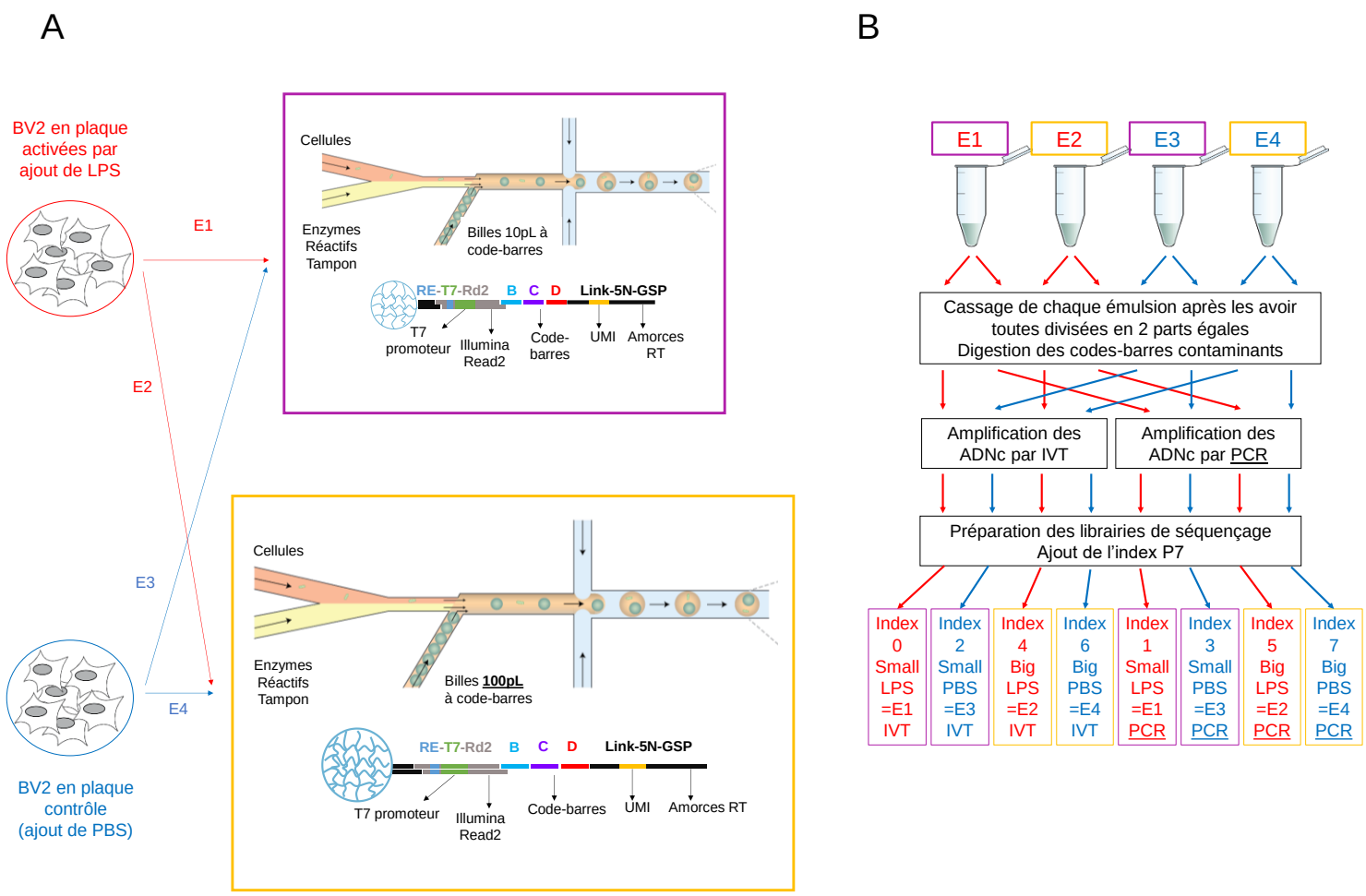


Figure 35 Schéma expérimentale de la RT en amorces ciblées en gouttes pour l'optimisation de plusieurs étapes du processus

2. Comparaison des résultats moyens obtenus en masse ou à l'échelle de la cellule unique

Nous avons voulu, dans un premier temps, comparer les résultats obtenus en tube à ceux obtenus en gouttes, afin d'évaluer leur similarité et leur cohérence biologique. Pour cela, nous avons effectué une réaction de RT en tube à partir d'ARN extraits de cellules BV2 activées au LPS ou contrôles (en présence de PBS) et de billes PA porteuses d'amorces à code-barres, spécifiques de 14 transcrits d'intérêt. Les billes utilisées pour l'expérience en masse sont les mêmes que celles utilisées pour les expériences en gouttes de 100pL (Emulsions 1 et 3, Figure 35) et l'ARN a été extrait à partir des mêmes cellules que celles utilisées pour les expériences d'encapsulation. Un million de cellules BV2, activées ou non, ont en effet été congelées avant extraction en parallèle de l'expérience de RT en gouttes. L'analyse moyennée de l'expression relative des différents transcrits cibles entre des cellules activées ou non est similaire dans le cas en masse ou en gouttes (Figure 36), preuve que la réaction en gouttes fonctionne convenablement et permet d'arriver à des conclusions biologiques faisant sens.

Afin de comparer des expériences en masse à des expériences à l'échelle de la cellule unique, les expressions relatives moyennes des différentes cibles ont été analysées, soit par RT-qPCR dans le cadre de l'expérience en masse, soit par quantification de la proportion d'UMI associés à chaque transcrit par rapport au nombre d'UMI totaux pour chacun des 8 échantillons séquencés. Les résultats sont présentés sous forme de *clustermap* dans la Figure 36, un graphique combinant une dimension *heatmap*, permettant de visualiser sous forme de niveau de couleurs les niveaux d'expression, à une dimension dendrogramme, qui identifie des groupes similaires, que ce soit au niveau des cellules ou des transcrits.

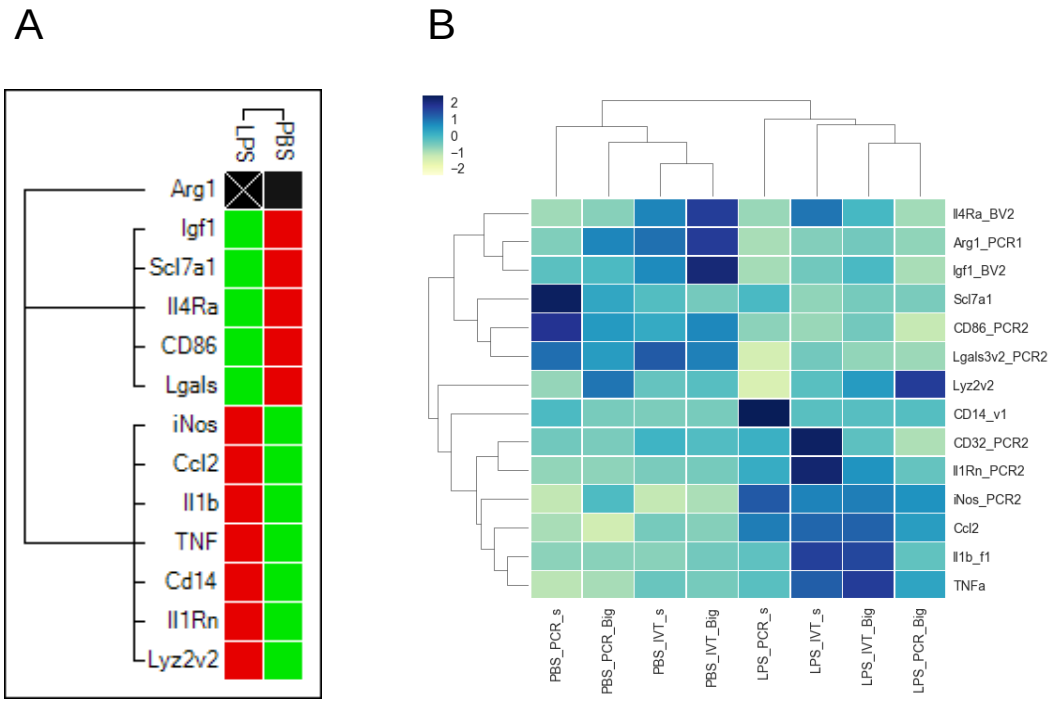


Figure 36 Analyse moyenne de l'expression relative de gènes cibles de l'inflammation afin de comparer une expérience en masse (A) aux expériences à l'échelle de la cellule unique (B) à partir des mêmes échantillons de cellules BV2 activées (LPS) ou non (PBS), A= Mesure quantitative de l'expression relative de cible par RT qPCR sur ARN extraits en amorces spécifiques, les résultats sont présentés sous forme de Clustermap généré grâce à l'outil CFX Manager après normalisation par rapport à l'expression d'un gène de référence (Rpl13a) dans chaque échantillon, B= Mesure de l'expression relative moyenne sur les 8 échantillons séquencés après normalisation du nombre d'UMI associés à chaque transcrit par rapport au nombre total d'UMI dans l'échantillon. Les résultats sont présentés sous forme de Clustermap, après homogénéisation du niveau de chaque transcrit, Légendes : PCR_s = Encapsulation de cellules BV2 dans des gouttes de 100pL et amplification par PCR, PCR_Big = Encapsulation de cellules BV2 dans des gouttes de 2nL et amplification par PCR, IVT_s = Encapsulation de cellules BV2 dans des gouttes de 100pL et amplification par IVT, IVT_Big = Encapsulation de cellules BV2 dans des gouttes de 2nL et amplification par IVT.

Le *clustemap* A (Figure 36) a été généré à partir des expériences de RT-qPCR en masse à l'aide du logiciel CFX Manager, après normalisation en fonction de l'expression d'un gène de référence Rpl13a dans chacun des échantillons (LPS et PBS). Une amorce spécifique de ce gène de référence a en effet été ajouté dans chacun des 2 tubes en plus des billes porteuses des amorces spécifiques de RT à codes-barres, autorisant une normalisation des échantillons entre eux, nécessaire pour pouvoir effectuer des analyses de l'expression génique relative à l'aide du logiciel CFX Manager. On observe que les cibles se divisent en 2 groupes principaux ; le premier est associé à une diminution de l'expression des gènes Arg1, Igf1, Il4Ra, Lgals3, Scl7a1 et Cd86 chez les cellules après activation à l'aide de LPS. Ces gènes

(hormis Cd86) sont associés à un phénotype de type M2 d'inflammation alternative ou à un état contrôle M0 (Tableau 3). Le second groupe est associé à une surexpression de marqueurs de l'inflammation M1, ou pro-inflammation, survenant au début d'une réaction inflammatoire lors d'une infection, à savoir ; iNos, Ccl2, Il1b, TNFa, Cd14 et Lyz2. On remarque que le gène Il1Rn, associé au phénotype M2, est lui aussi surexprimé par ces cellules ayant été mises en contact avec du LPS, signe que les 2 réponses inflammatoires sont initiées en réponse à un tel signal, avec néanmoins une réponse M1 plus marquée.

Le *clustermap* B (Figure 36) a été produit à partir des données de séquençage issues des 8 échantillons décrits dans la Figure 35, correspondant à l'encapsulation de cellules BV2 activées au LPS ou non activées (PBS) dans des gouttes de 100pL (s) ou de 2nL (Big), et dont le produit de RT en gouttes a été amplifié soit par PCR soit par IVT. Après démultiplexage des codes-barres, et identification des transcrits cibles et des UMI, puis filtration des codes-barres contaminants, le nombre d'UMI associés à chaque transcrit a été comptabilisé pour chaque échantillon. Une matrice a été générée, où chaque ligne correspond à un transcrit cible, chaque colonne à l'un des 8 échantillons et les valeurs dans la matrice à la proportion d'UMI attribuée à un transcrit donné dans l'échantillon concerné. Afin de faciliter la visualisation, cette matrice a été normalisée par ligne, de sorte à homogénéiser les niveaux d'expression entre les différentes cibles.

On observe que les échantillons se regroupent en fonction de l'état d'activation des cellules, avec un groupe correspondant aux échantillons où ont été encapsulées des cellules mises en contact avec du LPS et l'autre groupe à ceux où ont été encapsulées des cellules contrôles, mises en contact avec du PBS. La taille des gouttes dans lesquelles se fait la réaction de capture des ARNm sur les amorces à codes-barres puis leur rétrotranscription en ADNc, de même que le mode d'amplification de ces ADNc, par IVT ou PCR, n'a donc pas d'impact sur la réponse moyenne observée ; les différents échantillons à l'origine cellulaire commune s'organisent en un même groupe. Les transcrits cibles s'organisent également en 2 groupes principaux. Le premier groupe comprend essentiellement des marqueurs d'une inflammation de type alternatif (M2) ou d'un état contrôle (M0), à savoir Il4Ra, Arg1, Igf1, Lgals3 Scl7a1 et Cd86 (Tableau 3) et le second à des marqueurs de la pro-inflammation M1 à savoir Lyz2, Cd14, Cd32, iNos, Ccl2, Il1b, TNFa. On y retrouve également le marqueur d'une

inflammation de type M2, Il1Rn, de la même manière que ce qui a été observé dans l'expérience en masse. Les deux groupes sont donc les mêmes que ceux observés par l'analyse en groupe d'après les données en tube et font biologiquement sens.

Quelque-soit la taille des gouttes dans lesquels s'est fait la réaction de RT ou le mode d'amplification des ADNc lors de la préparation des librairies de séquençage, les analyses moyennées sont donc similaires entre elles ainsi qu'à celles faites à partir des échantillons en tube, et sont biologiquement cohérentes.

3. Comparaison de l'efficacité réactionnelle en fonction du volume des gouttes

L'influence du volume de la goutte dans laquelle se fait la réaction de capture des ARNm puis leur rétrotranscription en ADNc a ensuite été testée. L'impact de la taille et la capacité fonctionnelle des billes porteuses des amorces à code-barres sur l'efficacité réactionnelle a également été vérifié.

Les codes-barres ont été construits sur des billes PA de 10pL ou de 100pL en suivant le protocole de ligation par *split-and-pool* optimisé décrit au cours du chapitre précédent. Une amorce polyT a ensuite été liguée sur un extrait de chacun des 2 types de billes. Une sonde fluorescente FAM venant s'hybrider sur l'amorce polyT a permis d'évaluer la capacité fonctionnelle des billes en se basant sur une référence de billes commerciales, les billes 1CellBio, qui sont des billes PA de 120pL, porteuses de 10^9 amorces polyT à codes-barres (Zilionis *et al.*, 2017). Les résultats de cette analyse par microscopie à épifluorescence sont présentés dans la Figure 37.

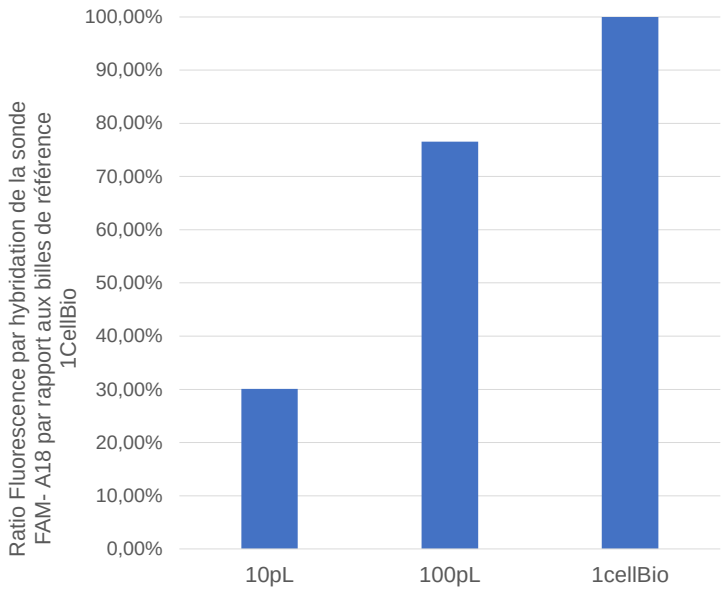


Figure 37 Mesure du nombre relatif d'amorces à codes-barres portées par des billes PA de 10pL ou 100pL par rapport à des billes commerciales de référence, dont la capacité fonctionnelle est de 10^9 amorces par bille (1CellBio, #20075)

Les billes de 100pL ont une capacité fonctionnelle plus grande que celles des billes de 10pL ; elles sont respectivement de l'ordre de 7.10^8 et 3.10^8 amorces par bille. Les

essais d'optimisation, présentés au cours du chapitre précédent, nous ont permis d'évaluer que le nombre optimal d'amorces par cible pour assurer une réaction de RT efficace était de l'ordre de 10^8. Étant donné que nos billes portent une proportion identique des 14 types d'amorces différentes pour nos gènes d'intérêt, nous sommes légèrement en dessous de ces capacités, mais tout de même dans le même ordre de grandeur. Comparées aux billes de 10pL, les billes de 100pL sont plus proches de l'optimum visé.

Les échantillons séquencés 1 et 5 ont été comparés (Figure 35). Ils correspondent respectivement aux cas où des cellules activées au LPS ont été encapsulées dans des gouttes de 100pL avec des billes de 10pL ou de 2nL avec des billes de 100pL. Ces 2 échantillons ont subi le même type d'amplification lors de l'étape de préparation des librairies au séquençage. Ils diffèrent donc juste au niveau du volume des gouttes produites et des billes qui ont été utilisées. Les caractéristiques de chaque échantillon, à savoir le nombre de cellules encapsulées, le nombre de codes-barres comptés avant et après application d'un filtre pour éliminer les codes-barres contaminants, ainsi que le nombre d'UMI par codes-barres sont présentés dans le Tableau 4.

Echantillon	**Nombre de *reads* alignés**	**Nombre de cellules attendues**	**Nombre de BC avant filtre**	**Nombre de BC après filtre**	**Nombre moyen d'UMI par BC**	**Nombre moyen de *reads* par UMI**
1 / petite goutte	13 710 464	15 000	121 268	13 604	70	14
5 / grosse goutte	23 908 159	2 000	16 513	2 122	265	42

Tableau 4 Caractéristiques numériques des échantillons 1 et 5 séquencés, correspondant à des cellules BV2 activées au LPS, encapsulées respectivement dans des gouttes de 100pL ou 2nL respectivement, Légende : BC= codes-barres, UMI = Identifiant Moléculaire Unique

En s'intéressant au nombre de cibles capturées par goutte, on constate que celui-ci est plus important dans le cas de gouttes plus volumineuses et pour lesquels la bille utilisée montre une plus grande capacité fonctionnelle. Ces résultats font sens avec les observations faites précédemment.

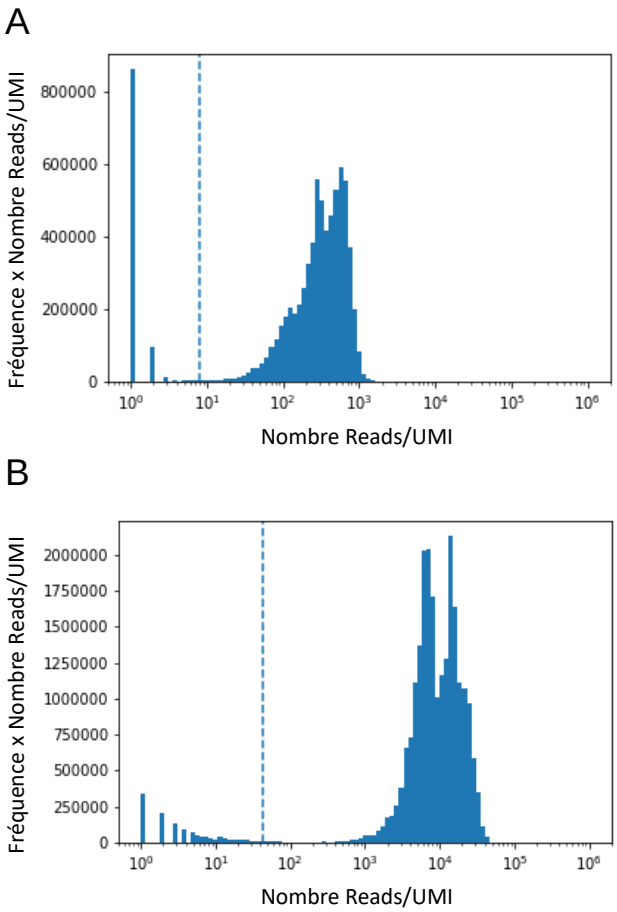


Figure 38 Histogramme de la fréquence du nombre de reads par UMI au sein de l'échantillon 1 (A) correspondant à l'encapsulation de cellules BV2 activées au LPS dans des gouttes de 100pL avec des billes de 10pL ou dans l'échantillon 5 (B), correspondant à l'encapsulation de cellules BV2 activées au LPS dans des gouttes de 2nL avec des billes de 100pL. Ces graphes ont été générés après identification des gènes, BC et UMI et filtration des BC contaminants comptant respectivement moins de 300 ou 100 reads pour les échantillons 1 et 5. Un poids a été appliqué en multipliant la fréquence d'une occurrence par l'occurrence elle-même pour faciliter la visualisation. La ligne en pointillé représente le nombre de reads par UMI moyen dans l'échantillon. Légende : UMI = Identifiant Moléculaire Unique

On constate que l'échantillon 5, correspondant aux cellules encapsulées dans des gouttes de 2nL, compte plus d'UMI par BC, signe d'une réaction de capture et RT en gouttes plus efficace. Cet échantillon compte moins de cellules ; cela est lié à la fréquence de production des gouttes, qui est 10 fois plus élevé dans le cas de l'échantillon 1. Il est donc techniquement difficile de collecter le même nombre de cellules dans les 2 cas, car cela prendrait trop de temps pour des grosses gouttes ou au contraire serait trop rapide pour un millier de cellules dans le cas de petites gouttes. De ce fait, et du fait que les échantillons aient été mélangés de manière équimolaire après indexation avec l'index illumina et avant séquençage, la profondeur de séquençage est plus importante dans le cas des grosses gouttes (Figure 38). Cela pourrait contribuer au fait que plus d'UMI soient détectés, mais en observant la répartition du nombre de *reads* par UMI (Figure 38) ainsi que le nombre moyen de *reads* par UMI, il semble que les 2 échantillons aient été séquencés avec suffisamment de profondeur.

Le nombre de molécules cibles capturées dans chaque cellule unique a ensuite été analysé en traçant la courbe du nombre total d'UMI capturés par cellule dans un échantillon de 2 000 cellules pour chacun des cas étudiés (Figure 39) ainsi que l'histogramme de la fréquence d'occurrence du nombre d'UMI par cellule pour chaque transcrit cible (Figure 40).

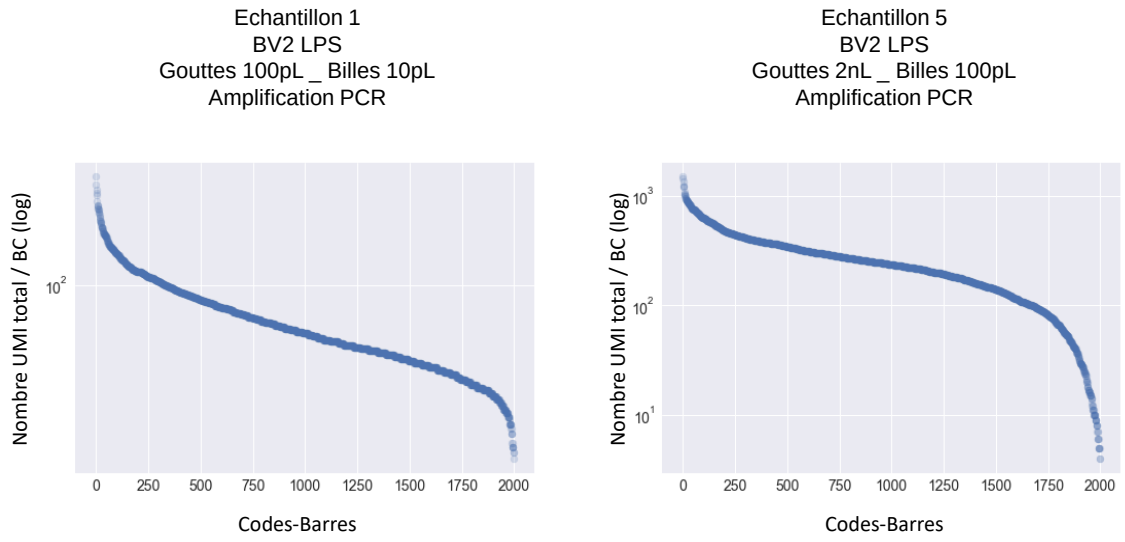


Figure 39 Quantification du nombre total d'UMI par BC sur un échantillon de 2000 cellules BV2 activées au LPS encapsulées dans des gouttes de 100pL (Gauche) ou de 2nL (Droite)

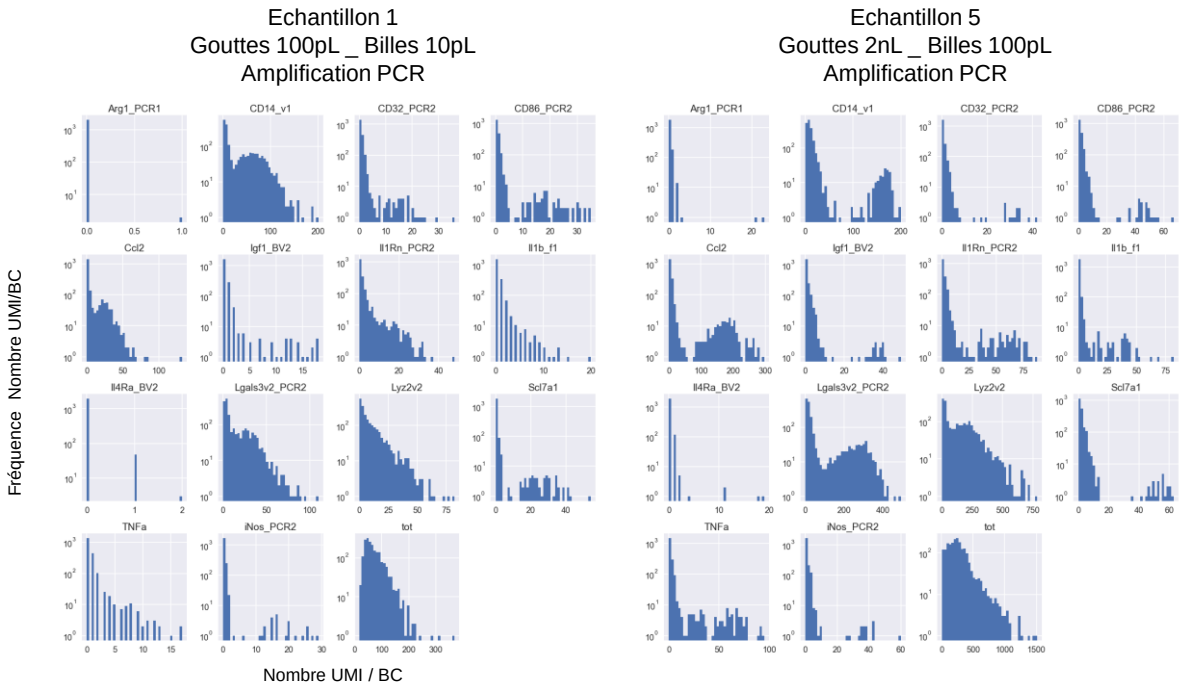


Figure 40 Comparaison du nombre d'ARN capturés par cellule (correspondant au nombre d'UMI) pour chaque transcrit cible entre l'échantillon issu de l'encapsulation de cellules BV2 activées avec du LPS dans des gouttes de 100pL (Echantillon 1) ou de 2nL (Echantillon 5). Ces graphes ont été générés à partir d'un échantillon de 2 000 codes-barres (cellules) dans chaque cas. Légende : UMI= Identifiant Moléculaire Unique

D’après la Figure 39, on constate qu’il y a presque un ordre de grandeur de différence dans le nombre d’UMI total quantifiés entre les 2 échantillons. La courbe est plus plate dans le cas de l’échantillon 5, montrant une capture plus homogène d’une goutte à l’autre, signe probable d’une meilleure efficacité réactionnelle et moins aléatoire entre gouttes. La dernière fenêtre de chaque graphe multiple de la Figure 40, et sous-titrée « tot », correspond à la fréquence du nombre d’UMI totaux quantifiés par cellule et est une combinaison des 14 autres graphes précédents. On remarque que jusqu’à 1000 UMI par cellule sont détectés dans le cas des gouttes de 2nL (Echantillon 5) contre 200 pour des gouttes de 100pL (Echantillon 1), indiquant une fois encore que la capture et la RT des ARNm cibles semble plus efficace dans des gouttes plus volumineuses permettant d’apporter des billes à la capacité fonctionnelle plus grande.

Ce constat peut également se faire pour chaque transcrit, un à un. On peut citer Ccl2, dont on compte jusqu’à 300 UMI par gouttes dans l’échantillon 5 contre 60 dans l’échantillon 1. De même, la répartition du nombre d’UMI capturés par cellule pour chaque cible est plutôt continue dans l’échantillon 1 et au contraire plutôt bimodale dans l’échantillon 5. On peut supposer que les molécules sont plus efficacement capturées dans l’échantillon 5, conduisant à un nombre d’UMI par transcrit quantifiés maximal dans chaque goutte si ce dernier est bien exprimé par la cellule. En revanche, le gène peut ne pas être exprimé dans certaines cellules, correspondant à la population vide la plus proche de l’axe dans les différents sous-graphes. Si la capture est dépendante d’un biais technique, on s’attend à observer une répartition plus continue du nombre d’UMI par cellule, car étant un phénomène à l’efficacité non optimal et donc plus aléatoire.

Finalement, des sous-groupes de cellules ont été recherchés à partir des données de scRNAseq ciblé sur cellules BV2 activées au LPS, par encapsulation dans des gouttes de 100pL ou de 2nL. Nous souhaitons ainsi à déterminer si, lors d'une réaction inflammatoire, les cellules microgliales expriment dans un premier temps un phénotype M1 puis évoluent vers un phénotype M2, impliquant l'existence d'état transitoire, ou si au contraire, des populations différentes entrent en jeu à différents moments.

Pour cela, différents graphes ont été réalisés, à partir de la matrice d'expression des cibles d'intérêt dans chaque cellule, elle-même générée après identification des codes-barres, gènes et UMI, de la filtration des codes-barres contaminants d'après un seuil défini graphiquement, et regroupement des *reads* par codes-barres, gènes et finalement, normalisation grâce aux UMI. Les résultats sont présentés en Figure 41 et en Figure 42.

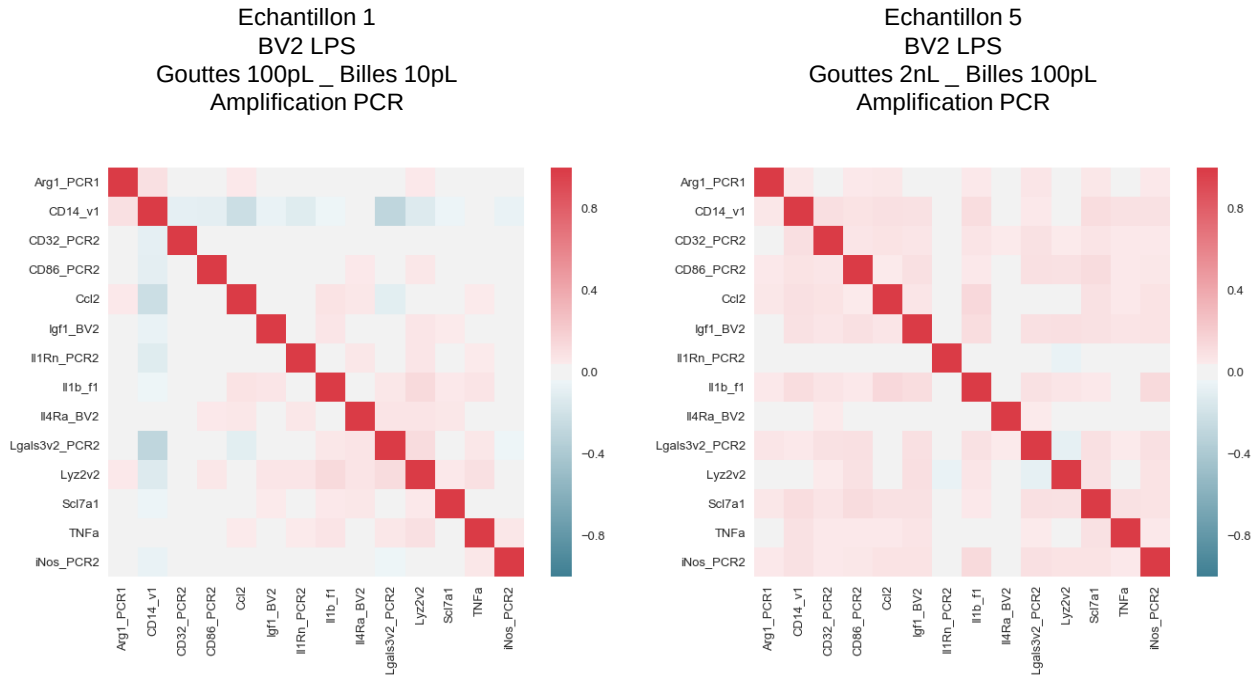


Figure 41 Matrice de corrélation de l'expression de gènes cibles au sein d'une même goutte de 100pL (Gauche) ou de 2nL (Droite)

La Figure 41 montre les matrices de corrélation générées à partir des matrices d'expression des échantillons 1 et 5, où la corrélation entre transcrit est calculée, et donne la probabilité de co-expression de transcrits au sein d'une même cellule. Ainsi, si des transcrits sont préférentiellement exprimés ensemble dans plusieurs cellules, un motif de corrélation apparaît. On constate cependant ici l'absence de motif de corrélation, et ce pour les 2 échantillons.

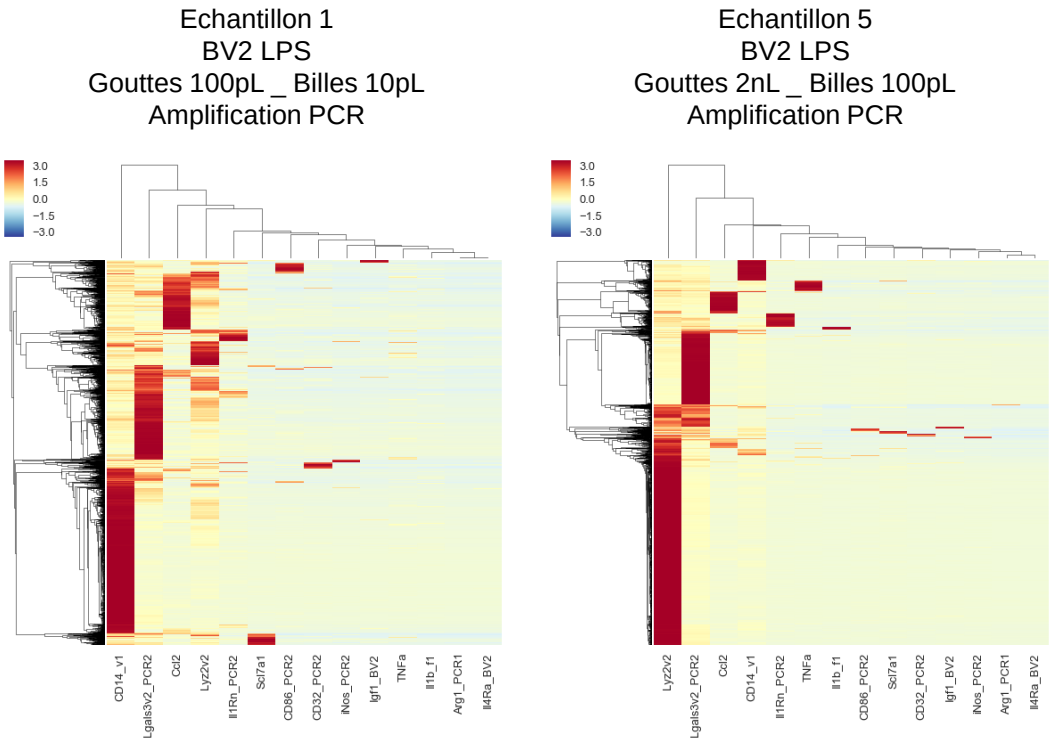


Figure 42 Clustermap représentant l'expression relative de transcrits cibles dans chaque cellule unique sous forme de heatmap ainsi que les sous-groupes présentant des traits communs sous forme de dendrogramme, dans des cellules BV2 activées au LPS, et encapsulées dans des gouttes de 100pL (Gauche) ou de 2nL (Droite)

La Figure 42 montre une représentation de la matrice d'expression sous forme de *clustermap*, qui associe une dimension *heatmap*, où les niveaux d'expression sont visualisés en gradient de couleur, à une dimension dendrogramme, permettant d'identifier des sous-groupes en fonction des similitudes. La matrice a été filtrée pour éliminer les codes-barres comptant moins de 10 UMI totaux avant de générer ce graphe. Une normalisation par ligne, c'est-à-dire par cellule, pour homogénéiser le niveau d'expression total des cellules entre elles, est également appliquée. Tout d'abord, la comparaison des 2 *clustermaps* montre que certains transcrits ne sont pas détectés dans les petites gouttes, alors qu'ils le sont, bien que faiblement, c'est-à-dire dans un petit nombre de cellules, dans les grosses gouttes, il s'agit de TNFa ou encore d'Il1b. Cette observation va dans le sens des conclusions faites précédemment, montrant que la réaction en gouttes est plus efficace dans des gouttes de 2nL et avec des billes de 100pL à la capacité fonctionnelle plus grande.

On constate dans les 2 cas qu'il est très difficile de définir des sous-groupes, que ce soit au niveau des cibles ou des cellules. Le *heatmap* montre d'ailleurs qu'il est assez peu fréquent de détecter plusieurs cibles au sein d'une même cellule. On observe des blocs, correspondant à la surexpression d'un transcrit donné. Ces blocs sont plus ou moins importants en fonction de la cible, donnant l'impression que les cellules ne peuvent exprimer qu'une seule cible à la fois, ou que le système n'est capable de capturer qu'une cible à la fois.

On peut supposer d'une part qu'il s'agit d'un biais technique. La capture des ARNm, n'étant pas efficace à 100%, fait que certains transcrits, bien que présents, ne sont pas détectés. Ce phénomène du « faux zéro », décrit dans la littérature au sujet des technologies scRNAseq total, est d'autant plus présent que le niveau d'expression d'une cible est faible. Considérons une efficacité de capture de l'ordre de 10%. Si un gène est faiblement exprimé, de l'ordre de 10 transcrits produits, il est plus probable de ne capturer aucun ARNm dans ce cas que dans celui d'un gène exprimant des milliers d'ARNm. Les blocs les plus grands correspondent d'ailleurs aux cibles exprimant le plus de molécules de transcrits, d'après les histogrammes traçant la fréquence du nombre d'UMI par cellule dans chaque transcrit (Figure 40).

Il convient également de considérer l'expression séquentielle des cibles d'intérêt. Ces gènes sont exprimés en réponse à un stimulus, et sont donc soumis à une expression limitée dans le temps, de même qu'à une possible dégradation rapide. L'image transcriptionnelle de chaque cellule est prise à un instant donné. Il est possible qu'à cet instant, tous les transcrits ne soient réellement pas tous présents.

Finalement, aussi bien une réalité biologique qu'un biais technique peut expliquer le phénomène observé. Une expérience en ARN synthétique, du type ERCC, permettrait d'évaluer le biais technique associé à notre technologie, l'efficacité de capture et la limite de détection. De plus, cibler plus de marqueurs d'intérêt pour chaque phénotype visé (M1 et M2 dans notre cas) pourrait faciliter la définition de sous-groupes, afin d'augmenter la probabilité de co-expression entre cibles.

4. Amplification des ADNc

Une autre étape clé des protocoles de scRNAseq est l'étape de préparation des librairies, avec amplification des ADNc à codes-barres, produits de la RT en gouttes. Cette amplification peut se faire de manière linéaire, permettant de limiter les biais d'amplification, ou de manière exponentielle, avec une grande sensibilité. Nous avons testé ces 2 méthodes en masse. Les résultats sont présentés au cours du chapitre précédent. Nous allons à présent les comparer dans des expériences à l'échelle de la cellule unique. Avant d'être cassée, chaque émulsion a été divisée en 2 parts égales (Figure 35).

L'une d'elle a suivi une préparation avec amplification par IVT et l'autre une amplification par PCR, comme présenté dans la Figure 43.

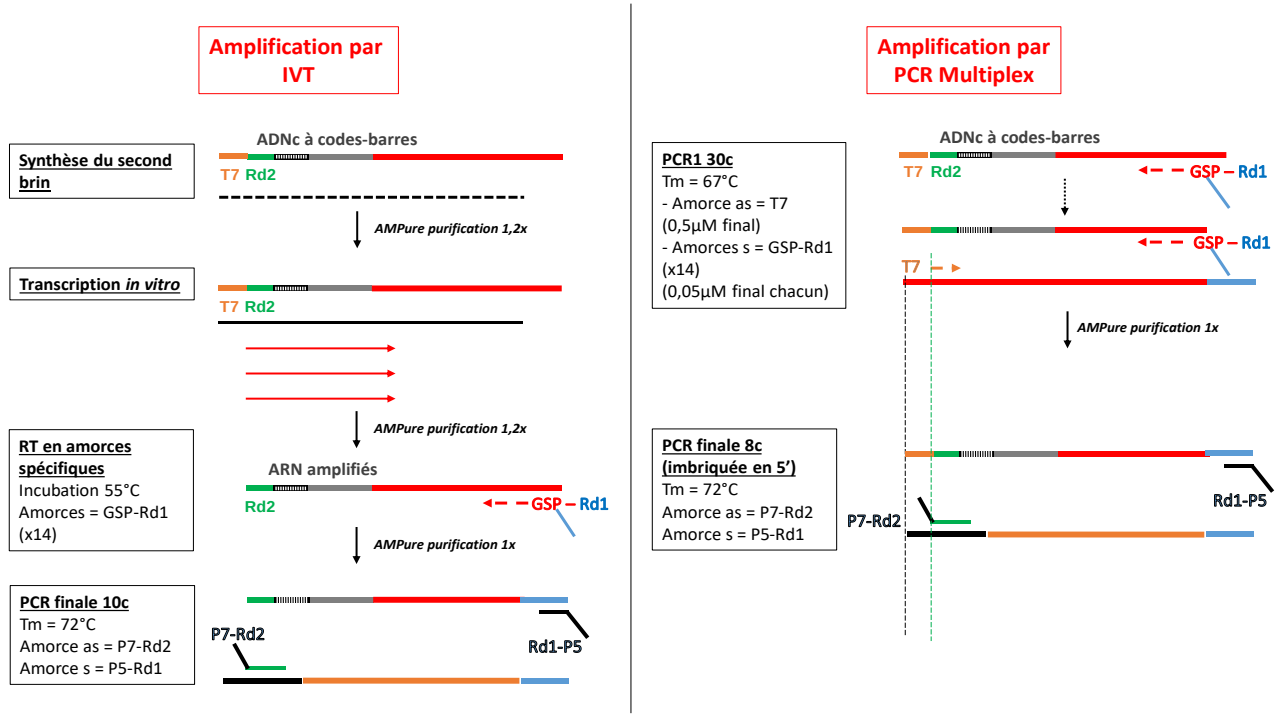


Figure 43 Schéma des 2 protocoles d'amplification comparés, Gauche = Amplification linéaire par Transcription in vitro, Droite = Amplification exponentielle par PCR

Après alignement contre le génome de référence contenant les 14 séquences d'intérêt et contre les 96 séquences d'index B, C et D, les codes-barres et transcrits de chaque *read* ont pu être identifiés. L'UMI quant à lui, est une séquence de 5bp, situé en aval d'une séquence consensus connue. La séquence de l'UMI est extraite en utilisant cette séquence consensus pour définir la position de départ de l'UMI. Aucun filtre n'est appliqué sur les UMI dans le script employé. Le taux d'alignement a été calculé à chacune de ces étapes d'identification, en comparant le nombre de *reads* pour lesquels successivement un transcrit cible, un code-barres et un UMI ont été correctement identifiés, par rapport au nombre de *reads* de départ dans chacun des 8 échantillons séquencés. Les résultats sont présentés dans la Figure 44.

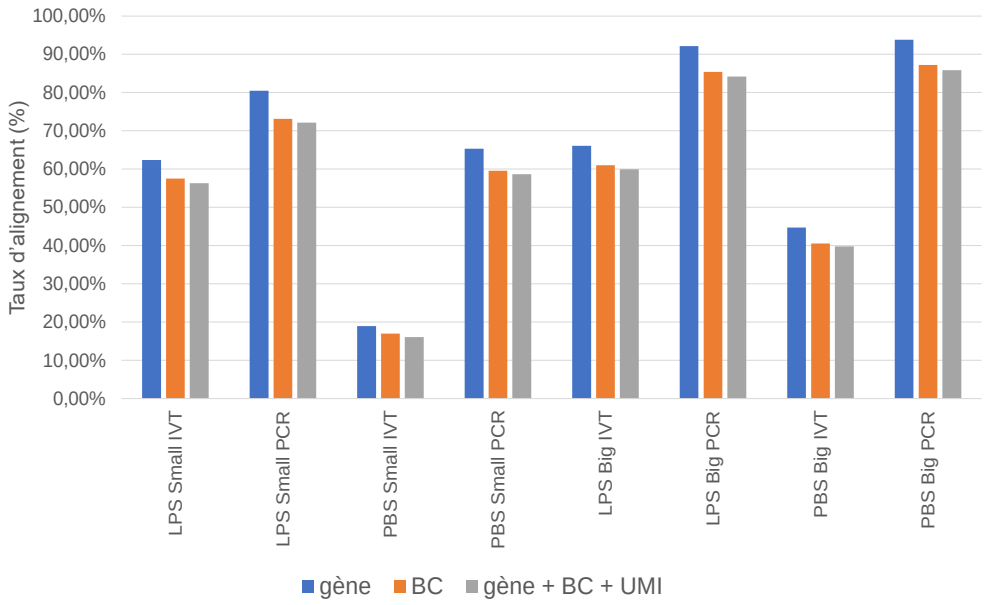
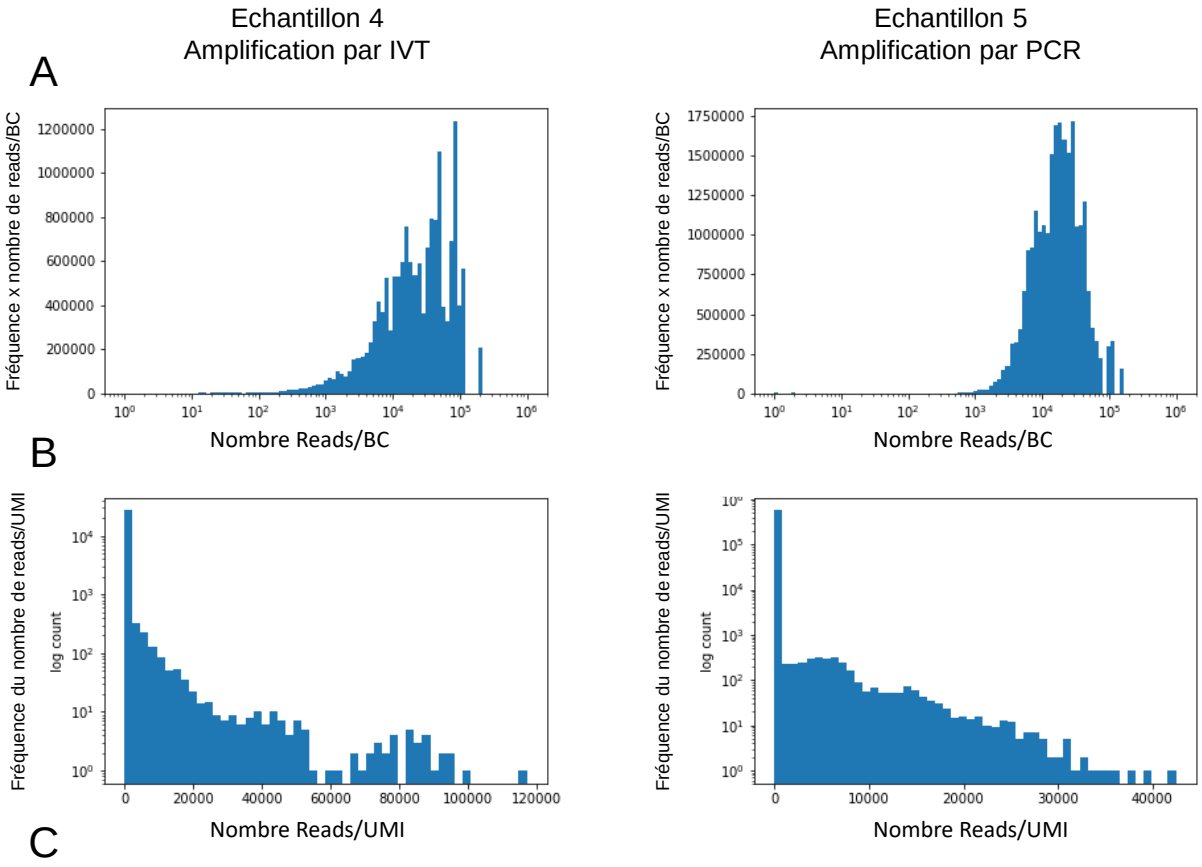


Figure 44 Taux d'alignement des 8 échantillons séquencés, correspondant à l'encapsulation de cellules BV2 activées (LPS) ou non (PBS) dans des gouttes de 2nL avec des billes de 100pL (Big) ou dans des gouttes de 100pL avec des billes de 10pL (Small) et ayant suivis une stratégie d'amplification linéaire (IVT) ou exponentielle (PCR). Ce taux a été mesuré (1) à l'étape d'identification des transcrits d'intérêt contre un génome de référence personnalisé à l'aide de l'outil bowtie2 (gène), (2) à l'étape d'identification des codes-barres composés de 3 indexes existant chacun en 96 versions à l'aide de l'outil bowtie2 (BC), et enfin (3) lors de l'extraction des UMI, une séquence de 5bp, par reconnaissance d'une séquence consensus située après le code-barres et en amont de l'UMI (gène + BC + UMI), Légendes : UMI = Identifiant Moléculaire Unique, BC = codes-barres

On constate que le taux d'alignement est globalement meilleur dans les échantillons correspondants à l'encapsulation des cellules activées au LPS par rapport à ceux où les cellules n'étaient pas activées. Ce taux est également meilleur dans les échantillons ayant suivi une préparation avec amplification par PCR par rapport à ceux où l'amplification se fait par IVT. Les échantillons amplifiés par IVT ont également été alignés contre le génome entier, pour vérifier si cette différence est due au fait que d'autres amplicons sont amplifiés du fait d'une spécificité moindre de la réaction de RT, qui se fait à 55°C. Les taux obtenus sont légèrement plus grands, de l'ordre de 5% de plus (données non présentées) que ceux observés ici. On sait que la réaction d'IVT nécessite un minimum de matériel de départ pour être efficace. Cela pourrait expliquer pourquoi l'amplification par IVT est moins efficace que celle par PCR, qui est connue pour être très sensible. D'ailleurs, les échantillons où des cellules PBS ont été encapsulées comptent bien moins de transcrits, étant donné que les cibles sont des marqueurs de l'inflammation dont l'expression est induite ou renforcée en réponse à un stimulus. Ces échantillons montrent un taux d'alignement particulièrement bas, probablement car le nombre d'ADNc néosynthétisés total, après cassage des émulsions, est trop faible. Enfin, le taux est légèrement plus grand dans les échantillons où les cellules ont été encapsulées dans des gouttes de 2nL. Finalement, l'échantillon montrant le plus faible taux d'alignement est l'échantillon 3, où des cellules non activées ont été encapsulées dans des gouttes de 100pL et dont le produit de RT a été amplifié par IVT ; ce taux est d'ailleurs particulièrement bas, de moins de 20% à l'issu des 3 étapes d'identification.

Nous nous intéressons dans un premier temps à l'échantillon E2 (Figure 35), correspondant au cas où des cellules activées au LPS ont été encapsulées dans des gouttes de 2nL avec des billes de 100pL.

La Figure 45 présente la fréquence du nombre de *reads* par code-barres (A) et du nombre de *reads* par UMI sous forme d'histogrammes, pour les 2 échantillons issus de l'émulsion E2 et ayant suivi une stratégie d'amplification par IVT (échantillon 4) ou par PCR (échantillon 5).



Echantillon / Amplification	Filtre *(sur le nombre de reads par BC)*	Nombre BC *(avant filtre)*	Nombre BC *(après filtre)*	Nombre total d'UMI *(après filtre)*	Nombre moyen de reads par UMI *(après filtre)*
4 / IVT	100	6652	2145	28 009	548
5 / PCR	100	16513	2122	564 380	43

Figure 45 Application de filtres pour éliminer les codes-barres contaminants sur les échantillons de cellules BV2 activées encapsulées dans des gouttes de 2nL, et ayant suivi une amplification par IVT (Gauche, échantillon 4) ou par PCR (Droite, Echantillon 5) et comparaison du nombre de BC, du nombre d'UMI et de la profondeur de séquençage, A = Histogramme de la fréquence du nombre de reads par BC après application d'un poids pour rendre le graphe plus facilement interprétable. Ce graphe sert de référence pour appliquer un filtre permettant d'éliminer les faux BC, B = Histogramme de la fréquence du nombre de reads par UMI, C = Tableau des données numériques pour chaque échantillon, à savoir le nombre de BC comptabilisés avant et après application du filtre sur le nombre de reads par BC, du nombre total d'UMI et de la profondeur moyenne (nombre moyen de reads par UMI), Légendes : BC = codes-barres, UMI = Identifiant Moléculaire Unique

Après élimination des codes-barres contaminants, correspondant à ceux comptant moins de 100 *reads*, seuil facilement identifiable grâce aux histogrammes A, les 2 échantillons comptent, comme attendu, environ 2 000 cellules chacun, signe que l'émulsion a bien été divisée en 2 parts égales. On peut noter que le nombre de codes-barres total avant application du filtre est environ 3 fois plus grand pour l'échantillon à l'amplification par PCR, montrant que le grand nombre de cycles d'amplification a été la source de plus d'erreurs.

Le nombre d'UMI total est en revanche différent d'un échantillon à l'autre. Il est environ 20 fois plus grand pour l'échantillon 5, issu de l'amplification par PCR, alors même que le nombre d'ADNc synthétisés dans les 2 échantillons, étant le fruit d'une seule et même émulsion divisée en 2, est censé être sensiblement identique. De ce fait, la profondeur de séquençage, définie par le nombre de *reads* par UMI, est plus grande pour l'échantillon 4, car autant de *reads* se répartissent en moins d'UMI.

Deux hypothèses peuvent être émises ;

- soit l'amplification par IVT, étant moins sensible et spécifique, conduit à un taux d'alignement de 20% inférieur à l'échantillon amplifié par PCR, conduisant à une perte partielle de signal
- soit l'amplification par PCR, comptant 38 cycles d'amplifications, a entraîné l'apparition d'erreurs dans les séquences et notamment au niveau des UMI, conduisant à l'apparition de faux UMI.

L'histogramme de la fréquence du nombre de *reads* par UMI de l'échantillon 5 (à amplification PCR), montre d'ailleurs qu'une grande majorité des UMI comptent très peu de *reads*. C'est le cas aussi pour l'échantillon 4 amplifié par IVT, mais avec deux ordres grandeurs de moins d'UMI pauvres en *reads*.

La Figure 46 présente la fréquence du nombre d'UMI par goutte pour chaque cible dans les 2 échantillons comparés.

Un filtre sur les UMI en fonction du nombre de *reads* qu'ils comptent a été appliqué sur l'échantillon amplifié par PCR afin de tester l'hypothèse de la présence de faux UMI. Le graphe du milieu correspond à l'histogramme de la fréquence du nombre d'UMI par cellule dans chaque transcrit sans appliqué de filtre sur les UMI. L'histogramme de droite correspond au cas où les données de l'échantillon 5 ont été filtrées ; les UMI comptant moins de 10 *reads* ont été éliminés.

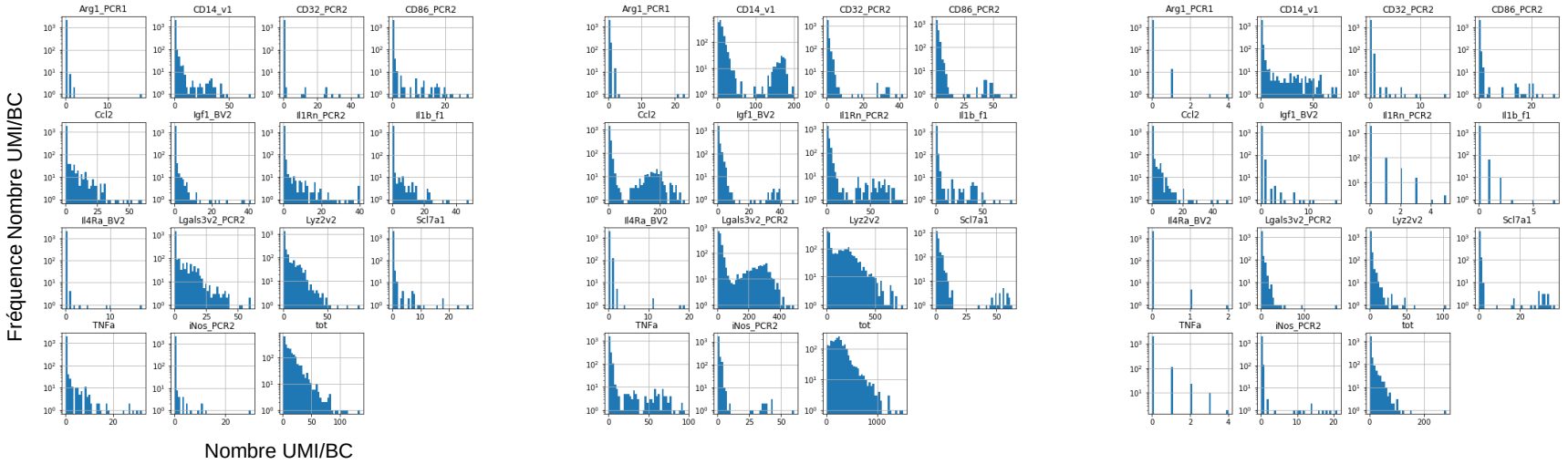
Le filtre sur les *reads* par UMI appliqué à l'échantillon ayant suivi une amplification par PCR permet de réduire le nombre d'UMI total dans l'échantillon 5 à un nombre comparable à celui de l'échantillon 4. Cependant, l'application de ce filtre entraîne une perte de signal pour certains transcrits, tels que Il1Rn, Il1b ou TNFa.

Une amplification par PCR en multiplex implique un biais lié notamment à l'efficacité variable des amorces utilisées. Ce filtre quelque peu arbitraire ne tient pas compte de ce paramètre. Ainsi, des cibles aux amorces moins efficaces, ayant été amplifiées plus tardivement au cours des 30 cycles d'amplification, et comptant de ce fait moins de molécules amplifiées, risquent d'être préférentiellement éliminées.

D'autres types de méthodes existent pour éliminer les signaux issus d'erreurs de séquençage ou d'introduction de mutations lors des étapes de PCR. Il est par exemple possible de calculer la distance de Hamming entre les différents UMI que comptent des *reads* d'amplicons potentiellement issus d'une même molécule d'ARN (un code-barres donné, une cible donnée) afin de regrouper ceux dont la distance est faible et pouvant avoir été créés via l'introduction d'une mutation ponctuelle (Smith, Heger, & Sudbery, 2017).

Echantillon 4
Cellules BV2 LPS
Gouttes 2nL – billes 100pL
Amplification IVT

Echantillon 5
Cellules BV2 LPS
Gouttes 2nL – billes 100pL
Amplification PCR



Filtre BC
2,145 BC
28,009 umi

Filtre BC
2,122 BC
564,380 umi

Filtre BC + reads/UMI
2,119 BC
16,150 umi

Figure 46 Comparaison de la fréquence du nombre d'UMI par BC quantifié pour chaque transcrit dans 2 échantillons ayant été amplifiés selon une stratégie par IVT (échantillon 4, graphe de gauche) ou par PCR (échantillon 5, graphes de droite). La fréquence de chaque occurrence, à savoir un certain nombre d'UMI par BC associé à un transcrit donné, est représentée sous forme d'histogramme. On en compte un par transcrit ainsi qu'un pour l'ensemble des transcrits. Ces graphes ont été générés après alignement, filtration des BC contaminants comptant moins de 100 reads et normalisation des UMI. L'échantillon 5, amplifié par PCR, a également été filtré (graphe à l'extrémité gauche) ou non (graphe du milieu) pour éliminer les UMI contaminants, comptant moins de 10 reads

On constate, comme évoqué dans le paragraphe précédent, que le nombre total d'UMI est un ordre de grandeur plus élevé dans l'échantillon 5, amplifié par PCR. En s'intéressant plus particulièrement aux histogrammes de chaque cible, on remarque que cette augmentation s'applique plus particulièrement pour les transcrits fortement exprimés, à savoir Cd14, Ccl2, Lgals3 ou Lyz2 mais que les ordres de grandeurs sont assez similaires d'un échantillon à l'autre pour les autres transcrits. Les histogrammes générés à partir des données de l'échantillon 5 amplifié par PCR montrent une distribution bimodale pour la plupart des transcrits, signe d'une surexpression significative de ces cibles dans une partie des cellules. Cette distribution en 2 populations est bien moins nette pour l'échantillon 4, amplifié par IVT. L'amplification par PCR permet d'obtenir des signaux significativement différents entre des sous-populations à l'expression variable.

L'analyse de la répartition des *reads* ou des UMI entre les différents transcrits dans les échantillons 4 et 5, amplifiés respectivement par IVT ou PCR, est présentée dans la Figure 47.

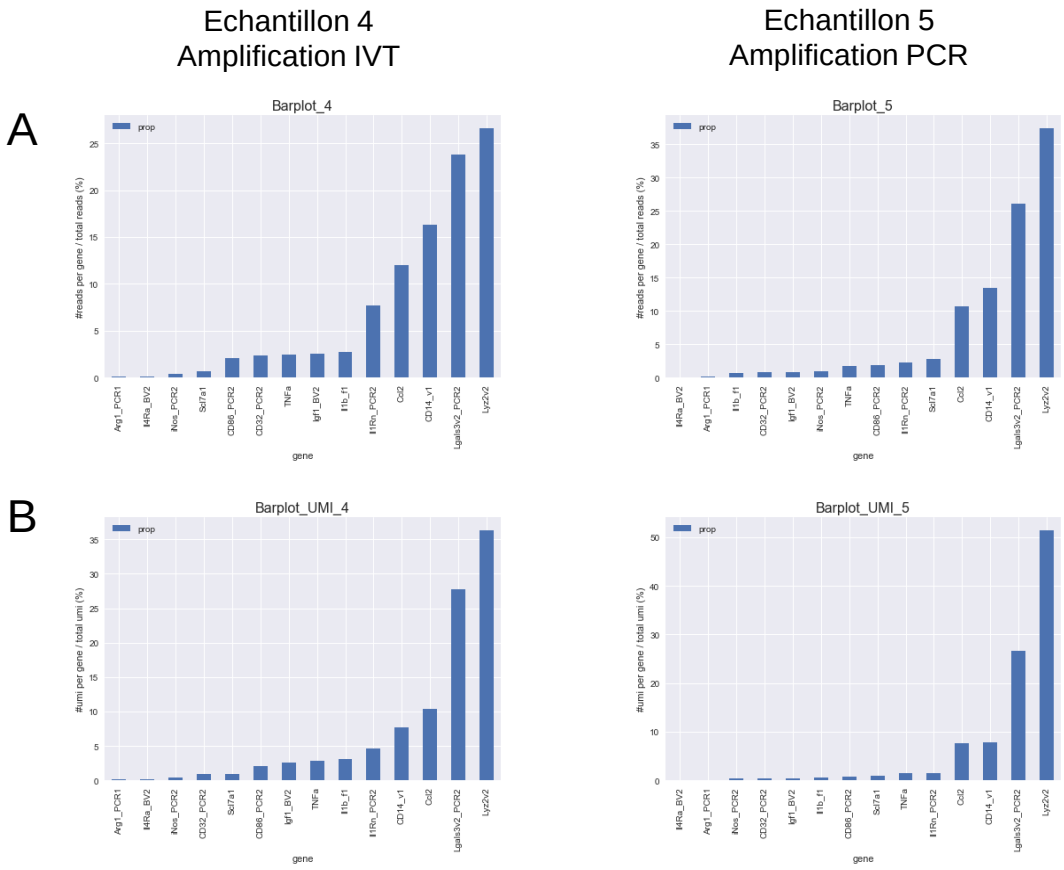


Figure 47 Distribution des reads et UMI attribués aux différents transcrits dans 2 échantillons séquencés, issus de l'encapsulation de cellules BV2 activées au LPS dans des gouttes de 2nL avec des billes de 100pL porteuses d'amorces à codes-barres spécifiques de 14 cibles d'intérêt. Après RT, l'émulsion a été divisée en 2 pour suivre une amplification par IVT d'une part (échantillon 4) ou par PCR d'autre part (échantillon 5).Les échantillons ont été séquencés et alignés contre des génomes de référence pour identification des cibles, des BC et des UMI. Les BC contaminants ont été filtrés par élimination des BC comptant moins de 100 reads, A = Histogramme en barres représentant la proportion de reads associés à chaque transcrit, B = Histogramme en barres représentant la proportion d'UMI associés à chaque transcrit, Légendes : UMI = Identifiant Moléculaire Unique, BC = code-barres

On observe une distribution plus homogène, aussi bien des *reads* que des UMI dans l'échantillon 4 que dans l'échantillon 5. Ce résultat fait sens avec les observations faites lors des expériences d'optimisation en RT-qPCR, qui montraient que l'expression relative des transcrits après une amplification par IVT était fidèle à celle observée par RT-qPCR sans étape d'amplification préalable. L'amplification linéaire par IVT limite donc la formation de biais en faveur d'une espèce dont l'amplification est plus efficace.

L'amplification par PCR est cependant plus sensible et malgré les biais éventuels permet de récupérer plus de signal par goutte, à un niveau significatif.

C. Comparaison de l'expression de transcrits cibles chez des cellules BV2 en état d'inflammation ou non, en utilisant un protocole de RNAseq ciblé en gouttes optimisé

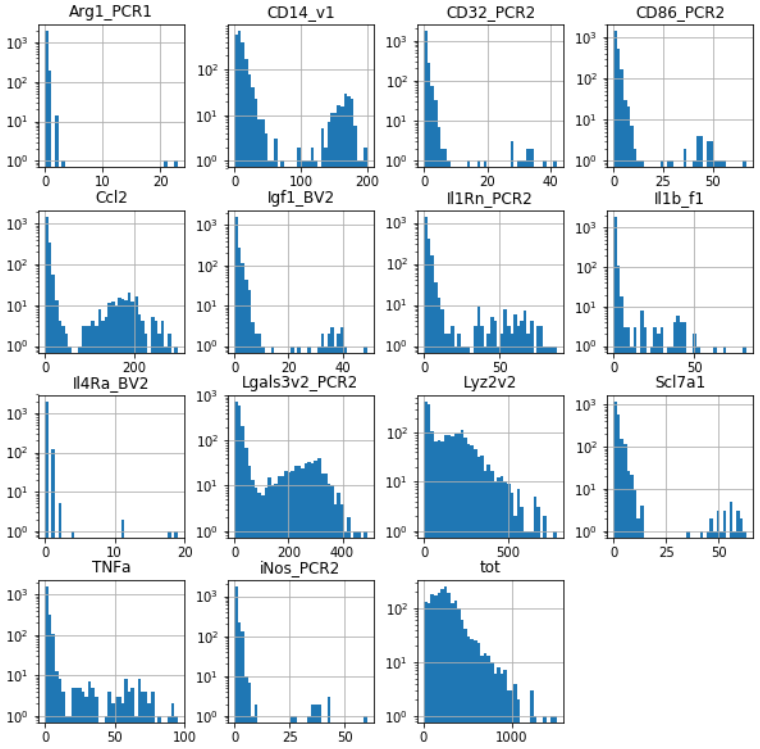
Le protocole optimal choisi au regard des résultats précédemment décrits est celui où des cellules sont encapsulées dans des gouttes de 2nL avec des billes de 100pL à amorces spécifiques à codes-barres, dont la capacité fonctionnelle est comparable à celle de billes commerciales 1CellBio, puis dont le produit de RT est amplifié selon une stratégie de PCR multiplex. Nous allons comparer les données issues du RNAseq ciblé en gouttes selon ce protocole appliqué à des cellules BV2 activées au LPS ou contrôle.

Les résultats précédents ont montré que la plupart des transcrits semble avoir une expression séquentielle, rendant une analyse par groupe délicate. L'interprétation est un réel défi pour discriminer une réalité biologique d'un biais technique. Les outils d'analyse par groupe du type Analyse en Composante Principale ou encore *Clustermap* ne sont donc pas adaptés à cette étude.

Nous pouvons cependant procéder à une analyse très fine de chaque transcrit un à un à l'échelle de la cellule unique, en observant la distribution des cellules en fonction du niveau d'expression d'une cible.

La Figure 48 présente la fréquence de cellules associées à un niveau d'expression donné à l'échelle de la cellule unique, définie par un nombre d'UMI par cellule, pour chaque cible d'intérêt, dans l'échantillon correspondant à l'encapsulation de cellules BV2 activées au LPS ou contrôle (PBS).

Echantillon 5
Gouttes 2nL – Billes 100pL
Amplification PCR
Cellules BV2 activées au LPS



Echantillon 7
Gouttes 2nL – Billes 100pL
Amplification PCR
Cellules BV2 non activées (PBS)

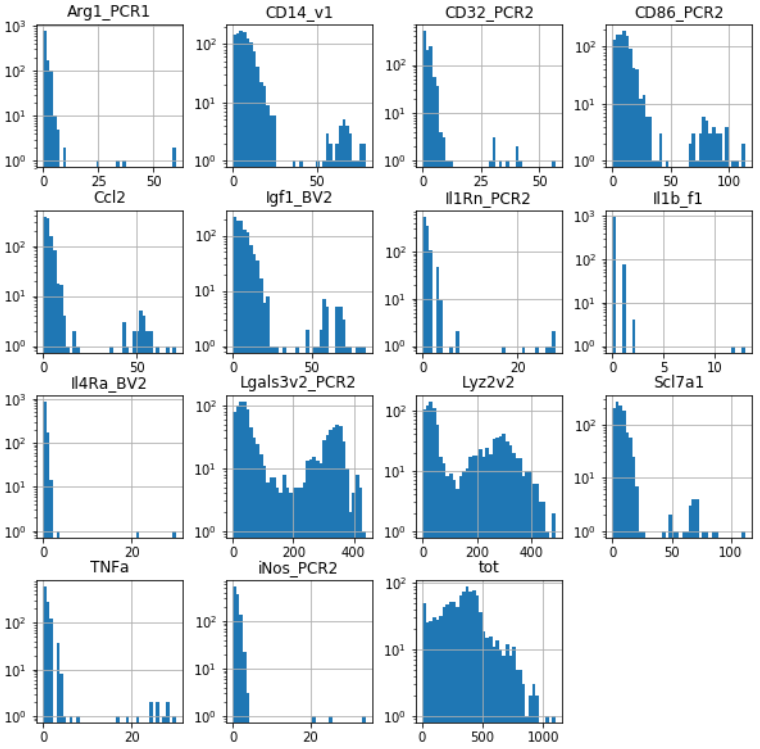


Figure 48 Comparaison de l'expression des différentes cibles à l'échelle de la cellule unique entre un échantillon de RNAseq ciblé en gouttes de cellules activées au LPS (Echantillon 5) ou de cellules contrôle (PBS). Les cellules ont été encapsulées dans des gouttes de 2nL avec des billes à amorces à codes-barres de 100pL et ont suivi une stratégie d'amplification par PCR. Chaque sous-figure représente la fréquence de cellules comptant un certain nombre d'UMI d'une cible donnée sous forme d'histogramme

Ainsi, l'expression du gène Cd14 dans des cellules non activées est très faible voire nulle, et seule une petite proportion des cellules montre une expression de l'ordre de 50 transcrits par goutte. Lors d'une activation, la proportion de cellules montrant une expression significative de ce gène est non seulement plus importante mais cette population est également décalée, avec un niveau d'expression de l'ordre de 100 à 200 transcrits par cellule. Plus de cellules sont donc recrutées et elles expriment d'autant plus de récepteurs Cd14, fixant justement le LPS, afin de permettre aux cellules de répondre plus efficacement à une infection bactérienne, ici simulée, créant une boucle de régulation positive. Le même genre de constat est fait pour les gènes Cd86, un autre récepteur à la surface des cellules microgliales, ou encore Ccl2, chimiokines induisant le recrutement des cellules microgliales au site de sécrétion.

Les gènes iNos, TNFa ou encore Il1b ne sont pas exprimés sans induction au LPS. Ces gènes sont des marqueurs typiques de la pro-inflammation et sont donc exprimés en fonction des besoins. Il est intéressant de noter que le gène Il1Rn est également exprimé en réponse à une induction au LPS, à un niveau de l'ordre de 50 transcrits par cellules, soit légèrement supérieur à celui de l'expression du gène Il1b, dont il inhibe pourtant l'effet. On fait ici face à une boucle de régulation négative, permettant de limiter les effets pro-inflammateurs, en préparation d'une réponse anti-inflammatoire.

Un marqueur typique de l'inflammation alternative qu'est Arg1 ne montre pas ou très peu d'expression en réponse à la présence de LPS. L'inflammation alternative semble encore être très limitée. Le type d'induction employé entraîne une réponse majoritairement pro-inflammatoire.

D'autres cibles, du type Lgals3 ou Lyz2 montrent une expression déjà importante même sans induction, et codent pour des protéines aux fonctions respectivement bactériolytique ou chimio-attractive et d'apoptose. Il est à noter que les fonctions de la protéine galectine 3 sont très variées. Ces protéines sont donc naturellement présentes dans près de la moitié des cellules microgliales et du fait de leur rôle important dans le processus pro-inflammatoire, elles sont exprimées par une proportion encore plus importante de cellules suivant l'activation au LPS.

III. Etude temporelle de l'inflammation chez la microglie

A. RNAseq sur transcriptome entier

Dans un premier temps, nous avons fait une étude temporelle de l'inflammation sur le transcriptome entier en masse, afin de définir des marqueurs temporels de l'évolution transcriptomique au cours du processus inflammatoire. Trois temps seront ainsi choisis pour notre étude de RNAseq ciblé à l'échelle de la cellule unique en gouttes.

Des cellules BV2 ont été déposées au fond de 5 puits de deux plaques 6 puits, et du LPS a été ajouté dans la première plaque pour induire une inflammation. Du PBS a été ajouté dans l'autre moitié, en contrôle. Le processus d'inflammation est stoppé à 2 heures, 4 heures, 8 heures, 16 heures ou 24 heures, en récupérant les cellules d'un puits dans chacune des 2 plaques à ces différents temps, puis en procédant à l'extraction de l'ARN total.

Un RNAseq total en tube est réalisé sur les 10 échantillons collectés, un index P7 est ajouté afin d'identifier chacun d'eux. A l'issu du séquençage MiSeq V2 en mode paire-ended 2x150 cycles, les 30 millions de *reads* sont alignés contre le génome de souris de référence mm9 à l'aide de l'outil HISAT2 via le serveur usegalaxy. La matrice d'expression de chaque échantillon est générée en comptant les *reads* associés à chaque gène à l'aide de l'outil featureCounts, puis en procédant à une étude de l'expression différentielle à l'aide de l'outil DESeq2, en considérant comme facteur de différenciation primaire le traitement appliqué sur les cellules (LPS ou PBS) et comme facteur secondaire le temps d'incubation. Cette analyse différentielle résulte en une matrice normalisée, où chaque colonne représente un échantillon et chaque ligne un gène. Des graphes sont également générés, permettant de constater les variations entre échantillons, dont l'un d'eux est une analyse en composantes principales. Chaque échantillon est un point du graphe et les 2 axes constituent les 2 premiers composants principaux, PC1 et PC2.

Les résultats de cette expérience sont présentés dans la Figure 49.

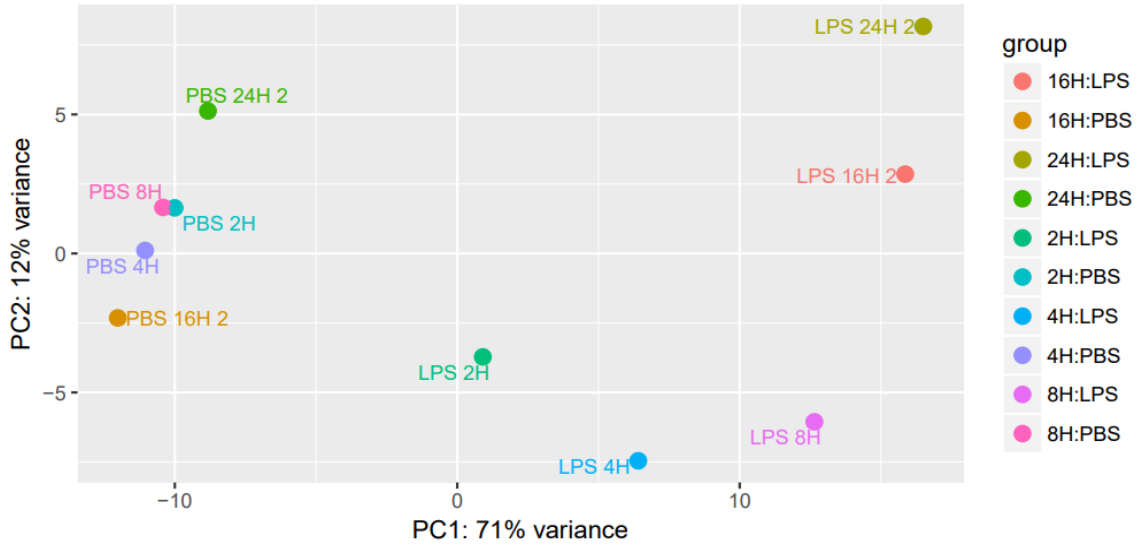


Figure 49 Analyse en composantes principales de cellules BV2 activées au LPS ou contrôle (PBS) pendant différents temps d'activation, à partir de données de séquençage de transcriptome complet en masse, en utilisant les outils fournis par usegalaxy.org, à savoir l'outil d'alignement , l'outil FeatureCounts pour construire des matrice d'expression et enfin l'outil DESeq2 permettant de définir des groupes de cellules et de déterminer les gènes différentiellement exprimés entre échantillons

On constate que l'axe PC1 est celui qui explique la majorité de la variation entre les échantillons. Au niveau de cet axe, les échantillons correspondant aux cellules non activées (PBS) sont regroupés tandis que les échantillons correspondant aux cellules activées se répartissent le long de cet axe, en fonction du temps d'activation. L'incubation avec du PBS, quel que soit le temps d'incubation, ne semble pas avoir entraîné de changements transcriptionnels majeurs. En revanche, les échantillons activés au LPS montrent une évolution dans leur transition en fonction du temps d'incubation, et qui semble atteindre un plateau à 16 heures d'activation.

A partir de la table d'expression normalisée, nous avons filtré les gènes les plus différentiellement exprimés, selon les critères définis, entre échantillons (*p-value* < 0,05 et *absolute fold-change* > 2) et obtenu une liste de 240 gènes. On y retrouve notamment certains de nos marqueurs d'intérêt à savoir TNFa, Ccl2, Cd14, Lgals3, Il1b et Il1Rn.

En nous basant sur ces résultats, nous avons choisi de nous focaliser sur une activation à 2 heures, 6 heures et 18 heures pour analyser l'évolution dans le temps de l'expression de marqueurs de l'inflammation après une induction au LPS de cellules BV2.

B. Analyse ciblée moyenne

Des cellules BV2 ont été activées pendant 2 heures, 6 heures ou 18 heures au LPS, ou bien ont été mises simplement en présence de PBS (contrôle négatif) et l'expression des 14 marqueurs d'intérêt utilisés jusqu'à présent a été mesurée en masse, afin d'évaluer si celle-ci est soumise à une variation dans le temps. La variation dans l'expression des cibles en fonction de l'état d'activation et du temps d'activation a été analysée par RT-qPCR à partir d'ARN extraits à partir de cellules activées pendant les différents temps d'intérêt ou contrôle ou en RNAseq ciblé en gouttes, selon le protocole optimisé, à savoir en encapsulant les cellules dans des gouttes de 2nL avec des billes de 100pL porteuses de 14 amorces spécifiques à codes-barres, et en suivant une stratégie d'amplification par PCR. Les billes utilisées sont les mêmes que celles décrites dans la partie précédentes (Figure 37).

L'ARN total a été extrait à partir de cellules BV2, activées avec du pendant LPS pendant 2 heures, 6 heures ou 18 heures ou de cellules contrôle (ajout de PBS dans le milieu), puis soumis à une RT-qPCR en tube, en utilisant des billes à amorces spécifiques à codes-barres, supplémentées avec une amorce spécifique au gène de référence Rpl13a libre. Les résultats de RT-qPCR ont été comparés aux données moyennes obtenues à partir du séquençage des échantillons correspondant à chaque point d'activation. Les résultats sont présentés dans la Figure 50.

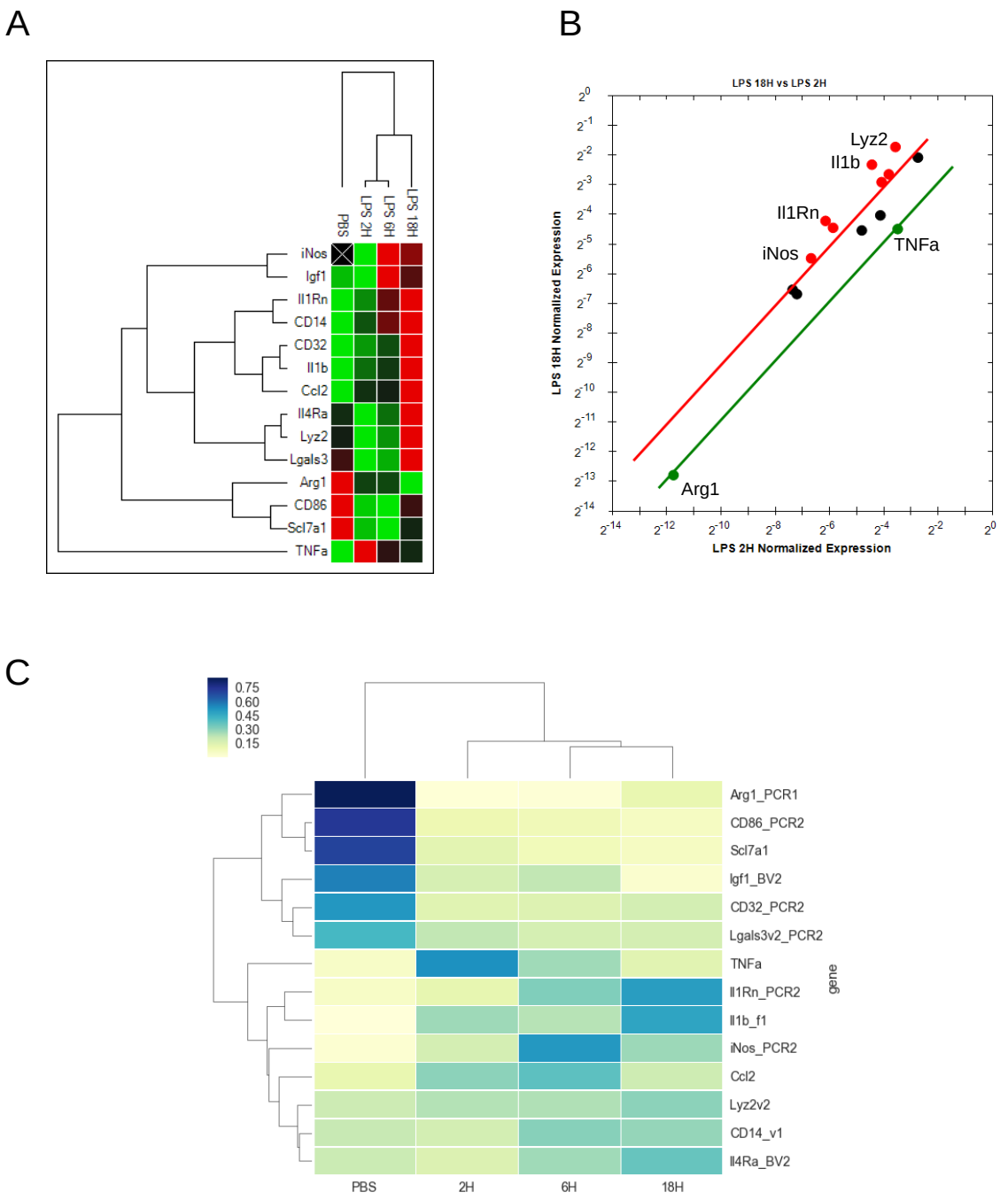


Figure 50 Influence du temps d'activation de cellules BV2 avec du LPS et comparaison entre des données issues d'expériences de RT-qPCR sur ARN extrait en tube et des données RNAseq ciblé en gouttes, A = Résultats de RT-qPCR en tube présentés sous forme de Clustermap, générés à l'aide du logiciel CFX Manager ; représentant les échantillons et les cibles, regroupés en fonction des similarités dans l'expression relative des gènes. Une expression plus forte est codée en rouge, une expression plus basse en vert et aucun changement d'expression en noir. Plus la couleur est claire, plus le niveau d'expression est élevé, B = Graphique en nuage de points, généré à l'aide du logiciel CFX Manager, permettant de déterminer les transcrits différentiellement exprimés entre des cellules activées 2h ou 18h au LPS, avec un seuil de variation supérieur à 2 ordres de grandeurs, d'après les données de RT-qPCR, C = Mesure de l'expression relative moyenne sur les 4 échantillons séquencés, correspondant à des émulions où ont été respectivement encapsulés des cellules BV2 activées au LPS pendant 2H, 6H ou 18H ou non activées (PBS). Une matrice est générée à partir de la proportion des différents transcrits

dans chaque échantillon. Les résultats sont présentés sous forme de Clustermap, après homogénéisation du niveau de chaque transcrit

Le *clustermap* en A a été généré à l'aide du logiciel CFX Manager, après normalisation par rapport à l'expression du gène de référence Rpl13a dans chaque échantillon. Ce graphe représente les liens entre cibles ou entre échantillons en fonction de profils d'expression communs sous forme de dendrogramme. On constate ainsi que les échantillons correspondant à une activation au LPS se regroupent. La représentation en niveau de couleur ajoute une dimension supplémentaire à l'analyse et révèle le niveau d'expression relatif de chaque cible dans les différents échantillons ; le rouge indique une augmentation de l'expression, le vert une diminution et le noir aucun de changement. Le gène TNFa par exemple montre un comportement différent des autres cibles ; l'ajout de LPS active son expression, et celle-ci culmine 2 heures après l'induction puis diminue. Les gènes Il1Rn, Cd14, Cd32, Il1b, Ccl2 sont également surexprimés en réponse au LPS mais montrent quant à eux un pic d'expression 18 heures après induction. D'autres gènes voient leur niveau d'expression diminuer suite à l'ajout de LPS ; les gènes Arg1, Cd86 et Scl7a1.La représentation en *clustermap* présentée en A homogénéise le niveau d'expression de chaque cible de sorte à faciliter la visualisation et ne permet pas de comparer les niveaux d'expression des cibles entre elles ni de se rendre compte du degré significatif dans les variations observées

Le graphe B de la Figure 50 représente les différences d'expression des cibles entre l'échantillon correspondant aux cellules activées 2 heures ou 18 heures au LPS et le degré significatif dans les variations observées en fonction d'un seuil défini. Les points rouges indiquent une surexpression de plus de 2 ordres de grandeur dans l'échantillon 18 heures par rapport à l'échantillon 2 heures tandis que les points verts indiquent une sous expression de plus de 2 ordres de grandeur. Les points noirs indiquent qu'il n'y a aucune différence d'expression significative entre les 2 échantillons, à savoir au-dessus du seuil de 2 ordres de grandeur fixé. On constate ainsi que les gènes Lyz2, Il1Rn, Il1b et iNos montrent une expression significativement plus importante en augmentant le temps d'activation de 2 heures à 18 heures tandis que le gène TNFa montre une expression réduite lorsque l'activation est plus longue.

Finalement, le graphe C de la Figure 50 présente les résultats moyennés des données issues du RNAseq ciblé sur les cellules BV2 activées au LPS 2 heures, 6 heures ou 18 heures ou contrôle, les mêmes cellules que celles à partir desquelles a été extrait

l'ARN utilisé pour les expériences de RT-qPCR. Chaque échantillon est codé à l'aide d'un index illumina pour les distinguer après le séquençage NextSeq High Throughput, en mode paire-ended. Après identification des gènes, codes-barres et UMI, les codes-barres contaminants, correspondant à ceux comptant peu de *reads*, sont filtrés et des matrices d'expression sont construites, après normalisation à l'aide des UMI. Une matrice moyenne est également générée, où les colonnes représentent les échantillons, les lignes les transcrits, et les valeurs la proportion d'UMI associés à chaque cible par rapport au nombre d'UMI totaux dans les différents échantillons. Un *clustermap* est construit à partir de cette matrice, après normalisation des lignes, afin d'homogénéiser le niveau d'expression des différentes cibles entre elles. Ce graphe offre deux dimensions d'analyse ; le *heatmap* permet de codifier le niveau d'expression sous forme de gradient de couleurs tandis que le dendrogramme hiérarchise les liens entre échantillons et transcrits. On observe le même regroupement des cellules activées au LPS entre elles que dans le graphe A obtenu à partir des données de RT-qPCR. La dimension *heatmap* montre également une augmentation du niveau d'expression des gènes Il1Rn, Il1b, et iNos avec le temps d'activation tandis que l'expression relative du gène TNFa est plus importante aux premières heures d'activation.

Les conclusions faites à partir des données en masse et des données en cellule unique moyennées sont donc encore une fois similaires et biologiquement cohérentes. Nous avons choisi de nous focaliser sur l'expression des gènes Il1Rn, Il1b, iNos et TNFa à l'échelle de la cellule unique car ces gènes montrent une évolution transcriptomique dépendante du temps d'activation.

C. Analyse ciblée à l'échelle de la cellule unique

Nous nous sommes ensuite focalisés sur 4 cibles d'intérêt afin d'étudier l'évolution de leur profil d'expression dans le temps à l'échelle de la cellule unique, à des expériences en RNAseq ciblé en gouttes. Nous avons procédé à des analyses qualitatives afin d'évaluer la dynamique d'activation de ces gènes, en regardant à la fois la densité de cellules exprimant le gène aux différents temps, et le niveau d'expression associé. Les résultats sont présentés dans la Figure 51.

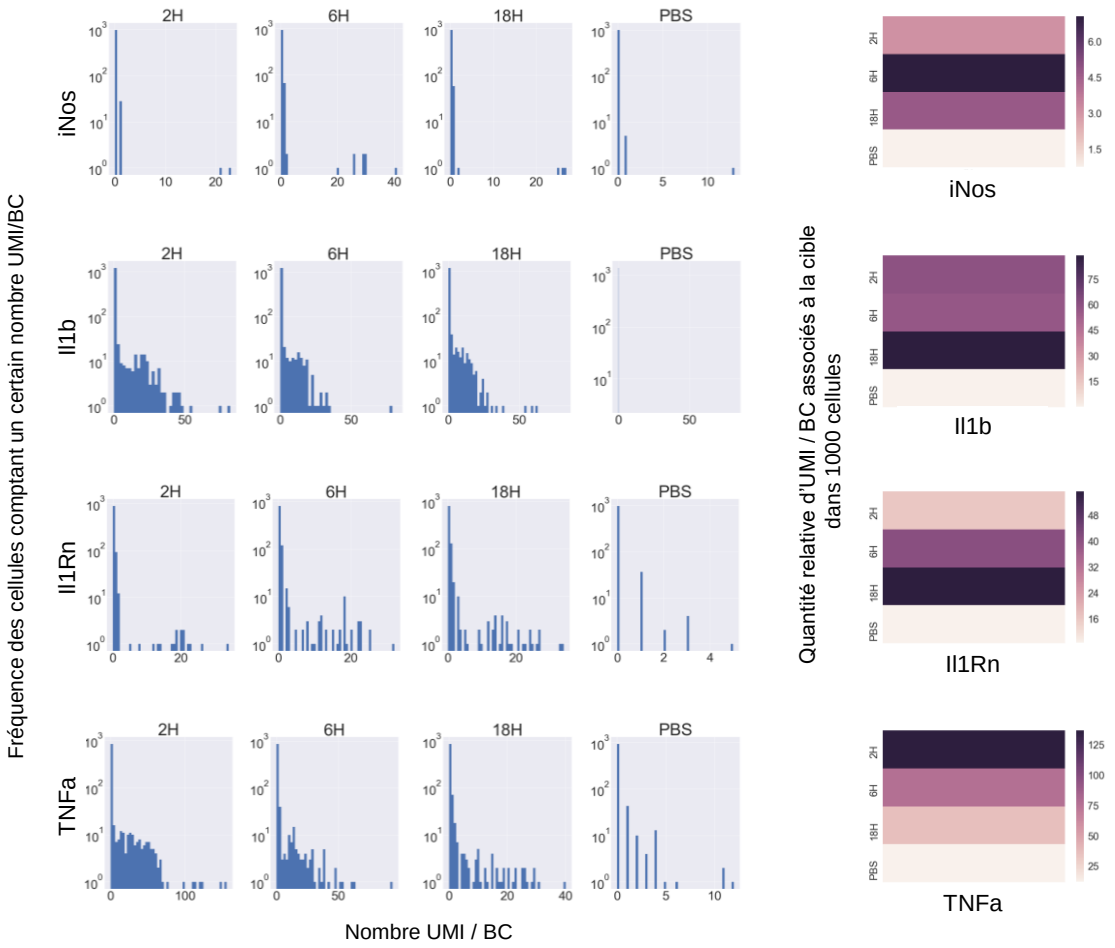


Figure 51 Analyse temporelle de l'inflammation à à l'échelle de la cellule unique par RNAseq ciblé en gouttes sur des cellules BV2 activées au LPS pendant 2 heures, 6 heures ou 18 heures de cellules contrôle (PBS). L'étude de 4 des 14 marqueurs ciblés sont ici présentés, au regard des analyses moyennes précdemment faites et montrant une expression varibale de ces dernières dans le temps ; iNos, Il1b, Il1Rn et TNFa. La fréquence de cellules comptant un certain nombre d'UMI d'une cible donnée à chaque temps d'activation est présentée sous forme d'histogramme (Gauche), et l'expression relative moyenne de chaque cible dans le temps est présentée sous forme de heatmap (Droite)

Les résultats sont présentés sous forme d'histogramme de la densité de cellules associées à un niveau d'expression donné (défini par le nombre d'UMI par code-barres), à chaque temps d'activation et pour chacune des 4 cibles d'intérêt. Un *heatmap* a également été généré, à partir des matrices d'expression normalisées par rapport au total d'UMI par goutte, afin de représenter la quantité total d'UMI présents à chaque point temps et pour chaque cible. Ce graphe donne une image moyennée des résultats à l'échelle de la cellule unique. Afin de comparer les données, 1000 cellules de chaque échantillon ont été considérées pour générer les différents graphes, et s'affranchir ainsi des différences dans le nombre de cellules collectées entre chaque échantillon.

La cible iNos montre l'apparition d'une petite sous-population de cellules exprimant de l'ordre de 20 transcrits par cellule en cas d'activation au LPS. Il n'y a aucune expression du gène en l'absence de LPS. Le niveau d'expression est constant d'un temps à l'autre mais on constate une densité plus importante de cellules après 6 heures d'activation. L'expression du gène semble donc On/Off, avec une distribution bimodale des cellules en fonction de l'expression de ce gène. L'augmentation du nombre de cellules exprimant iNos semble graduelle, jusqu'à atteindre un maximum à 6 heures puis à diminuer de nouveau. Les remarques faites à partir des histogrammes sont semblables à ce qu'on observe à l'aide du *heatmap*, à savoir l'expression de plus de transcrits à 6 heures d'activation.

En ce qui concerne l'expression du gène Il1b, celle-ci est inexistante sans induction au LPS. En revanche, dès 2 heures d'activation, on observe l'apparition de cellules exprimant jusqu'à 50 transcrits par cellule. Le niveau d'expression semble très variable d'une cellule à l'autre. La densité de cellules exprimant Il1b, de même que le niveau d'expression de ces cellules, semble constant au cours du temps.

Le gène Il1Rn présente un profil d'expression bimodal. Une sous-population de cellules activées, avec une expression de l'ordre de 20 transcrits, apparaît après 2 heures d'activation au LPS. La densité de cellules constituant cette sous-population augmente avec le temps d'activation, en accord avec le gradient observé sur le *heatmap.*

Finalement, la cible TNFa diffère des 3 autres. Elle est à son maximum d'expression 2 heures après activation au LPS puis tant la densité de cellules exprimant ce transcrit

que le niveau d'expression diminue. Cette diminution à double niveau laisse présager une cinétique de retrait de l'expression aussi rapide que celle de son apparition. Encore une fois, le *heatmap* montre également que le nombre d'UMI associés à ce transcrit est maximal 2heures après activation au LPS.

Nous avons ensuite voulu nous intéresser à l'expression des cibles Il1b et Il1Rn plus particulièrement, pour mesurer leur niveau d'expression relatif à l'échelle de la cellule unique. Nous nous sommes focalisés sur ces 2 cibles car elles sont étroitement liées : le gène Il1Rn code pour un co-récepteur d'Il1b ayant sur la cytokine un effet inhibiteur. C'est un antagoniste d'Il1b.

Les cellules exprimant l'une des 2 cibles, ou les 2, ont été extraites à partir des matrices d'expression de chaque échantillon, correspondant aux différents temps d'activation. La proportion de cellules positivement sélectionnée a été calculée dans chaque échantillon. La matrice à 2 colonnes résultantes a été normalisée par ligne, de sorte à connaître la proportion d'UMI attribuée à l'une ou l'autre cible dans chaque cas. Les résultats sont présentés sous forme de *clustermap* dans la Figure 52.

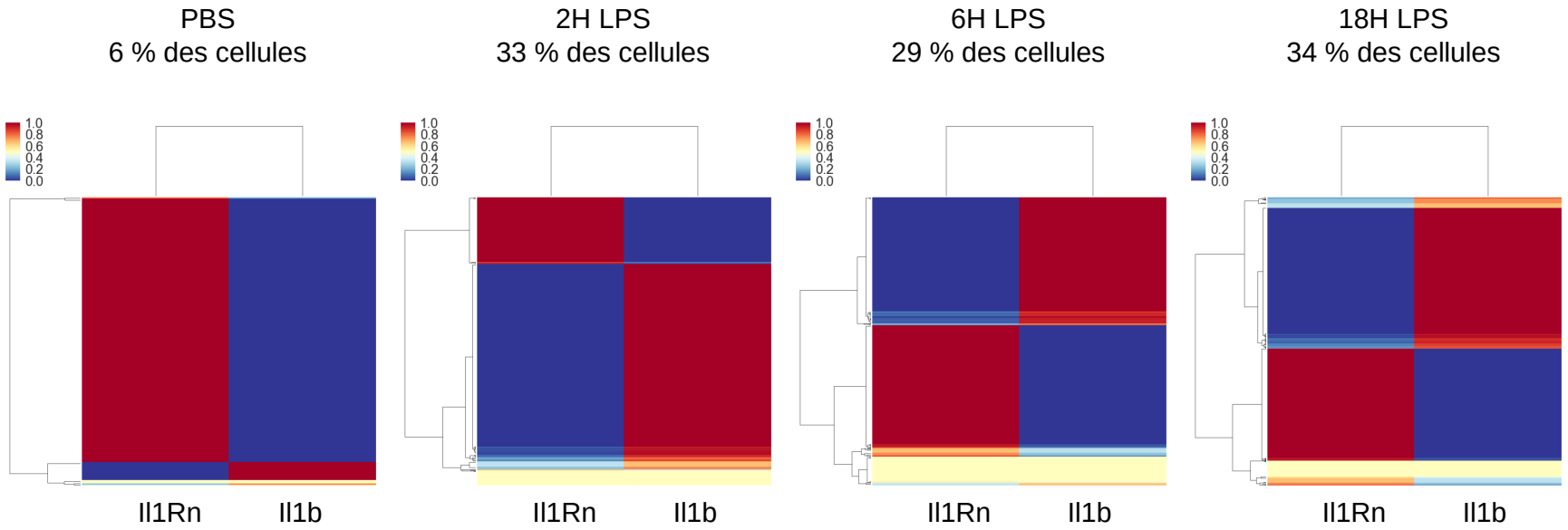


Figure 52 Expression relative du gènes Il1b et du gène Il1Rn au sein de cellules BV2 activées pendant différents temps au LPS ou non activées (PBS). Seules les cellules exprimant Il1b, Il1Rn ou les 2 ont été extraites à partir des matrices d'expression normalisées par rapport au nombre total d'UMI par ligne (par cellule), générées à partir du RNAseq ciblé en gouttes sur les 4 échantillons. Les données sont présentées sous forme de clustermap, montrant la répartition des UMI dans chaque goutte positive au filtre entre les 2 cibles étudiées. Le pourcentage indique la proportion de cellules conservées après application du filtre, c'est-à-dire la proportion de cellules exprimant Il1Rn, Il1b ou les 2.

On observe qu'en l'absence d'une induction au LPS, très peu de cellules expriment Il1Rn et encore moins Il1b. En effet, seules 6% des cellules totales expriment l'une des 2 cibles ou les 2.

En revanche, dès 2 heures d'activation, près de 30% des cellules expriment l'une des 2 cibles, parmi lesquelles 14% expriment les 2 cibles en même temps. Une plus large portion des cellules exprime Il1b, ce qui fait sens avec les observations précédentes, qui montraient que l'activation de l'expression du gène Il1b était efficace dès 2 heures d'activation puis se maintenait à un niveau plus ou moins constant, tandis que l'expression du gène Il1Rn n'avait pas encore atteint son pic culminant à 2 heures d'activation. A ce stade, les cellules exprimant les 2 gènes montrent un niveau d'expression des 2 cibles équivalent.

A 6 heures d'activation, la proportion de cellules exprimant Il1Rn est plus grande et est maintenant équivalente à celle de cellules exprimant Il1b. La proportion de cellules exprimant les 2 cibles et elle aussi plus importante ; elle représente 20% des cellules sélectionnées. Ces dernières s'organisent en 3 groupes d'importance équivalente ; le premier est associé à une expression d'Il1b majoritaire, le second à une expression d'Il1Rn majoritaire et le troisième, où les deux niveaux d'expression sont équivalents.

Finalement, à 18 heures d'activation, point culminant pour l'expression d'Il1Rn, les sous-groupes sont assez similaires à ceux observés à 6 heures d'activation. Parmi les 20% de cellules exprimant les 2 cibles en même temps, le groupe correspondant à une expression au même niveau des 2 cibles a vu sa proportion diminuer, au profit des sous-groupes où l'une des 2 cibles prend le dessus sur l'autre.

Ces analyses montrent que les 2 gènes ont une expression éclatée entre cellules et que la probabilité de co-expression est assez faible. En termes de nombre de cellules exprimant une cible, les 2 cibles semblent équivalentes. Le *clustermap* étant normalisé, il est difficile de conclure quand au niveau d'expression relatif de chaque gène. Finalement, il est délicat de conclure quant à l'hypothèse d'une expression alternée des 2 gènes au sein d'une même cellule ou à l'hypothèse de l'expression des 2 gènes par des cellules différentes.

IV. Remarques conclusives

Les comparaisons entre différents protocoles de RT en gouttes ainsi que d'amplification des ADNc ont permis d'établir un protocole optimal de scRNAseq ciblé en gouttes, permettant de détecter l'expression significative de gènes cibles à l'échelle de la cellule unique, d'une valeur typiquement supérieure à 20 transcrits par goutte. Point important, les données moyennées issus du séquençage RNAseq ciblé sont similaires à celles obtenues par des expériences en masse sur un même échantillon, montrant le maintien d'une réelle cohérence biologique.

Déjà mis en évidence lors des expériences d'optimisation, la réaction de capture des ARNm et de leur RT en gouttes est plus efficace lorsque les cellules sont encapsulées dans des gouttes de 2nL avec des billes de 100pL, dont la capacité fonctionnelle est supérieure à celles des billes de 10pL utilisées pour compartimenter les cellules dans des gouttes de l'ordre de 100pL. On dénombre en effet jusqu'à 5 fois plus de transcrits capturés par gouttes, et une distribution bimodale des cellules pour l'expression d'une cible donnée. Cette solution (gouttes de 2nL et billes de 100pL) s'accompagne cependant d'une limitation dans le nombre de cellules pouvant être analysées lors d'une encapsulation.

Le choix d'un protocole d'amplification des ADNc implique de définir précisément les besoins de l'expérience à réaliser. Une amplification linéaire par IVT permet une répartition des *reads* et UMI plus homogène entre les différents transcrits. Cependant, ce mode d'amplification souffre d'une spécificité et d'une sensibilité moindre que le protocole de PCR, entraînant une perte de signal lors des étapes d'alignement. L'amplification par PCR multiplex montre de très bon taux d'alignement, quelque-soit la quantité initiale de départ, et permet la capture de 10 fois plus de transcrits par gouttes. Une question demeure tout de même quant à la présence de faux UMI, augmentant artificiellement le nombre de molécules quantifiées par gouttes et issus d'erreurs introduites au cours des cycles répétés d'amplification ; Nous avons montré qu'une simple filtration en fonction du nombre de *reads* par UMI n'était pas optimale car elle entraîne une perte de signal pour les cibles à l'expression faible, qui ont potentiellement étaient amplifiés plus tardivement au cours de l'étape de PCR multiplex et donc pendant moins de cycles. L'utilisation d'autres modes de filtration est donc à prévoir, en combinaison avec cette stratégie d'amplification des ADNc.

Notre technologie de RNAseq ciblé en gouttes sur une dizaine de marqueurs de l'inflammation est donc prometteuse. Cependant, l'étude d'un nombre si restreint de transcrit, dont l'expression est soumise à induction et de ce fait séquentielle, ajouté aux limites techniques d'efficacité de capture des ARNm en gouttes, rend très complexe l'identification de sous-groupe de cellules. Pour les raisons évoquées, il est plutôt rare de détecter l'expression de plusieurs cibles dans une goutte, correspond à une image figée du transcriptome d'une cellule alors même que celui-ci évolue en permanence.

Finalement, une technologie d'analyse du transcriptome complet de la cellule semble plus adaptée à l'identification de sous-groupes de cellules sans à priori. Des outils informatiques puissants parviennent à établir des traits similaires entre cellules, permettant de définir des groupes et ce malgré une mauvaise efficacité de capture des ARNm au sein de la goutte. Les « faux-zéro » de la matrice d'expression, correspondant à des transcrits présents dans la cellule mais non capturés, sont compensés en fonction de l'appartenance d'une cellule à un groupe, dont l'empreinte transcriptomique a été établie. Notre technologie en revanche est plus adaptée à une étude avec à priori, sur le comportement de cibles d'intérêt. Nous proposons d'ailleurs une ébauche d'étude de la dynamique de l'inflammation sur des marqueurs d'intérêt à l'échelle de la cellule unique, en procédant à un RNAseq ciblé de cellules BV2 activées à différents temps.

Chapitre 5 : Les virus

I. Introduction

Les virus Influenza, plus largement connus sous le nom de virus de la grippe, s'attaquent aux voies respiratoires, provoquant des symptômes pouvant s'avérer mortels, en particulier lorsque des personnes fragiles sont touchées. Deux types de groupes pathogènes existent chez l'homme, les types A et B. Très facilement transmissibles, leur constante variabilité génétique leur confère un grand avantage évolutif. Ils constituent donc un problème de santé majeur. Leur évolution génétique peut se faire selon 2 stratégies ; par glissement antigénique ou par cassure (Institut Pasteur, 2014). La première consiste en des mutations ponctuelles des gènes codant pour les marqueurs de surface du virus, engendrant des modifications mineures au niveau des antigènes et une moins bonne reconnaissance du virus par les cellules du système immunitaire adaptatif. Ce phénomène est responsable de la grippe saisonnière, nécessitant la conception chaque année de nouveaux vaccins. La seconde stratégie consiste en un réassortiment génétique entre plusieurs souches virales, conduisant à des modifications drastiques des virus et de leurs antigènes, par la création de toutes nouvelles combinaisons. Cette évolution est possible uniquement chez les virus de type A. Le réarrangement entre souches d'origine animale distinctes est possible, et notamment chez le porc qui peut être infecté par des souches d'origine humaine ou encore aviaire, pouvant conduire à l'apparition de souches pandémiques. Ce phénomène est donc responsable de l'avènement de toutes nouvelles formes virales, dont le phénotype résultant est imprédictible et potentiellement différent de celui de ses 2 parents, et finalement à des pandémies (Lowen, 2017).

Le virus Influenza de type A (IAV) est formé par une capsule, fait d'une enveloppe lipidique recouverte des glycoprotéines antigéniques HA (Hémagglutinine) et NA (Neuraminidase), dont le rôle est de faciliter respectivement l'entrée et la sortie du virus de la cellule hôte (Figure 53). Ces protéines sont les plus variables et sont situées à la surface du virus. Elles sont les cibles antigéniques des cellules immunitaires et chaque souche virale est nommée en fonction du type de HA et NA qu'elle possède. Cette enveloppe contient l'information génétique du virus, sous forme d'ARN simple brin à orientation anti-sens, segmenté en 8 parties codant pour une dizaine de protéines dont

10 essentielles à savoir (les protéines sont données dans l'ordre du segment par lequel elles sont encodées, les différents segments sont séparés par une virgule) : PB2, PB1, PA, HA, NP, NA, M1 et M2, NS1 et NEP. PB1, PB2 et PA constituent les sous-unités formant l'ARN polymérase - ARN viral dépendante, responsable de la transcription et de la réplication du génome viral dans la cellule hôte. La nucléoprotéine NP se lie à chaque segment ARN viral, conjointement avec une copie de l'ARN polymérase, pour former une structure complexe nommée ribonucléoprotéine virale (vRNP), qui sera libérée dans le cytoplasme de la cellule hôte lors d'une infection avant de rejoindre le noyau pour être multipliée. M1 et M2 sont encodées sur le même segment (numéro 7). M1 est une protéine matricielle et M2 est un canal ionique à la membrane du virus, facilitant la libération des vRNPs dans la cellule hôte. La protéine NS a pour rôle d'inhiber la réponse antivirale dans la cellule hôte. La protéine NEP quant à elle, favorise l'exportation des vRNPs néosynthétisées hors du noyau. Une fois répliqué, l'ARNm viral est traduit en protéines dans le cytoplasme de la cellule hôte et les vRNPs néosynthétisées s'assemblent sous leur forme infectieuse, appelée virion, à la membrane de la cellule hôte pour créer un bourgeon et finalement une nouvelle particule virale, capable d'infecter les cellules voisines (Krammer *et al.*, 2018)(Samji, 2009).

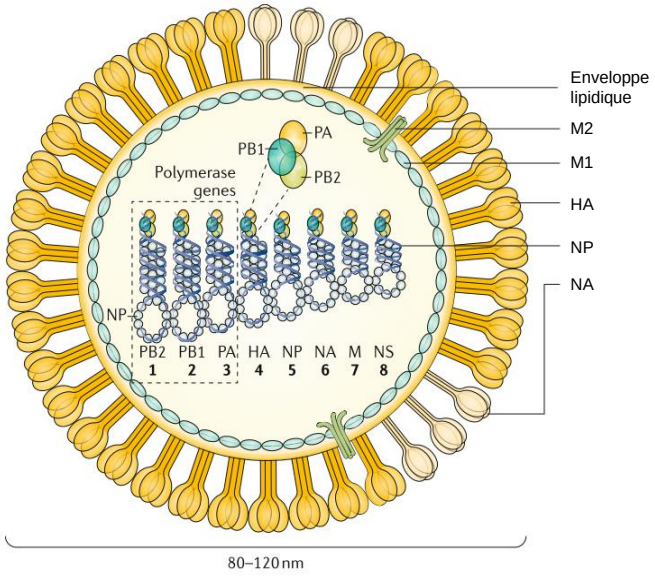


Figure 53 : Schéma de la structure du virus Influenza, extrait et traduit à partir de (Krammer *et al.*, 2018)*. Le virus est fait d'une double membrane lipidique dans laquelle sont insérées les glycoprotéines HA et NA. Le génome viral est codé sous forme d'ARN simple brin anti-sens, segmenté en 8 parties. Il est structuré en ribonucléoprotéine, par interaction avec la protéine NP et avec une copie par segment d'une ARN polymérase –ARN viral dépendante, formée par les 3 sous-unités PB1, PB2 et PA.*

Des particules virales dépourvues de segment ou avec une partie seulement des segments peuvent être exportées hors de la cellule hôte mais la plupart d'entre elles incorporent les 8 segments, un de chaque type. Seuls les virions possédant les 8 segments ARNs sont potentiellement infectieux. Des observations, faites en microscopie électronique à transmission, de coupes transversales au niveau des bourgeons viraux, ont montré que l'empaquetage des segments viraux sous-forme de virion suit un modèle précis, très conservé et né d'interactions se faisant entre segments adjacents, grâce à des signaux spécifiques présents sur chaque segment. Ce modèle est basé sur une structure dite « 7+1 » où 7 segments en entourent un seul, favorisant la formation de virions complets (Hutchinson, von Kirchbach, Gog, & Digard, 2010)(Noda *et al.*, 2006) (White & Lowen, 2018).

C'est grâce à sa structure segmentée que le virus peut évoluer par réassortiment génétique en cas de co-infection par deux souches différentes d'une cellule hôte, en créant de nouveaux virions. En théorie, en cas de co-infection par deux virus possédant chacun 8 segments différents, 256 (2^8) nouveaux assemblages pourraient ainsi être générés. Les analyses *in vitro* ou *in vivo* faites à ce jour n'ont cependant pas permis de détecter autant de combinaisons, et tendent vers l'hypothèse d'un biais favorisant certaines recombinaisons. Les signaux contrôlant l'assemblage des différents segments sont en effet spécifiques d'un segment et d'une souche donnés et vont impacter la probabilité d'assemblage entre segments d'origine différente. De même, la compatibilité entre protéines virales d'un nouveau virus entre en jeu dans l'efficacité d'infection d'une nouvelle cellule hôte. Certains hybrides sont ainsi négativement sélectionnés car n'étant pas complètement infectieux (Gerber, Isel, Moules, & Marquet, 2014)(White & Lowen, 2018).

Les techniques actuelles pour prédire les réassortiments entre deux souches virales sont à très bas débit (Isel, Munier, & Naffakh, 2016). Nos partenaires ont ainsi évalué que la prédiction de nouveaux virus à partir de 2 souches différentes nécessite d'analyser plus de 10^5 virus, de sorte à couvrir toutes les nouvelles combinaisons ainsi que leur quantité relative, soit leur probabilité d'occurrence, et à être statistiquement significative.

Le but de la collaboration entre l'équipe de Nadia Naffakh à Pasteur et notre équipe à l'ESPCI Paris est de mettre au point une technique pour l'étude à très haut débit de virus créés suite à une co-infection de cellules hôtes par deux souches distinctes, afin d'évaluer les virus infectieux réassortis les plus favorables. Cette technologie servira d'outil prédictif pour faciliter la conception de vaccin en amont et limiter le risque de pandémie.

II. Description de la stratégie expérimentale

Nous avons développé une technologie RNAseq ciblé capable de capturer des segments viraux et d'identifier leur origine, pour l'étude à très haut débit du phénomène de réassortiment génétique entre souches virales distinctes et la prédiction de nouveaux virus ainsi que leur probabilité. Une preuve de concept est réalisée en analysant le réarrangement entre 2 souches saisonnières du virus Influenza humain.

La première étape de cette preuve de concept, présentée au cours de ce chapitre, consiste à tester notre technologie de RNAseq ciblé sur les 2 souches virales, traitées séparément, c'est-à-dire que les cellules hôtes n'ont été infectées que par un type de souche à la fois (absence de co-infection). Nous avons voulu vérifier si l'origine de chaque segment est identifiable et si l'aspect cellule unique, ou ici souche unique, est conservé tout le long du protocole, afin d'évaluer les biais techniques pouvant engendrer l'apparition de fausses combinaisons. Le schéma de l'expérience est présenté dans la Figure 54.

Des cellules pulmonaires humaines A549 ont été séparément infectées avec deux types de souches virales, H1N1 ou H3N2. Cette infection a été réalisée à haut MOI (Multiplicity Of Infection), afin d'assurer un fort taux d'infection. Les cellules infectées par H1N1 ou H3N2 ont été fixées avec 80% de méthanol avant encapsulation afin d'inactiver les virus et pouvoir travailler en laboratoire L2. Elles contiennent donc les ARN viraux. Les cellules infectées avec la souche H1N1, qu'on nommera cellules H1 ou avec la souche H3N2, qu'on nommera cellules H3, ont ensuite été encapsulées en gouttes avec des billes à code-barres et porteuses d'amorces spécifiques des différents segments viraux. Trois encapsulations ont été réalisées, respectivement en utilisant des cellules H1, ou des cellules H3 ou un mélange de 50% cellules H1 et 50% de cellules H3 (Figure 54).

Les amorces ont été conçues de sorte à cibler des régions variables des 8 segments viraux, où la souche d'origine peut être identifiée distinctement. Deux stratégies ont été employées pour la conception des amorces ciblant les différents segments ; soit l'amorce spécifique de RT est commune aux 2 souches (segments NP, NS, PB1, PB2, PA, M1) avec une zone variable en 5' de l'amorce, soit 2 amorces de RT différentes pour chaque souche permettent de cibler un segment donné (segment NA et HA). Au total, chaque bille est donc porteuse de 10 amorces spécifiques.

Ces billes sont également porteuses de 4 segments d'index constituant le code-barres, A, B, C et D. Le dernier index D est unique et de séquence connue. Il permet de coder les différentes encapsulations ; ainsi les cellules H1 seules sont encapsulées avec des billes porteuses d'un index unique D1, les cellules H3 avec des billes porteuses d'un index unique D2 et le mélange H1 :H3 avec des billes porteuses d'un index unique D4.

Les 3 émulsions collectées sont cassées et les contenus des émulsions D1 et D2 sont mélangés avant amplification ou après amplification. Le mélange des 2 souches a donc été fait à 3 niveaux différents ;

- avant encapsulation,
- après encapsulation et avant amplification,
- après encapsulation et après amplification.

Chacun de ces mélanges différents a été codé par un index Illumina pour distinguer les molécules provenant de chaque mélange au moment du séquençage. Les cellules ont été infectées par une seule souche à la fois et ont ensuite été compartimentées dans des gouttes à hauteur d'une seule cellule par goutte. En théorie, à un code-barres de gouttes doit donc être associé les 8 segments provenant d'une seule souche.

Cette première expérience va permettre de répondre à plusieurs questions ;

- Les différents segments sont-ils reconnus ?
- Les différentes souches sont-elles identifiables ?
- L'aspect cellule unique, apporté par le code-barres, est-il conservé tout le long du processus ?

Pour répondre à cette dernière question, l'échantillon codé avec l'index i02 illumina, où les produits de synthèse issus des encapsulations en souche unique n'ont été mélangés qu'en fin de process, sert de référence. L'échantillon codé i03, correspondant aux mélanges des émulsions en souche unique avant amplification permet de vérifier s'il y a des échanges entre ADNc porteurs de l'information du code-barres et du segment, lors de l'étape d'amplification. En dernière instance, l'échantillon codé i05 permet de vérifier si l'encapsulation s'est bien faite à l'échelle de la cellule unique, de mesurer la proportion de double encapsulation et enfin de constater s'il y a eu des échanges d'ARN libres d'une goutte à l'autre.

La conception des amorces, les expériences préliminaires *in vitro* ainsi que la préparation des librairies de séquençage ont été réalisées par Kuang-Yu Chen de l'équipe « Influenza virus-host cell interactions » de l'institut Pasteur. J'ai participé à l'élaboration des protocoles de microfluidique et de préparation des librairies, à la préparation des billes à code-barres et à la manipulation en gouttes. Les analyses bio-informatiques (alignements des séquences, générations des matrices d'expression et analyses) ont été faites par Mathieu Bahin, de la plateforme bio-informatique de l'IBENS. Kuang-Yu et moi-même avons également participé aux analyses bio-informatiques. J'ai ainsi utilisé certains de mes scripts pour l'analyse des données de séquençage.

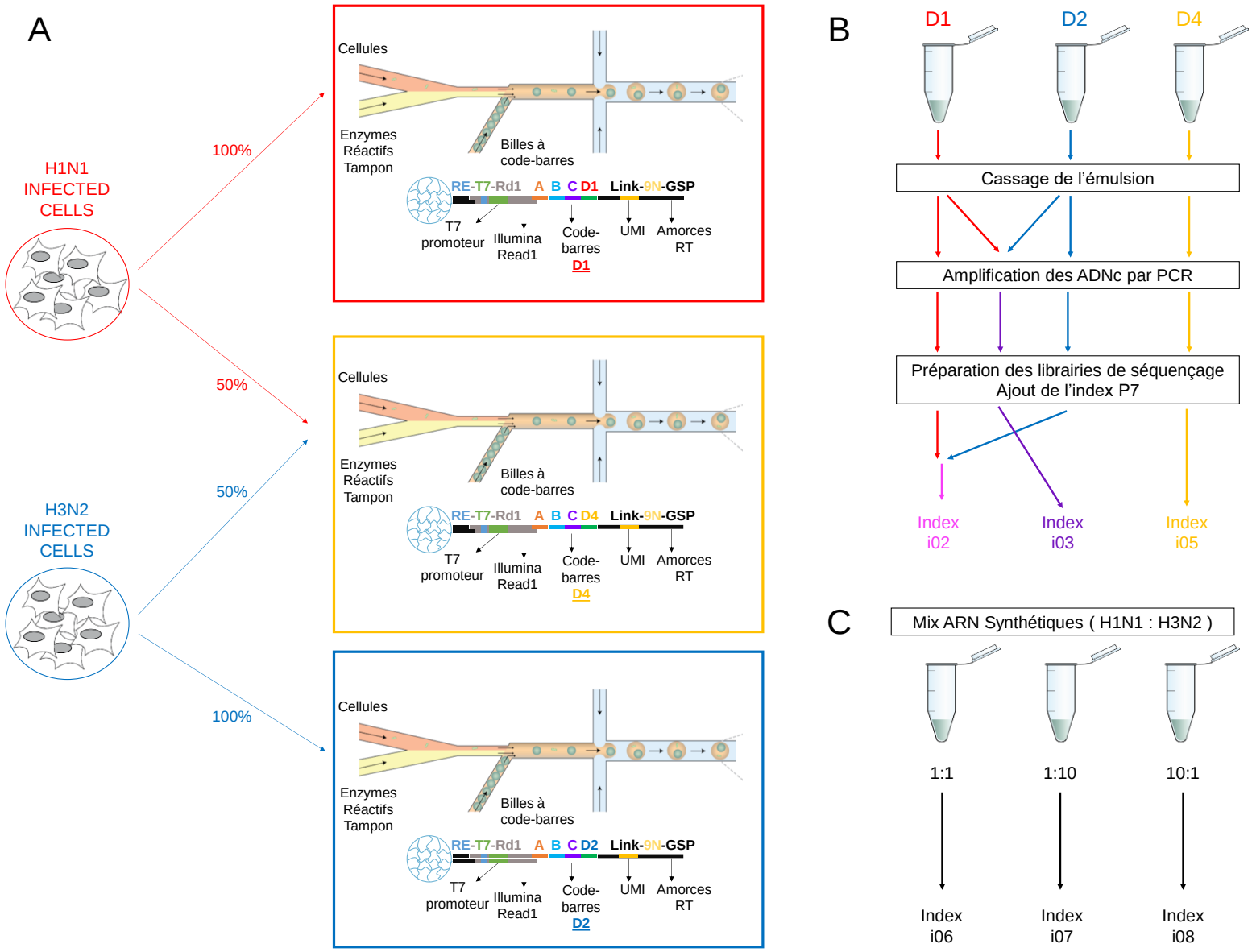

A
H1N1
INFECTED
CELLS
H3N2
INFECTED
CELLS
100%
50%
50%
100%
Cellules
Enzymes
Réactifs
Tampon
Billes à
code-barres
RE-T7-Rd1 A B C D1 Link-9N-GSP
T7
promoteur
Illumina
Read1
Code-
barres
D1
UMI
Amorces
RT
Cellules
Enzymes
Réactifs
Tampon
Billes à
code-barres
RE-T7-Rd1 A B C D4 Link-9N-GSP
T7
promoteur
Illumina
Read1
Code-
barres
D4
UMI
Amorces
RT
Cellules
Enzymes
Réactifs
Tampon
Billes à
code-barres
RE-T7-Rd1 A B C D2 Link-9N-GSP
T7
promoteur
Illumina
Read1
Code-
barres
D2
UMI
Amorces
RT
B
D1
D2
D4
Cassage de l'émulsion
Amplification des ADNc par PCR
Préparation des librairies de séquençage
Ajout de l'index P7
Index
i02
Index
i03
Index
i05
C
Mix ARN Synthétiques (H1N1 : H3N2)
1:1
1:10
10:1
Index
i06
Index
i07
Index
i08

Figure 54 Schéma des expériences préliminaires pour la mise au point d'une technologie de séquençage ciblé de l'ARN pour prédire les cassures génétiques chez le virus Influenza de type A, A= Encapsulations de cellules hôtes pulmonaires infectées par un virus de type H1N1 ou H3N2. Les cellules sont préalablement infectées à haut MOI puis fixées au méthanol avant d'être encapsulées dans des gouttes d'une centaine de picolitres avec des billes d'hydrogel porteuses d'amorces à code-barres (un code-barres unique par bille). Chaque bille est porteuse de 10 amorces spécifiques permettant de cibler les 8 segments viraux. Les billes utilisées pour chacune des 3 encapsulations ont été codées avec un index D afin de reconnaitre l'encapsulation d'origine. 3 émulsions sont produites, à partir des cellules infectées avec la souche H1N1 seule (Emulsion D1) ou à partir des cellules infectées avec la souche H3N2 seule (Emulsion D2) ou à partir d'un mélange homogène de ces 2 cellules (Emulsion D4). B = Les produits de synthèse de ces différentes émulsions sont ensuite amplifiés et préparés pour le séquençage. 3 échantillons sont finalement séquencés. Ils sont porteurs d'un index Illlumina permettant de les identifier après séquençage. L'échantillon i02 correspond à un mélange des 2 émulsions D1 et D2 après l'étape d'amplification, l'échantillon correspond au mélange des produits des 2 mêmes émulsions mais avant que ceux-ci aient été amplifiés. Les produits, issus des 2 émulsions, sont donc amplifiés en un seul tube, l'échantillon i05 correspond aux produits issus de l'émulsion D4. C = Une expérience parallèle en tubes a été effectué à partir d'ARN synthétiques et des mêmes billes à amorces que celles utilisées pour les 3 encapsulations.

III. Expériences préliminaires *in vitro*

Les expériences préliminaires *in vitro* ont pour but de valider les amorces de RT et de PCR, qui seront ensuite utilisées pour l'expérience de RNAseq ciblé en gouttes. Ces couples d'amorces doivent permettre d'une part la capture des segments viraux et leur rétrotranscription en gouttes et d'autre part l'amplification spécifique ADNc néosynthétisés issus des 8 segments de chaque souche et l'identification de la souche d'origine lors du séquençage des amplicons générés.

Dans un premier temps, nos partenaires ont effectué des essais en tubes pour valider la RT en amorces spécifiques sur des ARN synthétiques. Cette expérience a montré un déséquilibre de représentation entre les segments provenant des 2 souches, à l'avantage de la souche H3N2. De plus, certains segments semblent surreprésentés, tandis que d'autres ne fonctionnent pas. Il semble donc nécessaire de concevoir de nouvelles amorces, voire de changer de stratégie d'amplification des ADNc.

Des ARN synthétiques ont été conçus, de sorte à être porteur des mêmes séquences cibles que les ARN viraux, ainsi que des mêmes variabilités inter-souches. Les ARN synthétiques de type H1N1 ont été mélangés avec les ARN synthétiques de type H3N2 selon 3 ratios différents ;

- Les ARN sont tous en quantités équimolaires
  (H1 : H3 = 1 : 1)
- Les ARN de type H1 sont 10 fois moins concentrés que les ARN de type H3
  (H1 : H3 = 1 : 10)
- Les ARN de type H3 sont 10 fois moins concentrés que les ARN de type H1
  (H1 : H3 = 10 : 1)

Des billes porteuses des 10 amorces spécifiques des différents segments ont été ajoutées, avec enzymes, réactifs et tampons. Les ADNc une fois synthétisés ont été amplifiés selon une stratégie de PCR multiplex. La dernière étape de PCR permet l'ajout d'index Illumina de séquençage pour identifier chaque expérience au ratio différent (Figure 54, C). Après séquençage Illumina, l'outil « Kraken » a été utilisé pour déterminer la taxonomie des molécules séquencées. Ce système d'alignement Kraken assigne des étiquettes taxonomiques à de courtes séquences d'ADN, et vise à atteindre une sensibilité et une vitesse élevées en utilisant des alignements exacts de k-mers et un nouvel algorithme de classification.

La Figure 55 présente les résultats de cet alignement, ainsi que les résultats théoriques attendus. On constate que les ratios entre les deux souches sont similaires à ceux attendus, avec une légère surreprésentation de la souche H3N2, et ceux dans les 3 cas. Plusieurs hypothèses sont possibles pour expliquer cela ; une mauvaise quantification des ARN synthétiques, ayant conduit à une erreur lors de la préparation des mélanges, ou encore l'existence d'un réel biais technique, lié à l'efficacité de capture variable des différents segments ou encore à un biais d'amplification des différentes espèces (lié à l'efficacité des couples d'amorces).

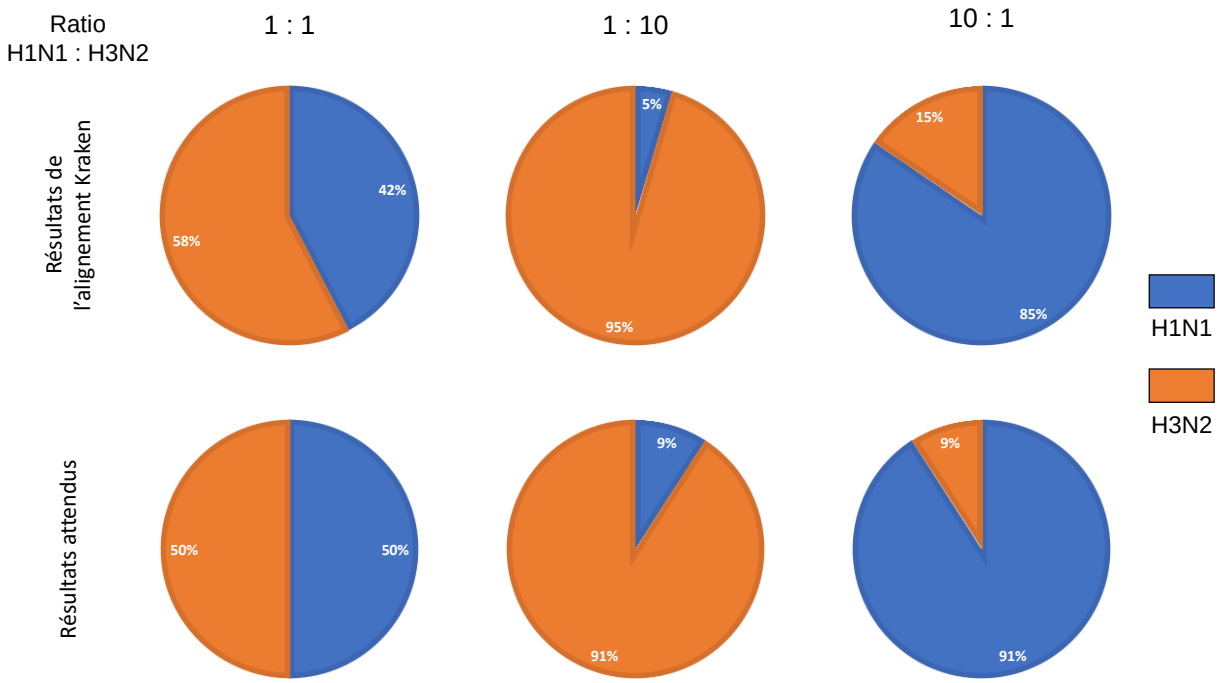


Figure 55 Détermination de l'origine taxonomique des différentes séquences après alignement avec l'outil Kraken sur les produits de séquençage issus d'une étude in vitro, où des ARN synthétiques correspondant aux 8 segments viraux de chaque souche de type H1N1 ou H3N2 (soit 16 segments au total) ont été mélangés à différent ratios avant de subir une rétrotranscription en présence de billes à amorces spécifiques à code-barres. Les billes sont porteuses de 10 amorces spécifiques permettant de cibler les 8 segments de chaque origine. Haut = proportion des molécules assimilées à une origine H1N1 ou H3N2 après alignement à l'aide de l'outil Kraken, Bas = Répartition attendue en fonction du ratio utilisé lors du mélange des ARN synthétiques

Ces séquences ont ensuite été alignées contre un génome de référence, contenant les séquences des amplicons cibles pour chaque souche, à l'aide de l'outil STAR (*Spliced Transcripts Alignment to a Reference*). STAR est un outil d'alignement contre génome de référence ultra rapide, capable d'identifier des variants d'épissage, mais

nécessitant l'utilisation de beaucoup de mémoire. La répartition des séquences pour chaque segment a été calculée et est présentée, pour le cas de l'expérience où les ARN sont ajoutés en quantité équimolaire, dans la Figure 56. Les ARNs étant tous présents en même quantité, on attend une répartition homogène des *reads* entre les différents segments. On constate de nouveau un fort déséquilibre entre les souches H1N1 et H3N2, et ceux en particulier pour les segments HA, NP, PA et PB2, où aucune séquence n'a été identifiée pour la souche H1N1. De même, quasiment aucune séquence n'a été alignée contre le segment M1, et ce pour les 2 souches. Le segment NS semble quant à lui, surreprésenté. La disproportion entre H3N2 et H1N1 semble finalement plutôt liée à un défaut de capture de certains segments pour la souche H1N1. Il semble nécessaire de concevoir de nouvelles amorces pour les segments n'ayant pas été identifiés, et de vérifier leur efficacité par qPCR. Enfin, une stratégie d'amplification linéaire par IVT, comme expliqué au cours des chapitres précédents, pourrait permettre de limiter la surreprésentation de certains segments au détriment d'autres.

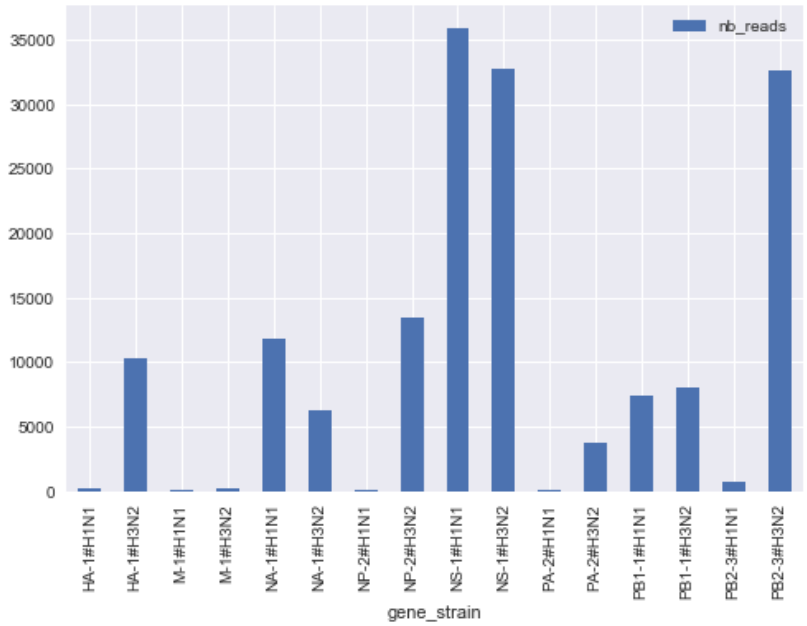


Figure 56 Distribution des reads associés aux différents segments cibles (pour chaque souche) dans l'échantillon indexé i06, où les ARN synthétiques de types H1N1 et H3N2 sont présents en quantité équimolaire (alignement à l'aide de l'outil STAR). Une répartition très hétérogène est observée alors qu'elle devrait théoriquement être homogène d'un segment à l'autre

IV. Expériences en gouttes

Les cellules infectées par chacune des 2 souches puis fixées ont ensuite été encapsulées suivant le schéma présenté en Figure 54. Dans un premier temps, les données ont été filtrées pour éliminer les codes-barres contaminants. Les données de séquençage ont ensuite été moyennées afin de les comparer aux données obtenues en tube. Les observations faites en tubes ou à partir des données moyennes de RNAseq ciblé en gouttes sont comparables et montrent une cohérence biologique.

Les analyses à l'échelle de la cellule unique des 3 échantillons séquencés ont permis de déterminer que les souches sont identifiables à l'échelle de la cellule unique. Elles ont également prouvé qu'il n'y pas d'échanges d'information entre les molécules d'ADNc à codes-barres lors de l'étape d'amplification, pouvant conduire à l'apparition d'hybrides. En effet à un code-barres donné n'est associé qu'une seule origine de souche, que ce soit dans l'échantillon où le produit des émulsions de chaque souche séparément a été mélangé en fin de préparation des librairies (après amplification), ou lorsque le mélange a été fait avant l'étape d'amplification. Enfin, si les 2 types de cellules sont encapsulés en même temps, on observe la présence de codes-barres associés à des segments issus des 2 souches simultanément, ce qui était attendu, étant donné la loi de Poisson suivant laquelle sont compartimentées les cellules et la présence d'agrégats de cellules dans la phase cellulaire.

A. Comment filtrer les données

Les manipulations en gouttes ont été réalisées en parallèle des expériences *in vitro* avec les mêmes billes, et donc les mêmes amorces, selon le plan expérimental décrit en Figure 54. La première étape de l'analyse, après avoir identifié gènes et codes-barres, consiste à filtrer les données pour éliminer les artefacts. Nos échantillons ont été sous-séquencés et il a été difficile de définir un seuil pour filtrer les données. Un contrôle qualité consistant à établir une matrice de corrélation nous a permis de valider le filtre appliqué.

Les échantillons ont été séquencés sur un séquenceur MiSeq v3, en mode paire-ended, sur 2x300 cycles, ainsi que 6 cycles pour l'indentification de l'index P7. À l'issue du séquençage, les données ont été alignées contre un génome de référence personnalisé avec les amplicons de chaque cible à l'aide de l'outil STAR pour l'indentification des gènes et souches d'origine. Les codes-barres ont quant à eux été

démultiplexés par alignement contre 4 fichiers contenant les 96 versions des index A, B, C et D, à l'aide de l'outil bowtie2. Environ 8 millions de séquences ont été obtenues à l'issue du séquençage contre 25 millions attendues. Cela peut être dû à une mauvaise quantification de l'échantillon. Parmi ces 8 millions de séquences, le taux d'alignement n'est que de 55% environ.

Nous souhaitons éliminer de l'analyse les codes-barres ne correspondant pas à une réalité biologique. Ces derniers peuvent être issus de codes-barres contaminants sur une bille, ayant conduit à la synthèse d'ADNc avec un mauvais code-barres au sein d'une goutte. On sait en effet que chaque bille consiste en des centaines de millions de molécules de code-barres, et on a évalué que 95% d'entre eux en moyenne sont identiques, il reste donc 5% de codes-barres différents par bille. Ils peuvent également être le fruit d'ARNs flottants ayant été retrotranscrits dans une goutte contenant une bille à code-barres mais pas de cellule. Enfin, ils peuvent apparaître suite à des erreurs survenant lors de l'étape d'amplification ou encore de séquençage. Ces artefacts correspondent à des faux codes-barres et ne vont compter que très peu de *reads*. Il est donc possible de les éliminer en supprimant les codes-barres contenant moins de *reads* qu'un seuil limite qui est déterminé à partir de l'histogramme de la fréquence d'occurrence du nombre de *reads* par code-barres. En théorie, deux populations distinctes sont définies. Cependant, nos échantillons ayant été séquencés avec très peu de profondeur, il est plus difficile d'isoler les vrais codes-barres, car ceux-ci ne comptent que peu de *reads* (Figure 57, A). Nous avons essayé de définir des seuils à partir des histogrammes tracés, représentés sur les graphes par une flèche rouge.

Afin de vérifier si ce seuil est correct, nous avons tracé une matrice de corrélation pour chaque échantillon. Elle permet de voir quels segments sont associés entre eux, c'est-à-dire fréquemment trouvés au sein de la même goutte, et lesquels sont au contraire anti-corrélés, et très peu rencontrés ensemble dans une même goutte. En théorie, les segments associés à une souche devraient être corrélés entre eux mais décorrélés de ceux provenant de l'autre souche. Ces matrices ont été calculées après application du filtre. Une corrélation des segments en fonction de leur origine est observée. L'information cellule unique semble donc avoir été majoritairement conservée dans tous les cas de figure, et si des échanges se sont produits à l'étape d'encapsulation ou d'amplification, ils restent minimes. La matrice de corrélation des codes-barres ayant été négativement filtrée a également été calculée pour l'échantillon i05. La

corrélation entre segments d'une même souche n'est pas conservée, signe que ces codes-barres ne sont pas consistants et peuvent donc être supprimés.

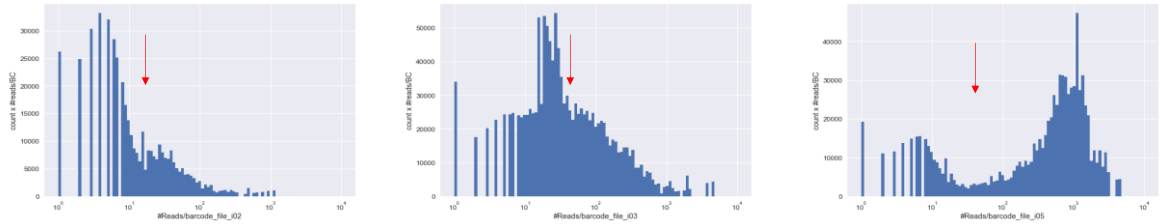
Finalement, le nombre de codes-barres total, correspondant au nombre de cellules encapsulées, est un bon indicateur de la justesse du filtre appliqué. Environ 1 5000 à 2000 cellules ont été encapsulées par émulsion. Les échantillons i02 et i03 correspondent au mélange des émulsions D1 et D2, on s'attend donc à retrouver 3 000 à 4 000 codes-barres. L'échantillon i05 quant à lui correspond à l'émulsion D4 seule, on espère retrouver 1 500 à 2 000 codes-barres. Le nombre de codes-barres avant filtration est bien au-delà de celui attendu, en revanche, il est similaire au chiffre attendu après application du filtre (Figure 57, D). Le nombre de cellules encapsulées a été estimé de manière approximative, en mesurant le temps de collection et en tenant compte de la probabilité de distribution des cellules selon une loi de Poisson, connaissant la densité de cellules. Le taux d'encapsulation effectif des billes et des cellules a été vérifié en procédant à un comptage sur vidéos prises à l'aide d'une caméra à haute vitesse pendant chaque manipulation microfluidique. Le marquage des cellules et des billes avec des fluorophores aux spectres séparés permettrait un suivi en temps réel plus précis du nombre de cellules encapsulées. Nos stations microfluidiques sont en effet équipées de 4 lasers, capables d'illuminer tout le contenu d'une goutte en une ligne, et dont la fluorescence réémise est renvoyée sur des photomultiplicateurs. Les gouttes contenant une bille et une cellule vont donc émettre un signal spécifique, qui peut être mesuré en temps réel, de sorte à déterminer avec précision la proportion de gouttes contenant une cellule, ou une bille, ou les deux.

i02
Emulsions D1 et D2
regroupées après
amplification

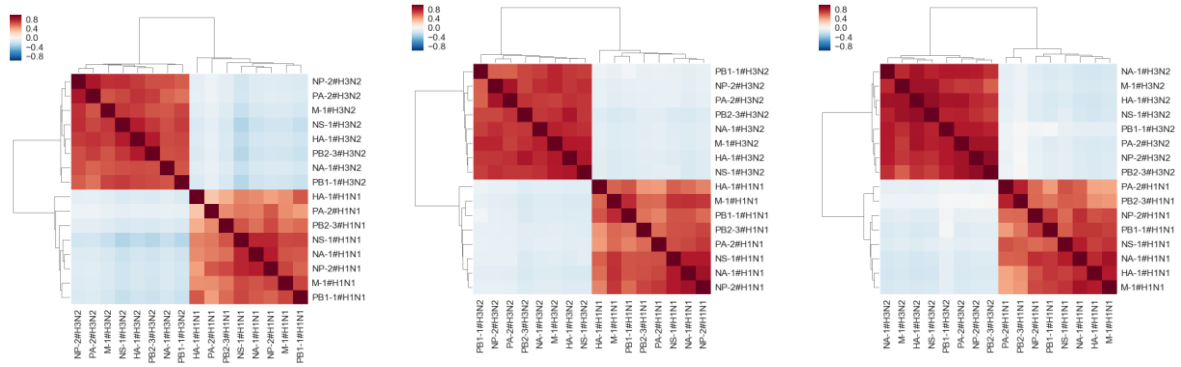
i03
Emulsions D1 et D2
regroupées avant
amplification

i05
Emulsions D4
Cellules H1 et H2
co-encapsulées

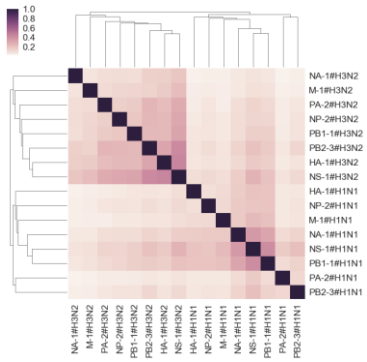
A



B



C



D

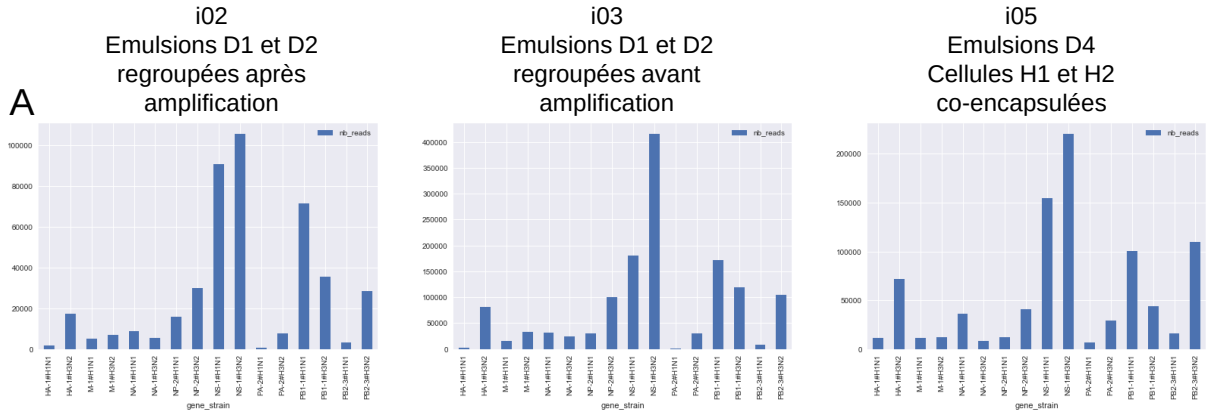
	i02	i03	i05
Filtre appliqué (nombre de reads/BC)	20	50	50
Nombre de BC avant filtre	84 844	109 087	50 595
Nombre de BC après filtre	3004	4214	1627

Figure 57 Filtrations des données de séquençage; A = Histogramme de la fréquence du nombre de reads par BC pour les échantillons (de gauche à droite) i02, i03 et i05, la flèche rouge représente le filtre appliqué, B = Matrice de corrélation à l'échelle de la goutte unique entre les différents segments après application du filtre sur le nombre de reads par BC pour les échantillons (de gauche à droite) i02, i03 et i05, C = Matrice de corrélation de l'échantillon i05 sur les BC contenant moins de 50 reads (soit les BC étant déplétés après application du filtre), D = tableau récapitulatif du nombre de BC identifiés avant et après application du filtre sur le nombre de reads par BC, Légende : BC = code-barres

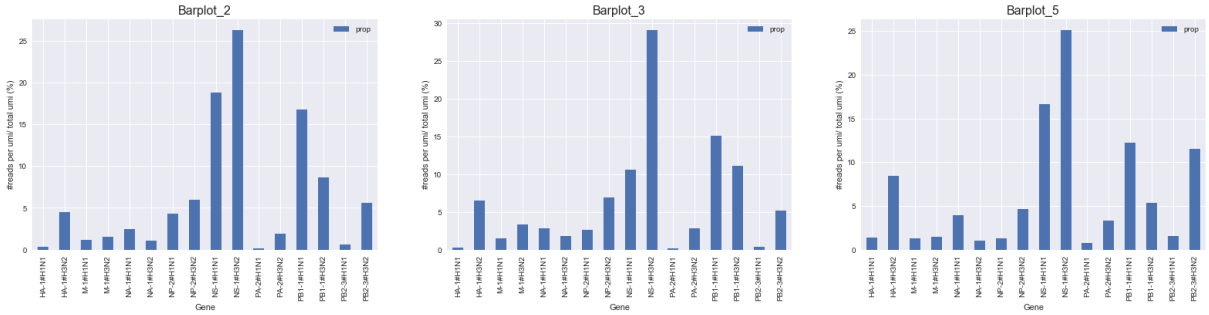
B. Analyses moyennes

Une fois les données filtrées, des analyses moyennées à partir des données de RNAseq ciblé en gouttes de chaque échantillon ont été faites, de sorte à vérifier si elles sont comparables à celles observées en tube, et donc si la RT en amorces spécifiques fonctionne efficacement en gouttes. Nous avons constaté les mêmes problèmes en gouttes que ceux observés en tube, à savoir une efficacité de RT non homogène d'un segment à l'autre, voire ne fonctionnant pas pour certains segments, alors même que le nombre de chaque segment dans les cellules est supposé entre homogène. Les analyses moyennes ont été faites à la fois sur le nombre de *reads* ou le nombre d'UMI. La quantification grâce à l'UMI permet en théorie de s'affranchir des biais éventuels d'amplification. Cependant, nous avons constaté que le séquençage trop peu profond de nos échantillons ne nous permet pas d'utiliser convenablement cet outil.

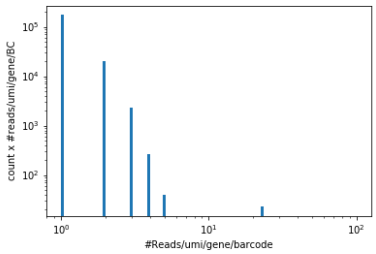
A



B



C



D

Echantillon	Nombre de BC	Nombre total de reads	Nombre total d'UMI
i02	3 004	123 668	116 056
i03	4 214	501 926	339 086
i05	1 627	664 710	619 128

Figure 58 : Analyses moyennes des données de séquençage des échantillons issus des manipulations en gouttes, A = Répartitions des reads entre les différents gènes ciblés pour chacune des souches, dans les différents échantillons séquencés, B = Proportion d'UMI comptés pour chaque segments cibles par rapport au nombre total d'UMI dans l'échantillon, C = Histogramme de la fréquence du nombre de reads par UMI dans l'échantillon i02, D = Tableau récapitulatif du nombre total de BC, reads et UMI dans chacun des 3 échantillons après avoir appliqué un filtre sur le nombre de reads par BC, Légende : BC = code-barres

Les résultats présentés en Figure 58 montrent la répartition des *reads* ou des UMI entre les différents segments dans chaque souche, et ce pour les 3 échantillons séquencés. Ils sont très similaires d'un échantillon à l'autre, preuve que le mélange entre les 2 souches à chaque niveau du processus a été fait de manière similaire. En s'intéressant à la répartition des *reads* entre les différents segments, on constate le même déséquilibre qu'observé *in vitro,* à l'avantage de la souche H3N2, hormis pour le segment PB1, qui est plus fortement capturé et amplifié chez la souche H1N1. De même, certains segments sont sous-représentés, c'est le cas des segments M1, NA et PA (pour les 2 souches), ou HA, NP et PB2 (pour la souche H1N1), tandis que d'autres sont surreprésentés (segment NS pour les 2 souches). Ces résultats sont logiques ; ces amorces ont déjà montré des faiblesses *in vitro*, et vont nécessiter la conception de nouvelles amorces. Les analyses sur la distribution des *reads* ou des UMI sont semblables. Cela n'est pas surprenant, car la plupart des UMI ne sont représentés que par un seul *read*, comme le montre l'histogramme de la fréquence du nombre de *reads* par UMI dans l'échantillon i02 (Figure 58,C). Le tableau D de la Figure 58 confirme cette observation ; les nombres de *reads* totaux et d'UMI totaux sont sensiblement identiques, et ce pour les 3 échantillons.

Finalement, chaque cellule ne compte que 40 à 400 *reads* (selon l'échantillon), et les UMI ne comptent en majorité qu'un *read*. Avec une profondeur si faible, la correction des biais d'amplification grâce aux UMI n'est pas envisageable. Il serait intéressant de séquencer plus profondément ces échantillons afin de vérifier si les déséquilibres observés entre segments peuvent être réduits et ainsi définir si l'origine de ce déséquilibre est liée à un biais d'amplification.

C. Identifications de sous-populations

1. Résumé des résultats obtenus

Les analyses à l'échelle de la cellule unique vont nous permettre de répondre à plusieurs questions ;

- Est-on capable d'identifier les 8 segments dans chaque cellule unique ?
- L'origine des segments est-elle correctement identifiée et les segments d'origines différentes sont-ils bien décorrélés ?
- Observe-t-on plus de code-barres avec des segments des 2 origines lorsque les échantillons ont été mélangés avant amplification, signe d'échanges intervenant lors de cette étape ?
- Quelle proportion de codes-barres est associée aux 2 souches en même temps dans le cas où les cellules H1 et H3 ont été mélangées avant encapsulation ? Ce chiffre correspond-il aux estimations faites d'après la loi de Poisson ?

Les analyses à l'échelle de la cellule unique ont ainsi montré que les cellules se divisent en 2 groupes, dont chacun est associé à une souche différente. En revanche, les 8 segments ne sont pas détectés dans la plupart des cellules. La distinction inter souche de certains segments est d'ailleurs ambiguë. L'échantillon i03, correspondant au cas où les émulsions des cellules H1 et H2 ont été mélangées avant amplification, ne montre pas de cellules associées aux 2 souches à la fois. L'étape d'amplification n'entraîne donc pas la formation d'hybride, par réaction de *template switching* par exemple. L'échantillon i05, où les deux cellules H1 et H3 ont été co-encapsulées montrent la présence d'une faible proportion de codes-barres pour lesquels des segments provenant des deux 2 souches sont trouvés ; ils correspondent à des gouttes où les 2 cellules ont été co-encapsulées.

2. Analyse à l'échelle de la cellule unique

La matrice d'expression a été représentée graphiquement sous forme de C*lustermap* (Figure 59), après normalisation des lignes et colonnes, afin d'homogénéiser les différences inter cellules et inter segments. Le *clustermap* consiste en une carte thermique (*heatmap*) combinée à un dendrogramme, permettant à la fois de visualiser une matrice d'expression en niveau de couleurs tout en ajoutant une analyse par groupes, pour ordonner les données en fonction des similitudes entre des groupes de

cellules. Chaque ligne représente une cellule tandis que chaque colonne représente un segment. La normalisation des colonnes est essentielle à la lisibilité du *heatmap*, du fait de la forte variation dans la détection des différents segments, liés à des biais techniques, comme expliqué précédemment. La normalisation en fonction des cellules est classique, et permet de s'affranchir des biais techniques qui créent une forte hétérogénéité d'une cellule à l'autre. Ces normalisations sont d'autant plus nécessaires que l'utilisation des UMI pour réduire les biais d'amplification n'est pas efficace, du fait de la trop faible profondeur de séquençage.

On observe la répartition des cellules en 2 groupes distincts majoritaires pour les 3 *clustermaps* issus des données provenant des 3 échantillons séquencés. Ces groupes sont chacun associés à une souche particulière. Les segments sont également regroupés en fonction de la souche d'origine, confirmant les observations faites précédemment, lors de l'analyse des matrices de corrélation (Figure 57).

L'échantillon i02 est ici la référence, car le mélange entre les émulsions de chacune des 2 souches (D1 et D2) n'a été fait qu'en fin de processus, soit après quantification des librairies finales issues de la préparation des produits de synthèse des 2 émulsions. Le *clustermap* issu de cet échantillon ne montre en effet pas de codes-barres (lignes) avec une détection de segments provenant des 2 souches. On observe des trous dans les groupes dessinés, en particulier dans le cluster de la sous-population formée par les cellules associées à un phénotype H1N1. Ils correspondent à la non-détection de certains segments ; HA, PA et PB2 pour la souche H1N1 et dans une moindre mesure le segment NA pour la souche H3N2, soit les mêmes segments déjà pointés pour leur faiblesse grâce aux analyses précédentes. Finalement, même après application des normalisations, la répartition des segments dans chaque cellule n'est pas homogène et les 8 segments ne sont pas forcément tous détectés dans une cellule donnée, en particulier dans le sous-groupe associé à la souche H1N1, du fait de failles techniques. Cette observation est d'ailleurs valable pour les 3 *clustermaps*. Une optimisation est primordiale car la prédiction du réarrangement entre segments provenant de 2 souches différentes nécessite la détection de tous les segments dans toutes les cellules analysées.

L'échantillon i03 correspond au cas où les émulsions de chaque souche encapsulée individuellement ont été mélangées avant amplification. Le *clustermap* est très similaire à celui de référence. On ne note pas l'apparition d'un sous-groupe de cellules

co-détectant les segments H1 et H3. Il n'y a donc pas d'échange notable entre molécules, par une réaction de *template switching* par exemple, lors de l'étape d'amplification.

L'échantillon i05 correspond au cas où les cellules infectées par chacune des souches ont été co-encapsulées. On note l'apparition d'un nouveau sous-groupe de cellules, associé à la co-détection de segments des 2 origines. Ce sous-groupe représente environ 8% des cellules, soit plus que le taux de co-encapsulation attendu avec un lambda de 0,1 qui est de moins de 1%. Cela peut s'expliquer par la présence de blocs cellulaires, observés sur les vidéos prises lors de l'encapsulation. Ces agrégats cellulaires pourraient être limités en passant les cellules sur filtre ou encore en procédant à un tri des cellules par FACS avant encapsulation.

En conclusion, la comparaison des échantillons i02 et i03, correspondant au cas où les émulsions de chaque souche unique ont été mélangées respectivement après ou avant les étapes d'amplification, a permis de montrer que l'étape d'amplification, après mises en commun du matériel provenant de toutes les gouttes, n'entraine pas l'apparition d'hybrides contaminants. On a en effet pu constater que les 2 échantillons i02 et i03 donnent des résultats très similaires, et montrent qu'il n'y a pas création d'espèces contaminantes lors de l'étape d'amplification. La comparaison des échantillons i02 et i05, correspondant au cas où les 2 souches ont été encapsulées séparément puis mélangées avant séquençage et au cas où les 2 types de cellules ont été encapsulées ensemble, a permis d'estimer la proportion de code-barres associés aux deux souches, et ainsi le taux de co-encapsulation. Cette proportion, bien que faible, est légèrement supérieure à celle attendue du fait du phénomène d'encapsulations multiples, directement lié à la loi de Poisson. Cela est certainement lié à la présence d'agrégats dans la phase cellulaire.

i02
Emulsions D1 et D2
regroupées après
amplification

i03
Emulsions D1 et D2
regroupées avant
amplification

i05
Emulsions D4
Cellules H1 et H2
co-encapsulées

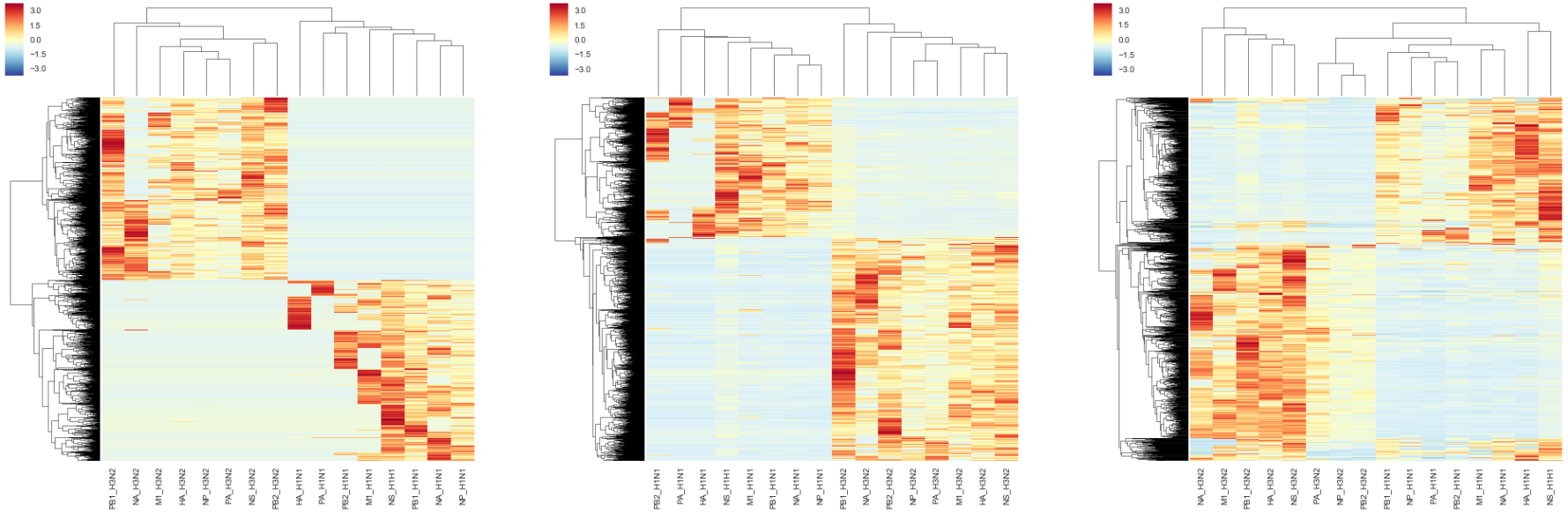


Figure 59 Identification de sous-populations cellulaires en utilisant l'outil clustermap sur les 3 échantillons séquencés, correspondant à une carte thermique permettant de représenter une matrice d'expression en niveau de couleurs, à laquelle a été ajoutée une dimension d'analyse par groupe sous forme de dendrogramme. Le clustering a été fait sur les différentes matrices d'expression après avoir normalisé l'expression des différents segments (normalisation par colonne) de même que les cellules entre elles (normalisation par ligne). Il permet d'identifier des sous-groupes de cellules partageant des profils d'expression similaires. Deux sous-groupes, associés à l'expression des segments de chaque souche, sont observés pour les 3 échantillons séquencés.

V. Conclusions et perspectives

Cette expérience préliminaire a montré la faisabilité du projet et a validé les premières étapes de la preuve de concept du développement d'une technologie à ultra haut débit pour l'étude du réassortiment génétique entre 2 souches virales. Les segments viraux sont détectables à l'échelle de la cellule unique et leur origine est identifiable. Il n'y a pas d'échange d'information entre molécules à code-barres après synthèse de l'ADNc à partir des amorces de RT. L'utilisation de notre technologie de séquençage de l'ARN ciblé à ultra haut débit faisant appel à la microfluidique en gouttes est donc adaptée au projet de prédiction des phénomènes de réassortiments génétiques entre différentes souches virales.

Certains points encore sont à améliorer, et notamment la conception de certaines amorces, ne fonctionnant ni en tubes, ni en gouttes. Il est primordial de détecter tous les segments au sein d'une goutte afin d'identifier de nouveaux génomes viraux, fruit du réassortiment entre deux souches. De même, il est important de contrôler l'agrégation des cellules hôtes. L'expérience où deux types de cellules hôtes, chacune infectée par une souche différente, ont été mélangés avant émulsification a montré un taux de co-encapsulation plus grand que ceux attendus d'après la loi de Poisson. Les vidéos prises au cours de l'encapsulation ont en effet montré la présence d'agrégats cellulaires. Finalement, on a pu constater un fort déséquilibre dans l'efficacité de détection des différents segments, alors qu'elle devrait être homogène. Hormis le problème lié à l'inefficacité de certaines amorces, on peut supposer qu'un séquençage avec plus de profondeur permettrait de corriger les éventuels biais d'amplification en utilisant le comptage des UMI. De même, une stratégie d'amplification linéaire par transcription *in vitro* pourrait être envisagée.

Enfin, certains aspects n'ont pas encore été abordés. La profondeur minimale de séquençage nécessaire pour la détection de tous les segments simultanément devra être définie, après avoir corrigé les points décrits dans le paragraphe précédent. Ce seuil nous donnera ainsi une idée du nombre maximal de cellules qui peuvent être séquencées en parallèle et donc du débit d'analyse. La prochaine étape est de travailler sur des cellules hôtes après co-infection. La première étape implique une infection à haut MOI de cellules hôtes par 2 souches en simultanée, de sorte à obtenir des cellules ayant été infectées par 2 souches différentes. Les virus issus de ces

cellules sont potentiellement génétiquement réassortis. Une seconde étape d'infection à bas MOI cette fois, permet d'infecter une cellule hôte par un seul virus et ainsi rechercher la présence d'éventuels virus hybrides et de quantifier ces sous-populations. L'infection à MOI bas est obligatoire mais implique que beaucoup de cellules hôtes ne soient pas infectées. Il convient donc de mettre en place une stratégie de tri afin de n'encapsuler que les cellules hôtes infectées.

Conclusions générales et perspectives

Nous avons développé et optimisé une technologie de RNAseq ciblé en gouttes, à l'échelle de la cellule unique, que nous avons appliqué à l'étude de l'inflammation chez des cellules microgliales ou à l'étude du réassortiment génétique chez les virus Influenza de type A. Une autre application est proposée en annexe ; il s'agit d'une étude *in vitro* sur l'évolution.

Deux étapes clés du protocole ont été optimisées, à savoir la capture des ARN cible puis leur RT en ADNc et l'amplification des ADNc (Kolodziejczyk, Kim, Svensson, Marioni, & Teichmann, 2015a). Plusieurs paramètres ont été testés de manière systématique pour l'amélioration de l'efficacité de la RT, tels que le tampon de lyse à utiliser ou encore la température d'incubation lors de la RT, jusqu'à identifier un point déterminant : le nombre d'amorces spécifiques présentes dans le compartiment où doit se faire la RT. Les billes d'hydrogel, sources des amorces à code-barres dans la goutte, ont donc naturellement été optimisées, afin d'augmenter leur capacité fonctionnelle. Deux méthodes d'amplification des ADNc ont ensuite été comparées ; l'amplification linéaire par IVT ou exponentielle par PCR. Tandis que l'IVT offre l'avantage de limiter les biais d'amplification (Hashimshony *et al.*, 2012), comme démontré lors de nos expériences d'optimisations par RT-qPCR en amorces spécifiques en tubes en comparant le profil d'expression de plusieurs cibles avant ou après amplification par PCR ou IVT, la PCR est quant à elle plus sensible et spécifique. Les expériences à l'échelle de la cellule unique nous ont permis de mettre en évidence les avantages et inconvénients offerts par chacune de ces 2 méthodes.

<u>La Microglie</u>

L'application à l'étude de l'inflammation chez la microglie a dans un premier temps permis de valider les optimisations faites en gouttes, en utilisant une lignée modèle de cellules microgliales souris, la lignée BV2, et en mimant une inflammation par ajout de LPS dans le milieu cellulaire. 14 marqueurs de l'inflammation ont été choisis d'après des analyses d'expression différentielles lors d'un état d'inflammation à partir de données en masse sur puces à ADN ou en RT-qPCR. Des amorces ont été conçues pour ces cibles et validées en tube avant de procéder aux expériences de RNAseq

ciblé en gouttes. Des expériences en tubes ont été réalisées en parallèle des expériences en gouttes, à partir d'ARN extrait des mêmes cellules que celles utilisées pour les encapsulations. Les résultats en masse sont similaires aux résultats moyens à l'échelle de la cellule unique et sont biologiquement cohérents, montrant une surexpression de marqueurs typique de ce type de stimulation au LPS.

Différents protocoles de RT et d'amplification ont ensuite été comparés en utilisant le modèle des cellules BV2 activées au LPS, afin d'induire l'expression des marqueurs de l'inflammation. Le RNAseq ciblé en cellule unique a été fait dans des gouttes de 2nL avec des billes de 100pL ou dans des gouttes de 100pL avec des billes de 10pL, dont la capacité fonctionnelle est 3 fois plus petite que celles des billes de 100pL. Les produits de synthèse de ces RT ont ensuite été amplifiés selon 2 stratégies possibles, une amplification linéaire par PR multiplex ou une amplification linéaire par IVT.

Ces expériences d'optimisation nous ont permis d'affiner le protocole, montrant que travailler dans des gouttes de plus grand volume avec des billes à la meilleure capacité fonctionnelle permet d'augmenter l'efficacité réactionnelle en goutte. De même, la préparation des librairies de séquençage et en particulier la méthode utilisée pour amplifier le matériel produit lors de la réaction en gouttes a un réel impact sur les résultats observés ; tandis qu'une amplification linéaire par IVT permet de réduire les biais d'amplification, une amplification par PCR semble plus sensible et spécifique.

L'amplification par PCR montre en effet de meilleur taux d'alignement lors de l'identification des gènes et fonctionne sur des échantillons pauvres en matériel. En revanche, une telle amplification semble créer de nombreux biais. Pour un même échantillon de départ, 10 fois plus de molécules sont quantifiées par goutte lorsque l'amplification se fait par PCR contre une amplification par IVT. 2 hypothèses s'offrent alors ; soit l'amplification très sensible par PCR assure moins de perte d'informations, soit le grand nombre de cycles a créé de faux UMI ; cette seconde hypothèse peut être vérifié en améliorant les scripts d'analyse et en utilisant des outils du type UMI Tools pour faciliter la filtration des faux UMI, issus d'erreurs survenant lors des cycles d'amplification répétés (Smith *et al.*, 2017).

Nous avons finalement défini un protocole optimal de RNAseq ciblé en gouttes ; la capture des ARNm puis leur RT est effectuée sur des billes de 100pL et dans des gouttes de l'ordre du nanolitre, tandis que l'amplification se fait selon une méthode de

PCR multiplex, en n'oubliant pas la nécessité d'implémenter nos scripts d'analyse avec une étape de filtration des UMI.

En utilisant ce protocole, nous sommes parvenus à détecter nos transcrits d'intérêt à l'échelle de la cellule unique. Une expression bimodale au sein de la population cellulaire est observée pour la plupart d'entre eux, où une sous population de cellules exprimant un gène cible à un niveau significatif de plus de 20 transcrits typiquement se distingue de la population vide. Nos analyses à l'échelle de la cellule unique ont par ailleurs montré que la majorité des transcrits semblent avoir une expression éclatée, ce qui, conjointement avec le nombre restreint de cibles étudiées en parallèle, rend difficile l'observation de corrélation d'expression ou l'identification de sous-groupes de cellules (Buettner *et al.*, 2015).

Alors que les technologies scRNAseq ont vu le développement de divers outils, tels que l'outil MAGIC (van Dijk *et al.*, 2017) pour corriger les « faux zéros » des matrices d'expression en comblant les vides d'après l'appartenance d'une cellule à un sous-groupe de cellules associées à une certain motif d'expression, il est difficile à notre niveau de faire la distinction entre réalité biologique ou bruit technique. Une technologie scRNAseq ciblant le transcriptome entier est finalement plus adaptée à l'identification de groupes de cellules et du motif d'expression qui leur sont associées dans le cadre d'une étude sans à priori, d'autant plus si les différences entre populations sont visibles à travers l'expression variables d'un grand nombre de gènes (de l'ordre du millier). Ce type d'étude pourra être fait an amont, de sorte à identifier des cibles d'intérêt et pouvoir les étudier de manière plus fine à l'aide d'une technologie ciblée, permettant d'augmenter la couverture pour les gènes d'intérêt (Torre *et al.*, 2018). De la même manière que la technologie ciblée à l'échelle de la cellule unique CytoSeq (H. C. Fan *et al.*, 2015), augmenter le nombre de transcrits analysés en parallèle à une centaine environ devrait aider à identifier des groupes de cellules, tout en se focalisant sur l'étude de marqueurs de l'inflammation par exemple.

Notre technologie apporte de la finesse dans l'analyse et nous a permis d'étudier la dynamique d'expression de certains gènes, tels que Il1Rn ou Il1b, de manière qualitative, à l'échelle de la cellule unique. Les données obtenues à partir du RNAseq ciblé sur des cellules BV2 activées au LPS à différents points temps devront être analysées plus en profondeur afin d'analyser les trajectoires d'expression des marqueurs de l'inflammation ciblée de manière quantitative, et de définir

éventuellement des paramètres cinétiques de la réponse dynamique impliquée dans l'inflammation. Ce type d'analyse cinétique ciblée à l'échelle de la cellule unique a déjà été publié en utilisant la technologie smRNA FISH (Schwabe & Bruggeman, 2014). Notre technologie offre ici la possibilité d'augmenter fortement le débit d'analyse de même que le nombre de cibles analysables en parallèle.

La prochaine étape va être de caractériser les performances de notre technologie. Tout d'abord, d'autres analyses plus poussées devraient également aider à définir la profondeur de séquençage suffisante pour significativement quantifier l'expression d'un gène cible en échantillonnant les données de manière graduelle. Ensuite, une comparaison de la couverture obtenue pour une cible d'intérêt à l'échelle de la cellule unique à l'aide de notre technologie ciblée par rapport à une technologie d'analyse du transcriptome entier du type inDrop (Klein *et al.*, 2015)(Zilionis *et al.*, 2017) sur un même échantillon cellulaire. Nous avons déjà réalisé ces expériences en utilisant le kit commercial de 1CellBio, basé sur la technologie inDrop. Les données de séquençage doivent encore être analysées et pourront servir de point de comparaison avec celles obtenues à partir de notre RNAseq ciblé sur cellules BV2 activées au LPS ou non. Une comparaison avec des données de séquençage sur le même type cellulaire et le même type de traitement (LPS) mais faisant appel à d'autres technologies pourront également être faites (Sousa *et al.*, 2018)(Das *et al.*, 2016). Le mix commercial d'ARN synthétiques ERCC Spike In est également un bon outil de comparaison permettant d'évaluer le gain dans la détection d'une cible d'intérêt en utilisant une technologie ciblée plutôt qu'une technologie faisant appel à des amorces polyT. C'est d'ailleurs cette stratégie qui a été employée par Mercer et son équipe pour valider leur technologie de RNAseq ciblé CaptureSeq (Mercer *et al.*, 2014). Cette expérience sur le mix ERCC Spike-In devrait également permettre de la sensibilité de notre technologie, définie par sa limite de détection, de même que l'efficacité de capture de cibles en fonction de leur niveau d'expression et la reproductibilité, d'une cellule à l'autre, définissant le bruit technique associé à une capture variable d'une goutte à l'autre (Svensson *et al.*, 2017).

Le virus Influenza de type A

Nous avons ensuite appliqué notre technologie de RNAseq ciblé à l'échelle de la cellule unique pour l'étude du réassortiment génétique chez IAV. Une preuve de concept à partir de 2 souches de virus IAV humain, H1N1 et H3N2, a montré que notre technologie permet de détecter une combinaison de segments viraux à l'échelle de la cellule unique et d'identifier la souche d'origine.

Nous avons constaté que l'aspect cellule unique est conservé tout le long du processus, allant de l'encapsulation des cellules infectées avec un type de souche, au séquençage des échantillons, en passant par une étape d'amplification par PCR des produits de RT en gouttes par PCR. Nous sommes ainsi parvenus à regrouper des cellules infectées par chaque souche (une seule souche à la fois) en fonction de la souche d'origine.

Quelques améliorations techniques sont à prévoir ; certains segments ne sont pas correctement détectés. La conception de nouvelles amorces pour ces segments devrait améliorer la détection de ces cibles. De plus, les échantillons ont été sous-séquencés, ne permettant pas d'utiliser les UMI pour corriger les biais d'amplification (Islam *et al.*, 2014) et ayant pu conduire à une perte d'information pour les segments étant amplifiés avec une efficacité moindre. Tous les segments sont normalement équimolaires dans la cellule hôte, à une quantité de l'ordre de 10^4 à 10^5 copies par cellule. Cependant, on observe un fort déséquilibre dans la détection des différents segments. L'utilisation d'une stratégie d'amplification par IVT pourrait être utile à cette étude, étant donné que tous les segments sont présents à une quantité initiale suffisante et homogène d'un segment à l'autre.

Le principal avantage offert par notre technologie est l'ultra haut débit. Dans le cadre de cette étude, 10^5 cellules hôtes infectées doivent être collectées pour assurer une couverture significative des différentes nouvelles combinaisons de segments créées par réassortiment génétique. Nous avons pu constater que l'efficacité de la RT sur des cellules microgliales est assez similaire dans des gouttes de 2nL ou de 100pL en ce qui concerne la détection de transcrits fortement exprimés. Etant donné les objectifs de collection et la fréquence d'encapsulation 10 fois plus grande dans le cas des gouttes de 100pL, le protocole à favoriser est celui d'un RNAseq ciblé sur billes de

10pL à amorces spécifiques à code-barres, suivi par une amplification linéaire par IVT ou PCR ;

Une fois les problèmes techniques résolus, et tous les segments viraux étant efficacement détectés, la suite de la preuve de concept à partir des 2 souches de virus humain pourra être réalisée, en effectuant une expérience de co-infection par les 2 souches et en quantifiant l'apparition de nouveaux virus issu d'un réassortiment génétique entre les 2 souches, à partir de l'encapsulation de quelques 10^5 cellules hôtes, présélectionnées selon sur le critère de co-infection par les 2 souches.

Notre technologie a donc le potentiel de détecter des transcrits non polyadénylés chez d'autres types d'organismes que les cellules eucaryotes.

Nous souhaitons prochainement appliquer notre technologie à l'étude des voies de régulations chez Escherichia Coli. La lyse des membranes bactériennes en gouttes pour réaliser une RT directe sur cellule a déjà été mise au point. Le prochain défi à résoudre est d'augmenter le nombre de copies d'ARN bactérien gouttes en les faisant se multiplier avant de procéder au RNAseq ciblé sur plusieurs gènes entrant en jeu dans les différentes voies de régulation de la bactérie.

Annexe 1 = Séquences utiles

Séquences consensus	
Promoteur T7	TAATACGACTCACTATAGGG
M13	CAGGAAACAGCTATGACC
Amorces Cellules BV2	
Amorce RT BV2	
SF_5PiEt-L5-iNos_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN ACATTGATCTCCGTGACAGCC
SF_5PiEt-L5-Cox2_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN AAGCGTTTGCGGTACTCATT
SF_5PiEt-L5-CD32_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN CCAATGCCAAGGGAGACTAA
SF_5PiEt-L5-CD86_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GGCTCTCACTGCCTTCACTC
SF-5PiEt_L5_TNFa_1	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GCTACAGGCTTGTCACTCGAA
SF-5PiEt_L5_IL1b_R1	/5Phos/ CAAC TACGCTACGGAACGA NNNNN ATGTGCTGCTGCGAGATTTG
SF-5PiEt_L5_Cd14v1	/5Phos/ CAAC TACGCTACGGAACGA NNNNN TCCTTGCAGCTGTACCCTTG
SF-5PiEt_L5_Lyz2v2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GTGCTTTGGTCTCCACGGTT
SF_5PiEt-L5-Arg1_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GCCAGAGATGCTTCCAACTG
SF_5PiEt-L5-Lgals3_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN ATTGAAGCGGGGGTTAAAGT
SF_5PiEt-L5-Ccl2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN ATTCCTTCTTGGGGTCAGCA
SF_5PiEt-L5-Scl7a1	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GGCACTGACCAGCTCATTCT

SF_5PiEt-L5-Igf1_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GCAACACTCATCCACAATGC
SF_5PiEt-L5-Il1Rn_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN TTCTCAGAGCGGATGAAGGT
SF_5PiEt-L5-Il4Ra_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN ACTCTGGAGAGACTTGGTTGG
SF_5PiEt-L5-Rpl13a_BV2	/5Phos/ CAAC TACGCTACGGAACGA NNNNN GAGTCCGTTGGTCTTGAGGA
Amorce PCR sens BV2	
iNos_PCR2	TGA CAC ACA GCG CTA CAA CAT CC
Cox2_PCR2	TGG GGG AAG AAA TGT GCC AAT TG
CD32_PCR2	CTG AGG CTG AGA ATA CGA TCA CCT
CD86_PCR2	ACACGAG CGG GAT AGT AAC GCT GAC AGA G
Arg1_PCR1	TGT GAA GAA CCC ACG GTC TGT GG
Il1b_PCR1_f1	TGCCACCTTTTGACAGTGATGAGA
TNFa-1_PCR2	ATC GGT CCCC AAA GGG ATG AGA
Lgals3v2_PCR2	TTC AGG AGA GGG AAT GAT GTT GCC
Igf1_BV2	TGGATGCTCTTCAGTTCGTG
Il1Rn_PCR2	TGT CTT GTG CCA AGT CTG GAG ATG
Ill4Ra_BV2	GGA TAA GCA GAC CCG AAG C
Cd14_PCR1	GTTCCCGACCCTCCAAGTTTTAGC
Lyz2v2_PCR1	AACGTTGTGAGTTTGCCAGAACTC
Ccl2_PCR1	CCTGCTGCTACTCATTCACCAGC
Slc7a1_PCR1	TACTCTTTGGTGGCTGCCTGTG
SF_Rpl13a_PCR2	AGG CCA AGA TGC ACT ATC GGA AG
Adaptateurs Illumina	
Rd1	ACACTCTTTCCCTACACGACGCTCTTCCGATCT
Rd2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
P5	AATGATACGGCGACCACCGA
P7	CAAGCAGAAGACGGCATACGAGAT
P5-Rd1	AATGATACGGCGACCACCGA GATCT ACACTCTTTCCCTACACGACGCTCTTCCGATCT
P7-Rd2	CAAGCAGAAGACGGCATACGAGAT GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

Code-barres	
Structure (brin top)	/5Acryd/ TCTTCACGGAACGA - ACTG GACTAGAA TGATCA Rd2 - TTCG index B - TGAC index C - ACCA index D
Ac-DNA	
AcRanA	/5Acryd/ TCTTCACGGAACGA
5PiOb-RanA_rev	/5Phos/ CAGT TCGTTCCGTGAAGA
Ac-RanA_Extended	/5Acryd/ TCTTCACGGAACGA TGATCA CGATGACG TAATACGACTCACT
iOvb-RanA_Extended_rev	CTAT AGTGAGTCGTATTACGTCATCG TGATCA TCGTTCCGTGAAGA
Premier adapteur (différentes versions)	
5PiOt-RanRE-SpT7-Rd2	/5Phos/ ACTG GACTAGAA TGATCA CGATGACG TAATACGACTCACTATAGGG GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
SF_5PiOt-RanRE-SpT7-1	/5Phos/ ACTG GACTAGAA TGATCA CGATGACG TAATACGACTCACT
SF_5PiOv1t-RanRd2	/5Phos/ ATAGGG GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT
Sondes fluorescentes (Qualité contrôle billes)	
FAM-RanAas	/56-FAM/TCGTTCCGTGAAGA
FAM_Rd2s6as	/56-FAM/ AGCACACGTCTGAACTCCAGT
FAM_B18A	/56-FAM/BAAAAAAAAAAAAAAAAA
Pureté du code-barres	
GSP	/5Phos/ CA ACT ACG CTA CGG AAC GA NNNNNNNN CC GGT TTG TGT GAC TTT CGC CAC TC
ARN synthétique	CCT TGC GCC GGG AAACCAGTTG XXXX CAAC TGGTTTCC GAGTGGCGAAAGTCACACAAAC

(XXXX en 8 versions)	
PCR sens (XXXXX en 12 versions)	TTCCCTACACGACGCTCTTCCGATCT XXXXXX GAG CCT TGC GCC GGG AAACCA
ARN Synthétiques	
SF_Igf1_UM_EcoRI	ACAGCTGGACCAGAGACCCTTTGCGGGGCTGAGCTGGTGGATGCTCTTCA GTTCGTGTGTGGACCGAGGGGCT TTTA CTTCAAGAATTCCACAGGCTATGGCTCCAGCATTCGGAGGGCACCTCAGAC AGGCATTGTGGATGAGTGTTGC
SF_Rpl13a_UM_EcoRI	ACAGCCACTCTGGAGGAGAAACGGAAGGAAAAGGCCAAGATGCACTATCG GAAGAAGAAGCAGATCTTGAGG TTACGGAATTCGGCAGAAAAGAATGTGGAGAAGAAAATCTGCAAGTTCACA GAGGTCCTCAAGACCAACGGACTC
SF_Igf1_T7_forward	CTGCAGAATTC TAATACGACTCACTATA GG ACAGCTGGACCAGAGACCCTTTG
SF_Rpl13a_T7_forward	CTGCAGAATTC TAATACGACTCACTATA GG ACAGCCACTCTGGAGGAGAAACG
SF_Igf1_pT_reverse	TTTTTTTTTTTTTTTTTT GCAACACTCATCCACAATGC
SF_Rpl13a_pT_reverse	TTTTTTTTTTTTTTTTTT GAGTCCGTTGGTCTTGAGGA

Annexe 2

Trade-offs between Darwinian properties in autocatalytic RNA networks

Sandeep Ameta[1,‡], Simon Arsène[1,‡], Sophie Foulon[1], Baptiste Saudemont[1], Bryce E. Clifton[2], Andrew D. Griffiths[1,], Philippe Nghe[1,*]*

Affiliations

[1]Laboratoire de Biochimie, CNRS UMR8231, Chimie Biologie Innovation, ESPCI Paris, 10 Rue Vauquelin, 75005, Paris, France

[2]Department of Chemistry, Portland State University, Portland, Oregon 97207, USA

[‡]Sandeep Ameta and Simon Arsène contributed equally to the work.

Correspondence to: andrew.griffiths@espci.fr, philippe.nghe@espci.fr

Abstract

Evolution *via* template-based replication[1-4] was probably preceded by a more rudimentary form of evolution based on networks of autocatalytic reactions[5-8]. However, reaction networks possessing the Darwinian properties of variation (in composition, the relative fractions of catalysts), differential reproduction (accumulation of products), and heredity (persistence of composition), have so far not been identified. Here we show that networks of catalytic RNAs can possess certain properties of Darwinian systems, that these properties are controlled by network structure, and characterize important trade-offs between them. By combining barcoded sequencing with droplet microfluidics, we screened ~20,000 reactions corresponding to more than 1,800 distinct networks of ribozymes that catalyze their own formation from RNA fragments. We found that more highly connected networks tend to reproduce more quickly (accumulate more ribozymes) and be more robust to perturbations, indicating a trade-off between variation and reproduction. Variations are strongest when adding upstream ribozymes with novel reaction specificities (innovations) which target weakly connected networks. In turn, innovations increase connectivity, thus buffer against further perturbations, highlighting a second trade-off between robustness and variation. Despite this, phases of compositional robustness can alternate with sudden variations of the composition across trajectories of network growth by accretion of novel species. Finally, heredity is found to be limited by the activity of self-assembled, non-covalent, ribozymes. Our findings show that connectivity in reaction networks not only determines the probability of autocatalytic sets in chemistries[8], but also their potential for evolution. They are directly relevant to scenarios where early evolution is partly driven by environmental changes[9,10], as evolution of chemical compositions depends both on robustness (persistence being necessary for selection to act) and susceptibility to environmental changes (in order to explore novel states). More broadly, they provide guiding rules for chemistries capable of Darwinian evolution and constrain scenarios of the origins of life.

Main

In the prebiotic world, the spontaneous appearance of an RNA polymerase ribozyme (a replicase) with sufficient processivity to allow self-replication and enough fidelity to avoid an error catastrophe[5,6] seems unlikely, given that known replicases are long (>165 nt) and structurally complex[1-4]. However, theoretical studies suggest that earlier modes of evolution could have been supported by autocatalytic sets, where reproduction results from networks of more rudimentary catalysts[11-13]. Consistently, recent experiments indicate that RNA polymerase ribozymes can assemble from catalytic networks of RNA oligomers[14], suggesting that replicases may have emerged as components of such networks. Nevertheless, sustaining evolution in reaction networks is by no means trivial, as the Darwinian properties of variation, heredity and selection are mediated by chemical compositions (the proportion of different chemical species) rather than the copying of a sequence. Furthermore, the possibility to evolve chemistries may be constained by trade-offs between these properties. For instance, robustness to environmental perturbations and persistence of compositions are necessary for selection to act, but must be balanced with variation to explore novel states. So far, none of these Darwinian properties, nor their interplay, have been experimentally studied at a large scale in a prebiotically relevant system.

We studied an experimental model of autocatalytic RNAs derived from the group I intron of the *Azoarcus* bacterium[15]. Fragments (denoted WXY and Z) assemble into non-covalent complexes that catalyze the formation of more efficient covalent ribozymes[16,17] (denoted WXYZ, Fig. 1a), which in turn catalyze with higher efficiency the formation of further covalent ribozymes, allowing the formation of autocatalytic networks[16]. Three nucleotide long sequences, called internal guide (IGS) and target (tag), located at extremities of the WXY fragments[16] (Fig. 1a, top left), determine catalytic specificity by base-pairing between a ribozyme IGS and a fragment tag. Combinations of such fragments yield networks of diverse connectivity (Fig 1a, bottom center)[18]. These networks can be represented in a coarse-grained manner by directed graphs (Fig. 1a, top right): a node representing both non-covalent and covalent ribozyme species with the same IGS and tag, and a directed edge points from a ribozyme species to the ribozyme species whose formation it catalyzes.

We developed a method to generate and measure a high diversity of such RNA reaction networks, using droplet microfluidics coupled to barcoded Next-Generation Sequencing (Fig. 1b-c, Extended Data Fig. 1). It consists of first producing a library of 5 pL droplets containing 24 different combinations of the 16 WXY fragments (Extended Data Table 1), together with hairpin RNA reporters that do not react but enable later identification of the mixture of WXY fragments by sequencing. The 24 initial combinations were selected among a random sample by minimizing their mutual overlap while covering as evenly as possible network sizes from 1 to 16 species after combinatorial droplet fusion (see below and Methods). Random sets comprising 1 to 5 droplets of this initial library were then electrocoalesced[19] with a 50 pL droplet containing the reaction buffer and Z fragments. After incubation at 48°C for 1 hour, the composition of each droplet was analyzed by droplet-level RNA sequencing adapted from single-cell transcriptomics[20]. Hydrogel beads carrying cDNA primers with a barcode specific to each bead (Extended Data Fig. 2) are encapsulated one-by-one in droplets which are fused with the RNA containing droplets, where the primers are released and prime cDNA synthesis (Fig. 1c, Extended Data Fig. 1, and Methods). Barcoded cDNAs are recovered and sequenced, and *reads* from the same droplet carry the same barcode (Extended Data Fig. 3a, b).

The initial combinations of fragments (encoded by hairpin RNA reporters, Extended Data Fig. 3, 4), and the final fraction of ribozymes produced during incubation were determined in 20,038 droplets, comprising 1,837 unique networks with on average 11 replicates each (Fig. 1d, Extended Data Fig. 5a), indicating a mean precision of ~6% in species fractions (Extended Data Fig. 5b, c). Repeatability was tested with a full experimental replicate (r=0.84 between species fractions), and droplet cross-talk was quantified to be ~13% (Extended Data Fig. 5d-f).

We observed highly heterogeneous patterns of covalent ribozyme accumulation, demonstrating the existence of interdependences between reaction network components (Fig. 1d). Indeed, if every ribozyme species were reacting independently, the relative rank between any pair of ribozyme species would be conserved across all networks. On the contrary, we observed that for 63% of ribozyme pairs, their relative ranking differed in at least 10% of the networks (Extended Data Fig. 6a), extending previous findings of non-conserved ranking in a small set of such networks[21]. This diversity in species fractions was well predicted by a kinetic model (Methods)

integrating the effective catalytic rates of non-covalent and covalent ribozymes measured for individual IGS-tag pairs (r=0.83, $p<10^{-3}$, Fig. 1 e-f, Extended Data Fig. 6 b-d).

Comparing networks generated from distinct substrate sets also revealed large differences in growth, quantified as the concentration of covalent ribozyme accumulated during the reaction relatively to hairpin reporters of known concentration (Fig. 2a): there was on average a 62-fold difference in growth between the slowest and fastest network for each size category comprising more than 100 networks (sizes 3 to 10 ribozymes, see Extended Data Table 2).

We then quantified variations in response to the addition of WXY fragments allowing the formation of a novel ribozyme. This corresponds to comparing networks G and G' differing by a single node a (networks connected by a curved line in Fig. 1d). Denoting V the set of nodes common to G and G', the total perturbation is defined as $p_{G\to G'} = \sum_{v\in V} |y'_v - y_v|$ where y_v and y'_v are the respective fractions of species v in G and G', both normalized within V (Fig. 2b). By definition, the maximum is $p_{G\to G'} = 2$ and is reached for a full switch in species composition, for example going from $(y_u, y_v) = (0, 1)$ to $(y'_u, y'_v) = (1, 0)$ in $V = \{u, v\}$.

Averaging for each network the effect of all perturbations (green line, Fig. 2c,) revealed large differences from marginally to strongly perturbable networks (mean $\bar{p}_{G\to G'}$ from 0.1 to 1). Furthermore, almost none of the networks combined a large yield and a large perturbability (Fig. 2d), indicating a trade-off between growth and variation.

We aimed to interpret the diversity of observed responses based on network connectivity. The kinetic model does not provide such a direct interpretation as it consists of a mixture between two regimes, where either only non-covalent or only covalent ribozymes are active. For non-covalent ribozymes only, species fractions should be predicted by in-degree centrality, a network-theoretic measure which accounts only for catalysis by directly upstream ribozymes[22]. In contrast, for covalent ribozymes only, species fractions should be predicted by eigenvector centrality, which accounts for longer catalytic chains[22]. Despite these differences, the in-degree centrality was found to be a good approximation of the eigenvector centrality (for our dataset, R=0.83, p-value$<10^{-5}$, Extended Data Fig. 7a), as is al*read*y known in

general[23], except for strongly selfish ribozymes which fraction is underestimated (13% of the dataset, Extended Data Fig. 7b).

The in-degree centrality approximation allowed an analytical derivation of the perturbations (Supplementary File 1),

$$p_{G\to G'} = 2n\frac{1 - n\,\frac{m}{\sigma_G/e}}{\sigma_G/e + n} \quad (1),$$

with 4 control parameters (Fig. 3a): n is the *perturbation breadth*, the number of targets of the new node a introduced in the network; m is the *catalytic novelty*, the number of catalysts al*read*y present in G with the same IGS as a (novelty being higher for lower *m*); σ_G is the *background strength*, the sum of all edge weights in the network, and; e is the catalytic strength e of edges outgoing from a as determined by its IGS.

The influence of parameters σ_G, e, m, and n on $p_{G\to G'}$ predicted by the in-degree approximation (Equation 1) was similar to those predicted by eigenvector centrality (Fig. 3b), and, more importantly, were well verified experimentally (Fig. 3b, compare left and right panels). First, perturbation strength was impacted much more strongly by catalytic novelty m (Fig. 3b, bottom) than by perturbation b*read*th n for the values tested here (n = 1 or 2, Fig. 3b, top). Second, Equation (1) poses the condition $\sigma_G/e \leq n$ for significant variations, consistent with high values $p_{G\to G'}$ being observed only when $\sigma_G/e < 2$ (given that $n \leq 2$, Fig. 3b, top). Overall, the highest perturbations $p_{G\to G'}$ required a catalytic innovation ($m = 0$, i.e. a new IGS) to be combined with a low normalized background strength ($\sigma_G/e \leq 1$). Perturbations were indeed up to 10 times higher in such cases, compared to cases where either $m > 0$ or σ_G/e > 2 (Fig. 3b, bottom). Interestingly, the model indicated that perturbations should peak at intermediate values of σ_G/e when *m* > 0 (purple line, Fig 3b, right), which was consistently observed in the data (purple line, Fig 3b, left), despite the corresponding perturbation amplitudes being within the noise. The predictions of Equation 1 were also verified for other values of the parameters (Extended Data Fig. 8) and at the level of single nodes (Extended Data Fig. 9). In addition, the attenuation of perturbations with large σ_G/e combined with the weak but significant correlation between growth and σ_G (R=0.5, $p<10^{-5}$, Fig. 3c) partly explain the trade-off between growth and perturbability reported in Fig. 2d.

To test the interplay between robustness and variation in scenarios where novel species would appear either spontaneously[13] or due to changes in substrates provided from the prebiotic milieu[9,10], we analyzed cumulative perturbations across trajectories of network growth, starting from networks with three nodes, randomly adding one node at a time (Fig. 4a). We have seen that strong perturbations require catalytic innovations ($m = 0$). In such case, Equation 1 reduces to $p_{G \to G'} = 2/(1 + \sigma_G/ne)$. The ratio σ_G/ne expresses a second trade-off, between the robustness induced by the background strength of the network σ_G, and the variation induced by the novel catalyst as characterized by n and e.

As for a large perturbation to occur, ne must be comparable to the σ_G of the perturbed network, σ_G would roughly double after a strong perturbation, enhancing robustness to further variation. Consequently, strong variations should be followed by small ones, and result in inflexions (change in curvature) in cumulative perturbation trajectories. We quantified the corresponding inflexions by their sharpness (third derivative, Extended Data Fig. 10a), and categorized them as strong when in the top 25% of sharpness (Fig. 4b). Comparing the distributions of sharpness for all inflexion points showed that they are, as predicted, significantly sharper for catalytic innovations (Fig. 4c). Consistently, the number of strong inflexions correlates with the number of catalytic innovations per trajectory (Fig. 4d).

Although introducing a node causing a strong perturbation buffers the resulting network against further perturbations, subsequent variations along trajectories are still possible, as exemplified in Fig. 4ab. By computationally analyzing trajectories for an extended repertoire of specific interactions, we found that sustained variation requires perturbing species of increasing targeting b*read*th n (Fig. 4e) and increased waiting time between variation events (until the diversity of IGS/tag pairs saturates) Fig. 4f). The former allows the ever larger σ_G values to be overcome, while the latter corresponds to the build-up of weakly connected nodes that can become targets. Interestingly, the trade-off between σ_G and ne translates at the level of trajectories: chemistries with a high density of catalytic interactions lead to many trajectories with a few inflexions, while sparse chemistries lead to fewer trajectories but with many inflexions (Fig. 4g).

We have shown experimentally that RNA reaction networks can possess Darwinian properties, and that the structure of the catalytic networks imposes trade-offs between

these properties. What is missing, therefore, for such a system to evolve? Heredity, which in the context of autocatalytic networks depends on the persistence of the chemical composition over time is weakened due to catalysis by self-assembled, non-covalent, complexes. This leads to relaxation of chemical compositions toward states determined by the substrates rather than by the covalent catalysts transmitted from pre-existing networks (Extended Fig. 11). Reducing catalysis by self- non-covalent complexes may sufficiently enhance heritability for the system to evolve, if combined with selection and variation. On the other hand, self-assembled catalysts could facilitate the emergence of novel catalysts from variation in the substrates provided from environment. More generally, chemical networks that can maintain a balance between compositional robustness and variation would have been critical to allow the emergence of evolution.

References (ref 26-35 will be removed to methods)

1. Horning, D. P. & Joyce, G. F. Amplification of RNA by an RNA polymerase ribozyme. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9786-9791 (2016).
2. Johnston, W. K., Unrau, P. J., Lawrence, M. S., Glasner, M. E. & Bartel, D. P. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **292**, 1319-1325 (2001).
3. Kim, D. E. & Joyce, G. F. Cross-catalytic replication of an RNA ligase ribozyme. *Chem. Biol.* **11**, 1505-1512 (2004).
4. Wochner, A., Attwater, J., Coulson, A. & Holliger, P. Ribozyme-catalyzed transcription of an active ribozyme. *Science* **332**, 209-212 (2011).
5. Eigen, M. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465-523 (1971).
6. Eigen, M. & Schuster, P. The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercycle. *Naturwissenschaften* **64**, 541-565 (1977).
7. Hordijk, W., Steel, M. & Kauffman, S. The structure of autocatalytic sets: evolvability, enablement, and emergence. *Acta. Biotheor.* **60**, 379-392 (2012).
8. Kauffman, S. A. Autocatalytic sets of proteins. *J. Theor. Biol.* **119**, 1-24 (1986).
9. Abbott, S. S., Harrison, T. M., Schmitt, A. K. & Mojzsis, S. J. A search for thermal excursions from ancient extraterrestrial impacts using Hadean zircon Ti-U-Th-Pb depth profiles. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 13486-13492 (2012).
10. Stueken, E. E. *et al.* Did life originate from a global chemical reactor? *Geobiology* **11**, 101-126 (2013).
11. Jain, S. & Krishna, S. A model for the emergence of cooperation, interdependence, and structure in evolving networks. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 543-547 (2001).
12. Nghe, P. *et al.* Prebiotic network evolution: six key parameters. *Mol. Biosyst.* **11**, 3206-3217 (2015).
13. Vasas, V., Fernando, C., Santos, M., Kauffman, S. & Szathmáry, E. Evolution before genes. *Biol. Direct* **7**, 1; discussion 1 (2012).
14. Mutschler, H., Wochner, A. & Holliger, P. Freeze-thaw cycles as drivers of complex ribozyme assembly. *Nat. Chem.* **7**, 502-508 (2015).
15. Reinhold-Hurek, B. & Shub, D. A. Self-splicing introns in tRNA genes of widely divergent bacteria. *Nature* **357**, 173-176 (1992).
16. Draper, W. E., Hayden, E. J. & Lehman, N. Mechanisms of covalent self-assembly of the Azoarcus ribozyme from four fragment oligonucleotides. *Nucleic Acids Res.* **36**, 520-531 (2008).
17. Hayden, E. J., von Kiedrowski, G. & Lehman, N. Systems chemistry on ribozyme self-construction: evidence for anabolic autocatalysis in a recombination network. *Angew. Chem. Int. Ed. Engl.* **47**, 8424-8428 (2008).
18. Vaidya, N. *et al.* Spontaneous network formation among cooperative RNA replicators. *Nature* **491**, 72-77 (2012).
19. Chabert, M., Dorfman, K. D. & Viovy, J. L. Droplet fusion by alternating current (AC) field electrocoalescence in microchannels. *Electrophoresis* **26**, 3706-3715 (2005).
20. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187-1201 (2015).
21. Yeates, J. A., Hilbe, C., Zwick, M., Nowak, M. A. & Lehman, N. Dynamics of prebiotic RNA reproduction illuminated by chemical game theory. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5030-5035 (2016).
22. Yeates, J. A. M., Nghe, P. & Lehman, N. Topological and thermodynamic factors that influence the evolution of small networks of catalytic RNA species. *RNA* **23**, 1088-1096 (2017).
23. Valente, T. W., Coronges, K., Lakon, C. & Costenbader, E. How Correlated Are Network Centrality Measures? *Connect (Tor. Ont.)* **28**, 16-26 (2008).

24. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639-1645 (2009).
25. Arsene, S., Ameta, S., Lehman, N., Griffiths, A. D. & Nghe, P. Coupled catabolism and anabolism in autocatalytic RNA sets. *Nucleic Acids Res.* **46**, 9660-9666 (2018).
26. Duffy, D. C., McDonald, J. C., Schueller, O. J. & Whitesides, G. M. Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane). *Anal. Chem.* **70**, 4974-4984 (1998).
27. Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8**, 870-891 (2013).
28. Sciambi, A. & Abate, A. R. Generating electric fields in PDMS microfluidic devices with salt water electrodes. *Lab Chip* **14**, 2605-2609 (2014).
29. Anna, S. L., Bontoux, N. & Stone, H. A. Formation of dispersions using "flow focusing" in microchannels. *Appl. Phys. Lett.* **82**, 364-366 (2003).
30. Link, D. R., Anna, S. L., Weitz, D. A. & Stone, H. A. Geometrically mediated breakup of drops in microfluidic devices. *Phys. Rev. Lett.* **92**, 054503 (2004).
31. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
32. Anna, S. L., Bontoux, N. & Stone, H. A. Formation of dispersions using "flow focusing" in microchannels. *Appl Phys Lett* **82**, 364-366 (2003).
33. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060-1066 (2019).
34. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72-74 (2011).
35. von Kiedrowski, G. A Self-Replicating Hexadeoxynucleotide. *Angew. Chem. Int. Ed. Engl.* **25**, 932 (1986).

Figures

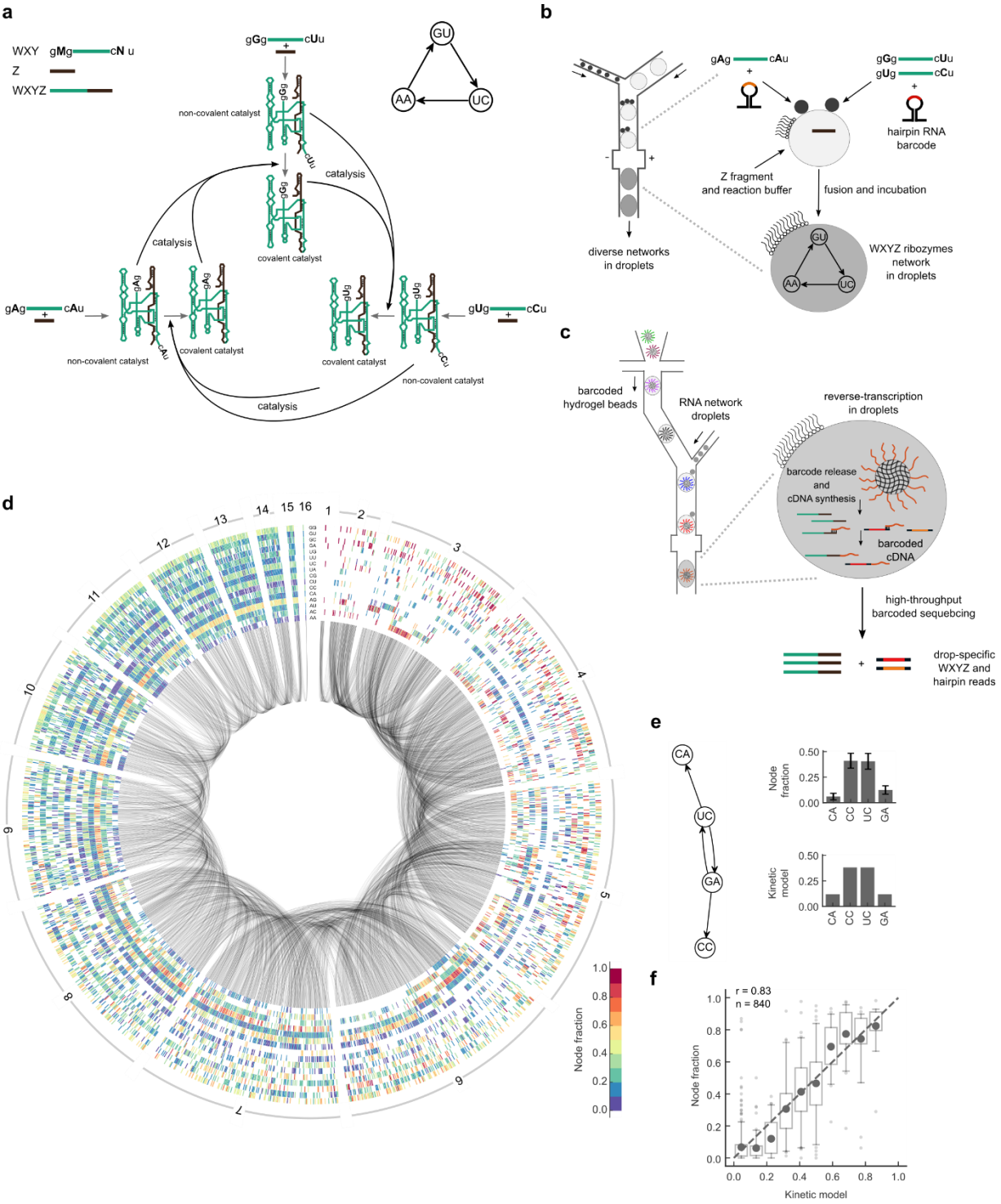


Fig. 1. Experimental set-up and RNA network compositional landscape. a, Top left: RNA fragments, named WYX and Z, can be assembled into full length ribozymes WXYZ (both non-covalent and covalent catalysts). WYX fragments each comprise 3 nucleotide long IGS and tag sequences, respectively denoted g**M**g and c**N**u, where **M**, **N** can be varied, resulting in 16 different combinations[18]. Bottom center panel: The base-pairing interactions between **M** and **N** determine the specificity of full length catalysts for their substrate fragments. In this example, each of the 3 different MN ribozyme species catalyzes the formation of another species from fragments provided as substrates, forming an autocatalytic cycle.

Note that for clarity, only Watson-Crick IGS-tag interactions are shown here. Any other non-Watson pair will have a weaker but non-zero activity (Extended Data Fig. 6c), therefore all networks possess some degree of autocatalysis. Top right panel: a simplified network representation shows catalytic species as nodes with their MN identity, and catalytic interactions as directed edges (corresponding to the black arrows of the bottom center panel). **b**, Random sets of ~3 droplets (small black disks), each containing different mixtures of WXY fragments, are fused by electrocoalescence in a microfluidic device with a droplet containing Z fragments and the reaction buffer (larger gray disks). This combinatorial process leads to a large diversity of catalytic networks, each contained in a different droplet (Methods). Substrate fragments are initially co-diluted with non-reactive hairpin RNA reporters that later allow the initial composition to be retrieved by sequencing. **c**, Droplet-level barcoded sequencing: each RNA-containing droplet is analyzed after incubation by fusing a sample of it to another droplet containing a hydrogel bead carrying DNA barcodes. These barcodes, specific to each bead, serve as primers for the reverse transcription of the RNAs, including the ribozymes formed during incubation, and the hairpin RNA reporters to identify initial condition. **d**, Compositional landscape of the 1837 unique measured networks[24]: Each ray is made of 16 boxes corresponding to the 16 possible MN species. A blank box means that the MN ribozyme is absent in this network, as determined from hairpin sequencing. Otherwise the box color indicates the measured fraction of the MN species. Grey arcs connect networks that differ from each other by a single node (addition or removal). **e**, Left: schematic of an *Azoarcus* network. Right: top, experimentally determined node fractions for the same network (error bars are ±1 s.d. over n=425 droplets); bottom, theoretical results with the kinetic model. **f**, Fraction obtained with the kinetic model versus measured fraction for every node that is a members of a 4 node-network. Bins with less than 10 points are discarded. Dark grey dots are means, quartile boxplots have 5^{th} - 95^{th} percentiles whiskers with flier points. The dotted grey line is the identity line. P-value < 0.001.

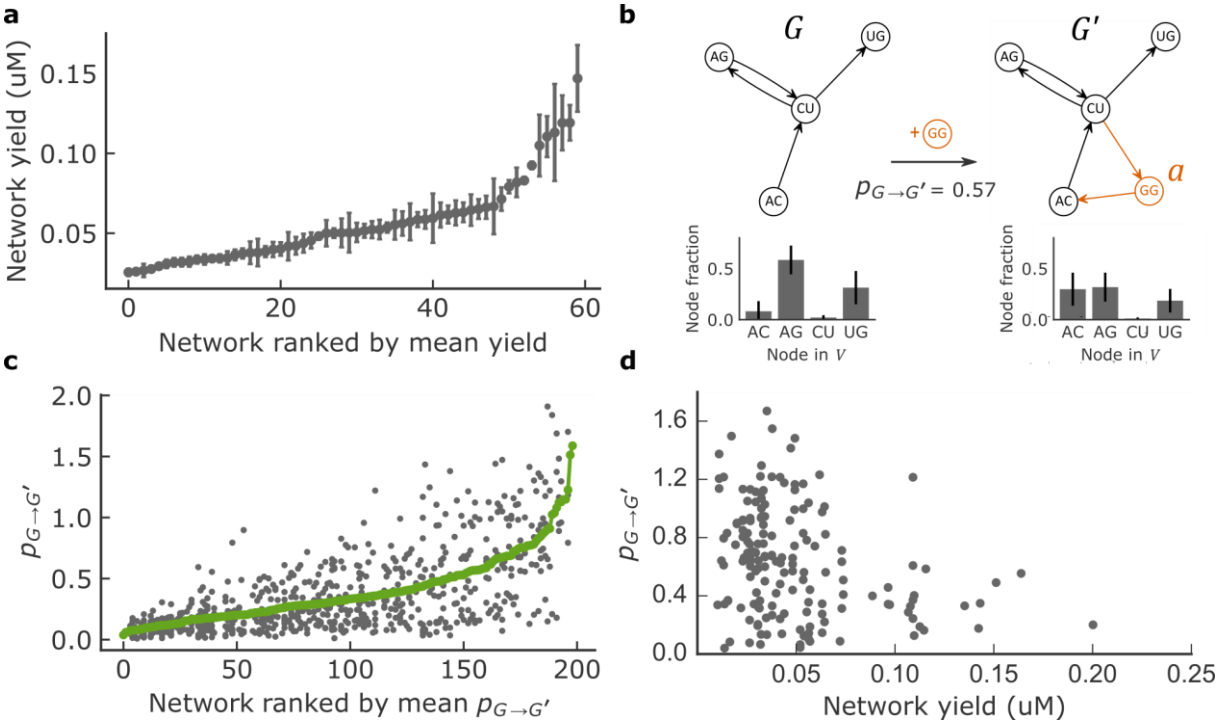


Fig. 2. Network growth and perturbation. a, Network total yield (concentration of WXYZ catalysts in μM) distribution of networks with four nodes and at least 10 replicates (for data on all sizes see Extended Data Table 2). Networks are ranked (x-axis) based on the mean across network replicates of the total WXYZ concentration (y-axis). The error bars show the Standard Error of the Mean. **b**, Example showing the addition of node a (GG, orange) to network G resulting in network G'. Top: schematic. Bottom: experimentally determined node fractions for the same network (error bars are ±1 s.d. over n=??? droplets (G) and n=??? droplets (G') **c**, Perturbation distribution of networks with four nodes (for all sizes aggregated, see Extended Data 6b). Networks are ranked (x-axis) based on the mean perturbation (y-axis). Each position on the x-axis corresponds to a certain network G, for which the perturbation has been computed for all possible single node additions present in the dataset, leading to several y-axis values corresponding to different G' networks. The green line is the average perturbation taken over all G' networks for each network G. **d**, Perturbation $p_{G\to G'}$ for networks with 3, 4 and 5 nodes plotted against network yield in μM for perturbations involving the addition of a novel catalysts (with G/C as IGS) with at least one target. There is a trade-off between perturbation and yield: odds ratio of 11 that $p_{G\to G'}$>8 for yields>0.08 (N=162, p<2.10^{-3}, right-sided Fisher exact test).

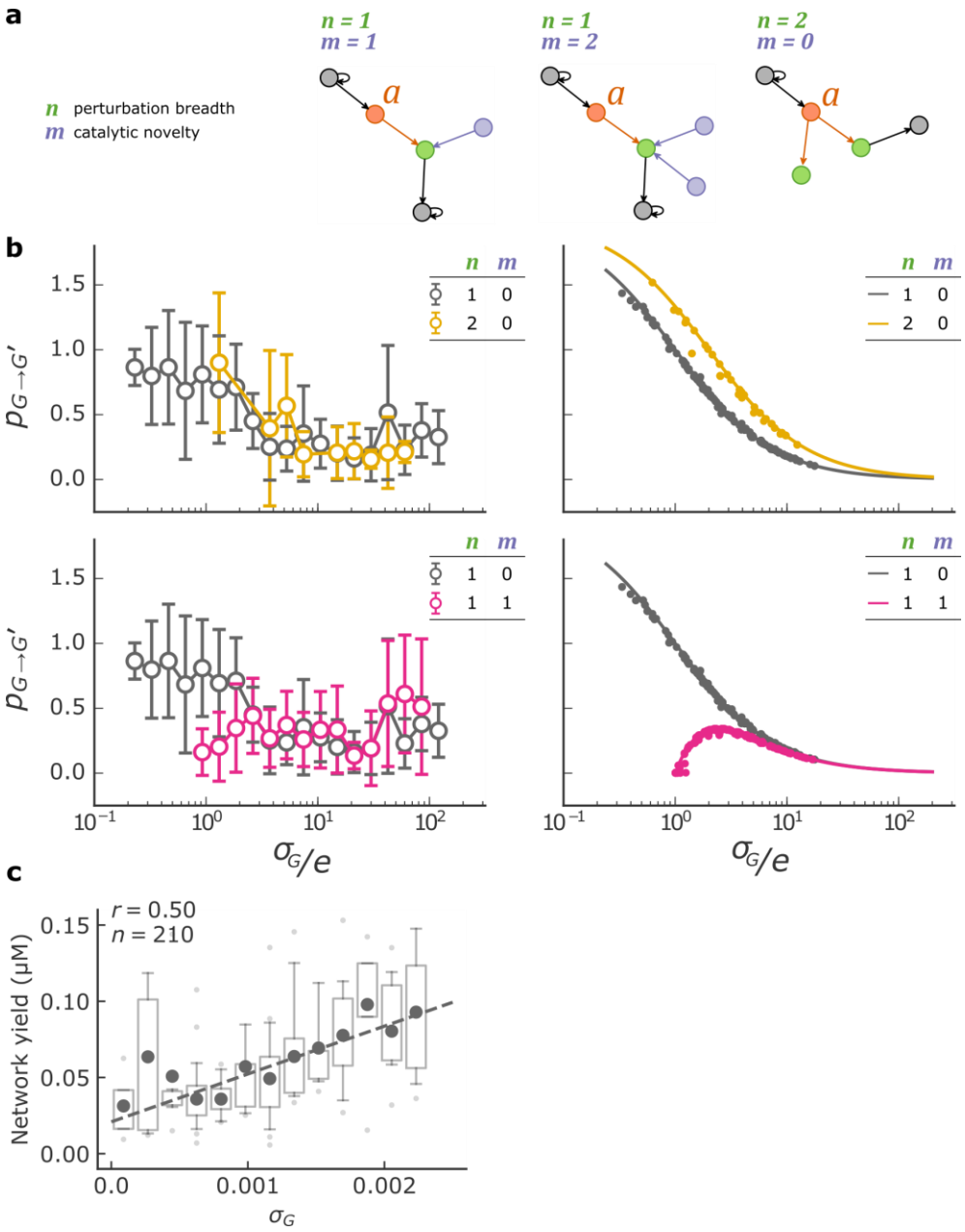


Fig. 3. Network perturbation analysis. a, Schematic of three situations where the parameters n, the number of targets of added node a, and m, the number of nodes sharing the same IGS as a, have different values. **b**, The perturbation $p_{G\to G'}$ is plotted against σ_G/e for different values of the parameters m and n for both the experimental data (left, mean and standard deviation per bin, bins with less than 3 points are discarded) and results predicted by in-degree centrality using the analytical expression (right, solid line) and with fractions predicted by eigenvector centrality (right, solid circles) for the networks with a number of nodes between 3 and 6 before addition. Results for other parameter values are shown in Extended Data Fig. 8. **c**, Network yield (μM) is plotted against σ_G for network with four nodes (see Methods for further details on how network yield is determined).

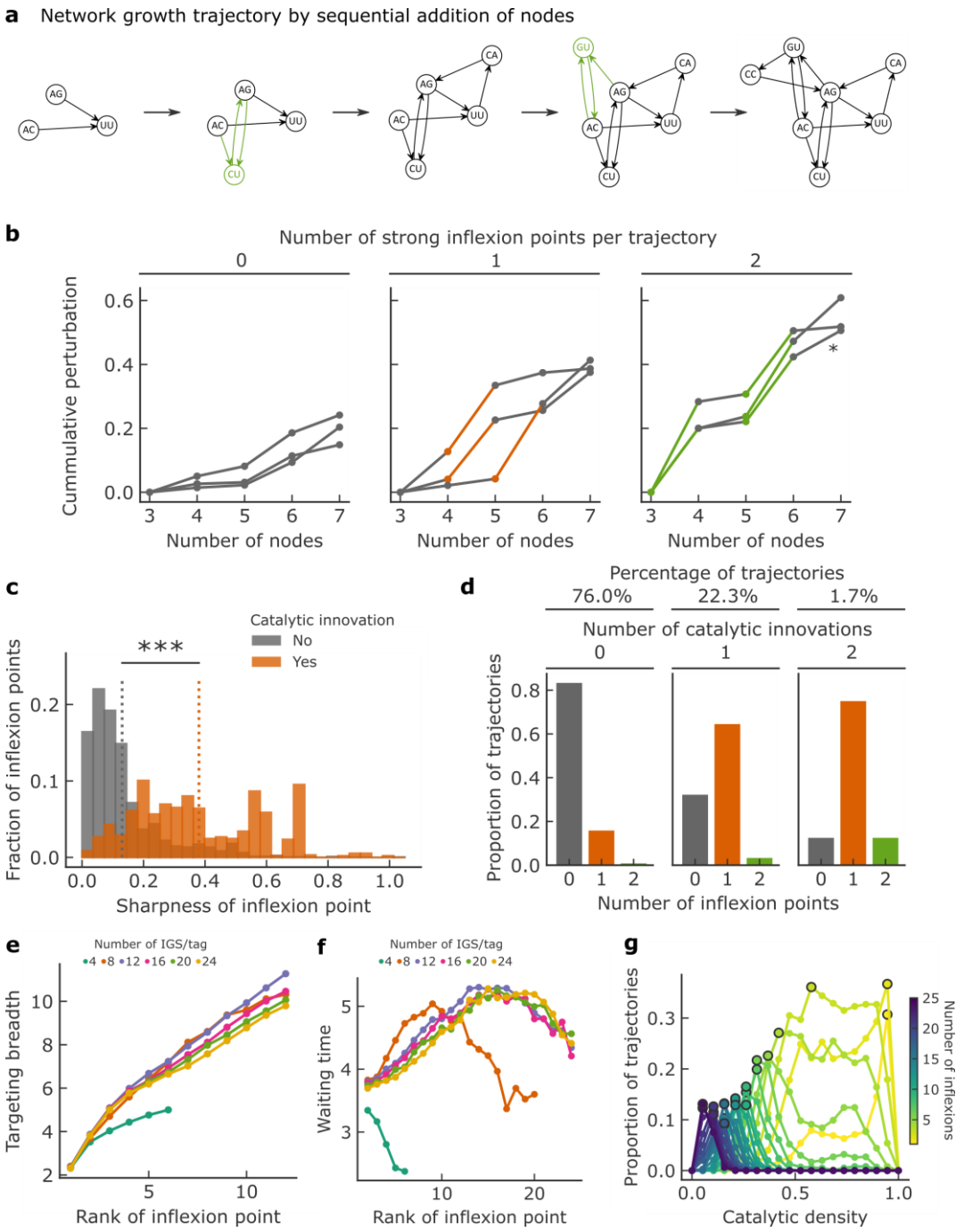


Fig. 4. Perturbation dynamics across network growth trajectories. a, Example of a network growth trajectory where at each step, a new node is added. Nodes resulting in strong perturbations (see panel b) are in green, and correspond to the IGS/tag pairs C-G and G-C. **b**, Examples of measured cumulative perturbation over trajectories, plotted against the number of node additions, by number of strong inflexions points (colored). The latter are determined as the top 25% in sharpness (absolute value of the third derivative, Extended Data Fig. 10a) measured over all trajectories. The asterisk (*) denotes the perturbation trajectory of the example shown in panel a. **c**, Distribution of sharpness for inflexion points associated with a catalytic innovation (orange, n=4,462) or not (grey, n=934). Catalytic innovations are defined as the introduction of strong IGS/tag interactions (CG or GC, Extended Data Fig. 6 b, c). that were not present until node addition. The dotted line is the mean of each distribution and the significance of the difference between the two distributions is reported (Mann–Whitney U test, $p < 0.001$). **d**, Distribution of the number of inflexion points within the 75th percentile sharpness per trajectory,

depending on the number catalytic innovations per trajectory. **e-g**, Computational study of network growth trajectories for chemistries with varying number of IGS/tag pairs and with varying degrees of catalytic density points are determined, as before, based on their sharpness (within the 75th percentile) here computed along a random sample of 1000 trajectories growing from 2 to 100 nodes. Catalytic densities are varied by random removal among the pool of all possible specific IGS/tag interactions. **e**,Targeting breath (number of targets) of catalysts causing strong perturbations, as the function of the inflexion rank. . **f**, Waiting time (number of species additions) between two strong inflexions, as a function of the inflexion rank (e.g.: rank 5 is the fifth inflexion observed along a growth trajectory) **g**, Proportions of trajectories with a given number of strong inflexion points plotted against catalytic density for a chemistry comprising up to 24 different IGS/tag pairs.

Methods

General material & methods. All experiments used DNAse/RNase free water (Thermo Fisher Scientific Product No.: 10977035). Chemicals were purchased from Sigma-Aldrich unless specified otherwise. 4-(2-Hydroxyethyl)-1-piperazinepropanesulfonic acid (EPPS) was purchased from Alfa Aesar (Product no.: J60511, CAS no.: 16052-06-5). Denaturing polyacrylamide gels (12%) were prepared using gel stock solution (Roth) and run in 1X Tris-Borate EDTA buffer. DNA oligonucleotides (Supplementary File 2) were obtained from IDT DNA technologies.

Transcription of RNAs. RNAs were *in vitro* transcribed as described in[25], using DNA templates produced as described in the Supplementary File 1. In the template for the WXY fragments the IGS is duplicated at position 25 to allow identification by sequencing. The RNAs were extracted with Phenol-Chloroform-Isoamyl alcohol, treated with DNase I (Thermo Fisher Scientific, Product No.: EN0521) and purified on polyacrylamide gels. RNA concentrations were measured using Qubit® RNA HS Assay Kit (Thermo Scientific, Product No.: Q32852).

Microfluidic devices and set-up. Droplet microfluidic experiments used HFE 7500 fluorinated oil (3M™ Novec™, Product No.:98-0212-2928-5) and a fluorosurfactant (RAN Biotechnologies, Product No.: 008-FluoroSurfactant). Devices were designed with AutoCAD software (Supplementary File 4), printed at Selba SA. PDMS (polydimethylsiloxane) devices were fabricated by soft-lithography as described in[26,27]. For electrocoalescence, indium tin oxide coated glass slides were used (Delta Technologies, Product No.: CG-90IN-S215), with channels containing 3 M NaCl liquid electrodes[28]. After plasma bonding, channels were made hydrophobic with 2% 1H,1H,2H,2H-perfluorodecyltrichlorosilane (ABCR, Product No.: AB111155) in HFE 7500. Droplet-based microfluidics experiments were performed using an inverted microscope (Nikon Eclipse *Ti*) as described earlier[27]. Flows were controlled by either syringe pumps (Harvard Apparatus Inc.) or by air-pressure control pumps (MFCS™-EZ, Fluigent SA).

Droplet microfluidic experiments. The complete experimental system consists of the steps **A** to **E** (Extended Data Fig. 1): (**A**) *Initial emulsions production:* Using all possible 16 WXY RNA fragments ($_{gMg}WXY_{cNu}$)[18], 24 unique different WXY fragments combinations were prepared in separate tubes (Extended Data Table 1). To determine

those combinations, tube contents were computationally drawn randomly among the 16 species. The total number of fragment types to distribute over all tubes (with possible redundancies) was set to 50, which was computationally found to maximize the number of distinct networks formed after random droplet fusions (10,000 iterations). Once this was fixed, network diversity was further maximized based on spectral graph criteria, by selecting the fragment assignment (among 100,000 random realizations) maximizing the Shannon entropy of the first eigenvalue and spectral gap distributions, across networks resulting again from the simulation of 10,000 random droplet fusions. Experimentally, for each combination, 1 µM of respective WXY fragment(s) was heated at 48°C and cooled down to room temperature over 10 min (0.1°C/sec) and mixed with hairpin RNA reporter at a final concentration of 0.03 µM. All 24 mixes were encapsulated individually in 5 pL droplets using flow focusing[29] on a dropmaker device (Supplementary File 4) using flow rates of 100-150 µL/h for the aqueous (RNA) and 150-170 µL/h for the oil (HFE 7500 with 2% surfactant) phases, each emulsion being collected for 5 min. All 24 emulsions were collected together in a 1.5 mL collection tube (with a PDMS cap) containing oil with 2% surfactant and mixed thoroughly. A 1.6 µM solution of Z RNA fragment in 1X *Azoarcus* reaction buffer (30 mMEPPS) pH 7.4, 20 mM $MgCl_2$), was folded in the same manner as WXY RNA, and encapsulated in 50 pL droplets using a 50 pL drop-making device (Supplementary File 4), at flow rates of 350 µL/h for the aqueous (RNA) and 250 µL/h for the oil (HFE 7500 with 2% surfactant) phases, droplets being collected in a 0.5 mL collection tube (with a PDMS cap). All flows were driven by syringe pumps. (**B**) *Electrocoalescence of initial emulsions*: The 5 pL droplets (containing WXY fragments) and 50 pL droplets (containing Z fragment and reaction buffer) were re-injected into separate channels of a droplet fusion device (Supplementary File 4). Both 5 pL and 50 pL droplets were spaced using HFE 7500 oil containing 2% surfactant and brought together into the same channel, where they were electrocoalesced[19,28] using an AC electrical field (Agilent 33522A waveform generator, sine function, 4 kHz, 400 mV, 50 Ω) amplified 10^3 times (TREK high-voltage amplifier). Flows were regulated such that, on average, between one and five 5 pL droplets coalesce with one 50 pL droplet (Supplementary File 5). This resulted in ~$1/10^{th}$ dilution of WXY RNA concentration in the fused droplets (the Z fragment is in stoichiometric excess). All flows were driven by a pressure control system (MFCS™-EZ, Fluigent SA), and the desired fusion frequency achieved by carefully adjusting and coupling the flows with pressure values of 380-400, 350-380,

470-520 mbar for 50 pL droplet emulsion, 5 pL droplet emulsion and spacer oil channel, respectively. Electrocoalescence and collection (in a 0.2 mL tube with a PDMS cap) was performed for ~3.5 h and the emulsions were stored on ice during the entire process. To monitor the fusion frequency during the collection, fusion events were counted from videos recorded every 20 min, showing very stable fusion frequency, closely comparable to what was obtained by sequencing hairpin RNA reporters (Extended Data Fig. 4). (**C**) *Incubation and splitting of droplets:* After collection, fused droplets were incubated at 48°C in a thermo-block for 1 h to allow the accumulation of ribozymes. In order to dilute the $MgCl_2$ during the reverse transcription (RT) step, these droplets (~60-65 pL) were re-injected (spaced with HFE 7500 oil containing 2% surfactant) and split by a T-junction device[30] (Supplementary File 4) into smaller droplets (~5 pL) prior to the RT step. Flow rates were maintained using pressures of 800-1000 mbar and 150-250 mbar, respectively for the oil separator channel and droplet inlet. (**D**) *Droplet barcoding and RT in droplet*: To sequence the RNA in each split droplet we developed a strategy combining droplet barcoding and RT in droplets, inspired by single-cell transcriptomics methods described earlier[20,31]. Here barcoded hydrogel beads are individually encapsulated in droplets with all the necessary reagents for RT and fused with RNA network containing droplets. The details of each step are as follows: (*a) Barcoded hydrogel bead synthesis*: Hydrogel beads were produced by co-encapsulating 10% (w/w) polyethylene glycol diacrylate (PEG-DA) 6000 (Sigma, Product No.: 701963), 1% (w/w) PEG-DA-700 (Sigma, Product No.: 455008), 400 µM of acrydite-modified dsDNA with a 4 base sticky-end (Oligo 10, top, carrying a 5'-acrydite modification, and Oligo 11, bottom, Supplementary File 2), 10 µM of FITC-Na (Fluorescein isothiocynate), 1% (v/v) photo-initiator (2-hydroxy-2-methylpropiophenone, Sigma, Product No.: 405655) in buffer (75 mM Trizma HCl pH 7.4, 50 mM NaCl) using flow focusing[32] on a dropmaker device (Supplementary File 4). 9 pL droplets were produced using flows of 150 µL/h for aqueous solution and 500 µL/h for oil (HFE 7500 with 2% surfactant). Droplets were polymerized by UV irradiation (UV omnicure AC475-365, Lumen Dynamics, ~360 mW exposure, λ_{nm} = 365 nm) through a PTFE tubing and collected in 5 mL tube (tubing internal diameter = 0.3 mm, distance from lamp 3 cm, tubing length = 12 cm; approximate exposure time = 47 s). Prior to UV irradiation, the complete experiment set-up was maintained in the dark to avoid any spontaneous polymerization. The collected droplets were washed once with 4 mL hexane to break the emulsion and then the beads washed 3 times with 4 mL binding-wash buffer (20 mM Trizma-

HCl pH 7.4, 50 mM NaCl, 0.1% of Tween 20) by centrifuging at 3000 g for 2 min. Beads were then filtered using Steriflip 20 µm Nylon filters (Millipore, Product No.: SCNY00020) and stored in binding-wash buffer supplemented with 1 mM EDTA. In contrast to earlier methods[20,31], barcodes were built on the beads using a split-and-pool ligation strategy[33] (Extended Data Fig. 2). For this, 250 µL (~10 million) of pelleted beads were washed 3 times with 4 mL of binding-wash buffer by centrifuging at 3000g for 2 min, and beads were subjected to a first ligation step after washing. Ligation was performed on a 2 mL reaction scale with 1X T7 DNA Ligase buffer (New England Biolabs), 4 µM dsDNA adaptor with two different 4 base sticky-ends, the first being complementary to the 4 base sticky-end on the beads, and containing a BclI restriction site and the *Read* 2 illumina adaptor sequence (Oligo 12, top and Oligo 13, bottom, Supplementary File 2), 30 U/µL of T7 DNA Ligase (New England Biolabs, Product No.: M0318L), at room temperature for 25 min with agitation at 600 rpm. Beads were then washed 5 times with 4 mL of binding-wash buffer and pelleted after the last wash. For the next ligation step, a master-mix (1.6 mL) was prepared to containing 1X T7 DNA Ligase buffer, 30 U/µL of T7 DNA Ligase and divided into the wells of a 96-well deep well plate (16 µL per well), pre-filled with 4 µL of 20 µM 1st barcode index (Index A) (Supplementary File 2), with a different index in each well. These indexes had two different 4 base sticky-ends, the first being complementary to the free sticky-end of oligos ligated in previous step (dsDNA, Oligo 12 and 13). The plate was sealed and incubated at 25°C with 600 rpm for 25 min. Then the ligase was heat inactivated at 65°C for 10 min and the content of all the wells was pooled after adding 200 µL of binding-wash buffer containing 1 mM EDTA. After pooling, the beads were washed 7 times with 4 mL of binding-wash buffer. The complete process of split-and-pool was repeated to ligate the 2nd (Index B) and 3rd (Index C) barcode indexes (Supplementary File 2) generating a total diversity of 8.8 x 10^5 (96^3) barcodes. After the final pooling step, a partially double-stranded sequence was ligated to all the beads using the same ligation protocol (Oligo 14, top and Oligo 15, bottom, Supplementary File 2). This sequence comprises a sticky-end complementary to the free sticky-end on the 3rd index, and a single-stranded 33 base long 3'-overhang at the other end (Extended Data Fig. 2b). Out of these 33 nucleotides, the first 8 are random nucleotides used as unique molecular identifiers (UMIs)[34] and remaining 25 nucleotides, which function as a primer for RT, are complementary to the 3' end of the WXYZ ribozyme and hairpin RNA reporter (which both contain the same primer binding region). The final barcoded hydrogel bead library

was washed 5 times with 4 mL of binding-wash buffer supplemented with 1 mM EDTA and re-suspended in the same buffer. (b) *Encapsulation and fusion*: The barcoded beads (50 µL) were washed 5 times with 500 µL of binding-wash buffer by centrifuging at 11866 g for 1 min. Washed beads were then mixed with 2.5 µL of 10 mM dNTPs, 10 µL of 5X SuperScript III reaction buffer (Invitrogen), 2.5 µL of 100 mM dithiothreitol (DTT), 0.4% of Tween20 (10%), 100 units of SUPERase·In^{TM} inhibitor (Thermo Fisher Scientific, Product No.: AM2694) and incubated at 37°C for 30 min to ensure diffusion of all reactants within the beads. After this, the beads were centrifuged as above and excess of liquid over the beads was removed. Then, at 4°C, 500 units of reverse transcriptase (SuperScript III, Thermo Fisher Scientific, Product No.: 18080044) and 25 units of BclI restriction enzyme (New England Biolabs, Product No.: R0160L) were added to the beads and mixed thoroughly. These hydrogel beads together with the other reagents were encapsulated individually in ~50 pL droplets (Supplementary File 6), fused with RNA droplets on the same microfluidic device (Supplementary File 4). The injection of close-packed deformable beads resulting in >99% of droplets containing a single barcoded hydrogel bead. The 5 pL RNA containing droplets and the 50 pL bead containing droplets were fused at a 1:10 ratio to ensure that, in the majority of cases, no more than one of the former is fused with one of the latter. For this, flows were pressure controlled using, respectively, 550, 225, 650, 450 mbar for hydrogel beads, 5 pL droplets, oil to control droplet spacing (HFE 7500 with 2% surfactant) and oil for beads encapsulation (HFE 7500 with 2% surfactant). Electrocoalescence was performed as described above. Fused droplets were collected in 0.2 mL collector tube (with PDMS cap) containing oil with 2% surfactant. (c) *cDNA synthesis and extraction*: The fused droplets were then incubated at 60°C in a thermo-block for 1 h to release the DNA barcodes using the BclI restriction enzyme and perform cDNA synthesis in droplets. The emulsion was broken post incubation by adding 2 volumes of 1*H*, 1*H*, 2*H*, 2*H*-Perfluoro-1-octanol and 100 µL water. The pooled barcoded cDNAs are then isopropanol precipitated and re-suspended in 40 µL of water. **(E)** <u>*Sequencing sample preparation*</u>: In order to block extension of the non-elongated primers, 20 µL of cDNA was mixed with 1X TdT reaction buffer (New England Biolabs), 0.25 mM $CoCl_2$, 0.4 mM ddCTP (Roche CustomBiotech, Product No.: 12158183103) and 0.4 U/µL of Terminal Transferase (TdT, New England Biolabs, Product No.: M0315L). The reaction was carried out in 50 µL volume and incubated for 30 min at 37°C before heat inactivation at 70°C for 10 min. The TdT treated cDNAs

were purified using magnetic beads (AMPure XP, Beckman Coulter, Product No.: A63881) following the manufacturer's protocol. Sequencing adaptors were then appended using two sequential PCR steps (Extended Data Fig. 3a). The first PCR was performed in multiples of 20 µL volume where 2 µL of TdT treated reverse transcription reaction was mixed with 0.5 µM of Oligo 16 (also containing a 6 nucleotide stretch used as sample barcode to multiplex different samples for the sequencing run) and 0.5 µM of Oligo 17 (Supplementary File 2) as forward and reverse primers in 1x PCR buffer (Thermo Scientific), 0.2 mM dNTPs, 0.01 U/µL of polymerase (Thermo Scientific Phusion Hot Start II, Product No.: F459) using the following protocol: initial denaturation 98°C/30sec, then 18 cycles of denaturing 98°C/10 s, annealing and extension 72°C/1 min, and a final extension of 72°C/5 min. PCR products were purified using AMPure XP magnetic beads and eluted in 34 µL of water. The second PCR was performed as above, but in 40 µL volume with 4 µL as PCR I purified product as template, using Oligo 18 and Oligo 19 (Supplementary File 2) as forward and reverse primers, and using the following protocol: initial denaturation 98°C/30sec, then 10 cycles of denaturing 98°C/10 s, annealing 56°C/30 s, extension 72°C/30 s, and a final extension of 72°C/5 min. The final sample was purified using AMPure XP magnetic beads, resuspended in 34 µL of water, analysed on TapeStation (Agilent 2200 TapeStation, using high sensitivity D1000 ScreenTape®, Product No.: 5067-5584) and quantified by Qubit® dsDNA HS Assay Kit (Thermo Scientific, Product No.: Q32854). Libraries were sequenced using the Illumina NextSeq 550 system in 2*150 High Output mode at the Genotyping and Sequencing Core Facility, ICM Paris (iGenSeq, Institut du cerveau et de la moelle épinière).

Sequencing data processing. Custom software was written to process sequencing data and further data analysis steps. The libraries were sequenced in paired-end mode with *read* 1 and *read* 2 of 180 bp and 120 bp, respectively. *Read* 1 was used to determine the sample barcode (used for multiplexing samples for sequencing) and the RNA molecule identity (either an *Azoarcus* ribozyme with specific IGS-tag pair or a hairpin RNA reporter). *Read* 2 was used to determine the UMIs[34] and the droplet barcode. The structure of *read* 1 and 2 is summarized in Extended Data Fig. 3b. *UMI normalization*: Four meta fields were associated for each pair of *reads*: sample barcode, droplet barcode, UMI and RNA molecule identity (details in Supplementary File 1). *Reads* missing any of these were discarded (~30%). The *reads* with the same

meta information were merged and the *read* count was added as an extra meta field. This allows to filter out noise due to sequencing errors, given that *read* counts resulting from such errors are significantly lower (threshold indicated on Extended Data Fig. 3e). The filtering thresholds for the minimum number of *reads* per UMI were determined separately for the *Azoarcus* ribozymes and hairpin RNA reporters based on the shape of the distribution of number of *reads* per UMI (Extended Data Fig. 3e-f). *Final data processing*: Filtered UMI-normalized *reads* were then clustered by droplet barcode to count the number of UMIs per type of RNA molecule per droplet barcode (Extended Data Fig. 3f). For each droplet barcode, only hairpin RNA reporters where UMIs made up ≥7.5% of the total of hairpin UMIs were retained. This threshold value was set to optimally match the fusion distribution measured by video (Extended Data Fig. 3g). Within each droplet, ribozyme sequences that did not corresponding to a correct hairpin reporter were discarded. Then, only the droplet barcodes with more than 10 UMIs associated with hairpin RNA reporters and 20 with ribozymes were retained and the fraction of each node in the network was computed. As before, these thresholds were chosen to closely match the observed distribution of drop fusion determined from videos acquired during the experiments (Extended Data Fig. 3h, i). Processed network compositions and mean node fractions are publicly available for all unique networks (Supplementary File 3). To determine network yield per droplet, first we computed a droplet-specific UMI to concentration conversion rate by taking into account total hairpin reporter UMIs as well as number of different hairpin reporters identified. This was then used to convert total number of ribozyme UMIs into a yield value.

Control experiment to measure cross-talk and quantification bias during droplet-level sequencing. Four 5 pL emulsions (A, B, C and D) were prepared, each with: a distinct pair of hairpin RNA reporters at a final concentration of 5 nM, a set of 4 WXYZ ribozymes with the same IGS but in different proportions (total concentrations summing to 100 nM) and 1.6 µM of Z fragment. The four emulsions were mixed and then singly fused to droplets containing hydrogel beads as described above to perform droplet level sequencing. The data was processed as above using the same thresholding strategy. We quantified the bias as the proportion of droplets containing pure (A, B, C or D) and compared to mixed populations (AB, AC, AD, BC, BD, CD, ABC, ABD, BCD and ABCD, Extended Data Fig. 5e).

Experimental measurement of catalysis by non-covalent and covalent ribozymes. Self-assembly of *Azoarcus* ribozymes with different IGS/tag pairs were re-measured in 20 mM $MgCl_2$, using the strategy developed by von Kiedrowski[35], as described[21]. For this, the initial rate of formation of WXYZ from WXY and Z fragments was measured as a function of the concentration of doped covalent WXYZ ribozymes (with same IGS and tag as the WXY fragment). The concentration of RNA fragments were identical to the droplet experiments (0.1 µM of WXY fragment and 1.6 µM of Z fragment) with additional but minor amounts (0.001 µM) of WXY fragment, radiolabeled using T4 Polynucleotide Kinase (New England Biolabs) and γ-32P ATP (Perkin-Elmer). In accordance to previous measurements[21], the initial rate of formation of WXYZ (v_0) was found to depend linearly on the concentration of the doped WXYZ. This linear relationship can be described with the following linear equation $v_0 = \alpha x + \beta$ (Extended Data Fig. 6 b,c) where, x is the concentration of the covalent ribozyme, α is the slope for an IGS/tag pair, which quantifies the WXYZ synthesis by covalent ribozymes, and β is the intercept, which quantifies the WXYZ synthesis by non-covalent ribozymes (trans-catalysis by non-covalent complex of WXY and Z fragments)[18]. Wobble pairs (GU and UG) and one of the mismatch IGS/tag pairs (AG) were also measured and found to have negligible values for α and β.

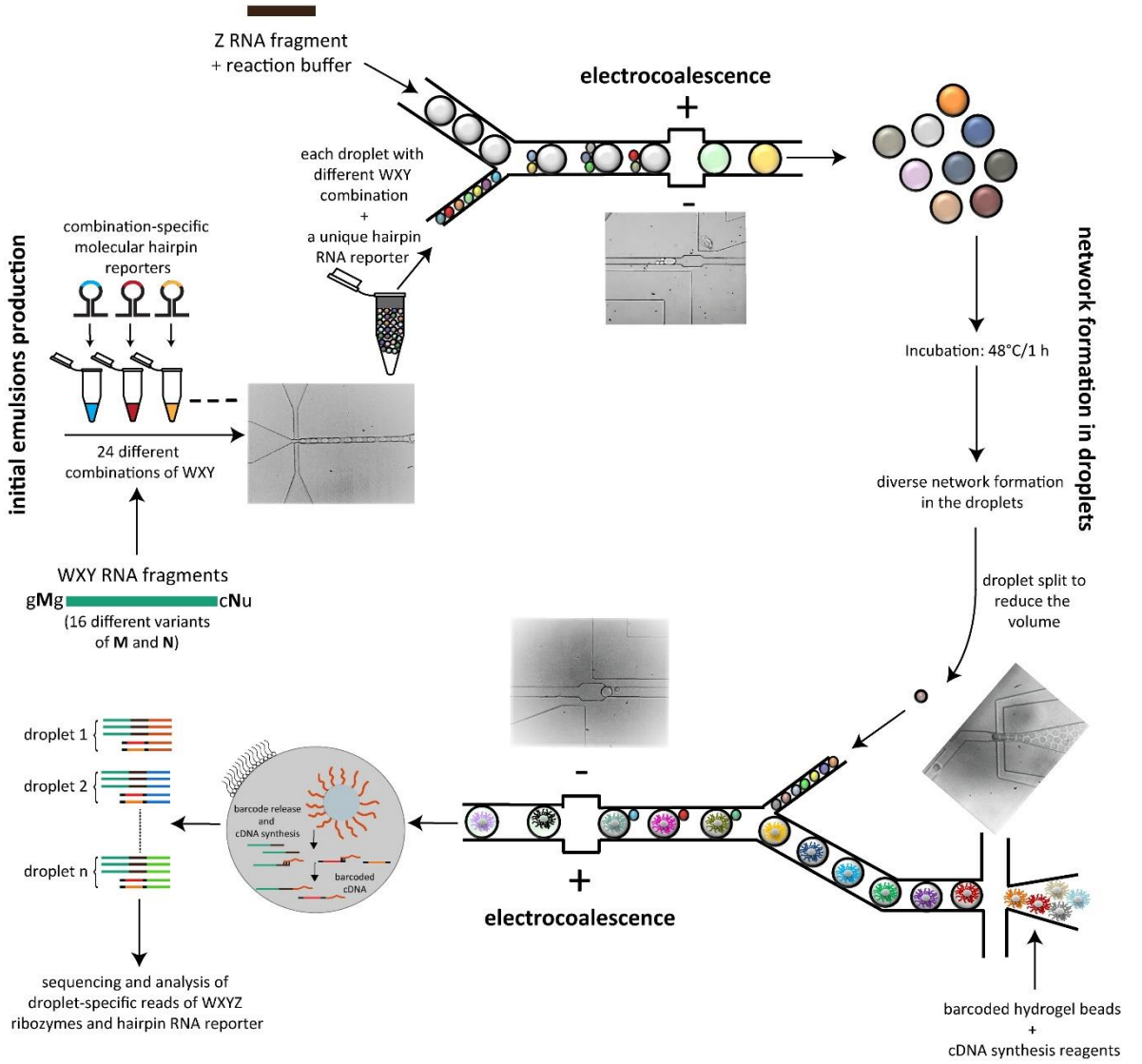
Kinetic model. In a network, covalent catalysts from species j, whose concentration is x_j, contributes to the rate of formation of covalent catalysts from species i with the term $\alpha_{ij}x_j$ while non-covalent catalysts contributes with the constant term β_{ij}. α_{ij} and β_{ij} terms were measured experimentally for single IGS-tag pairs (Extended Data Fig. 6b-c). For each species i, summing these two terms over all nodes gives the total rate of formation. This results in a linear system of ordinary differential equations (ODE) per network which was solved to obtain concentrations of all species in a network after incubation of the reaction.

Additional references

26. Duffy, D. C., McDonald, J. C., Schueller, O. J. & Whitesides, G. M. Rapid Prototyping of Microfluidic Systems in Poly(dimethylsiloxane). *Anal. Chem.* **70**, 4974-4984 (1998).
27. Mazutis, L. *et al.* Single-cell analysis and sorting using droplet-based microfluidics. *Nat. Protoc.* **8**, 870-891 (2013).

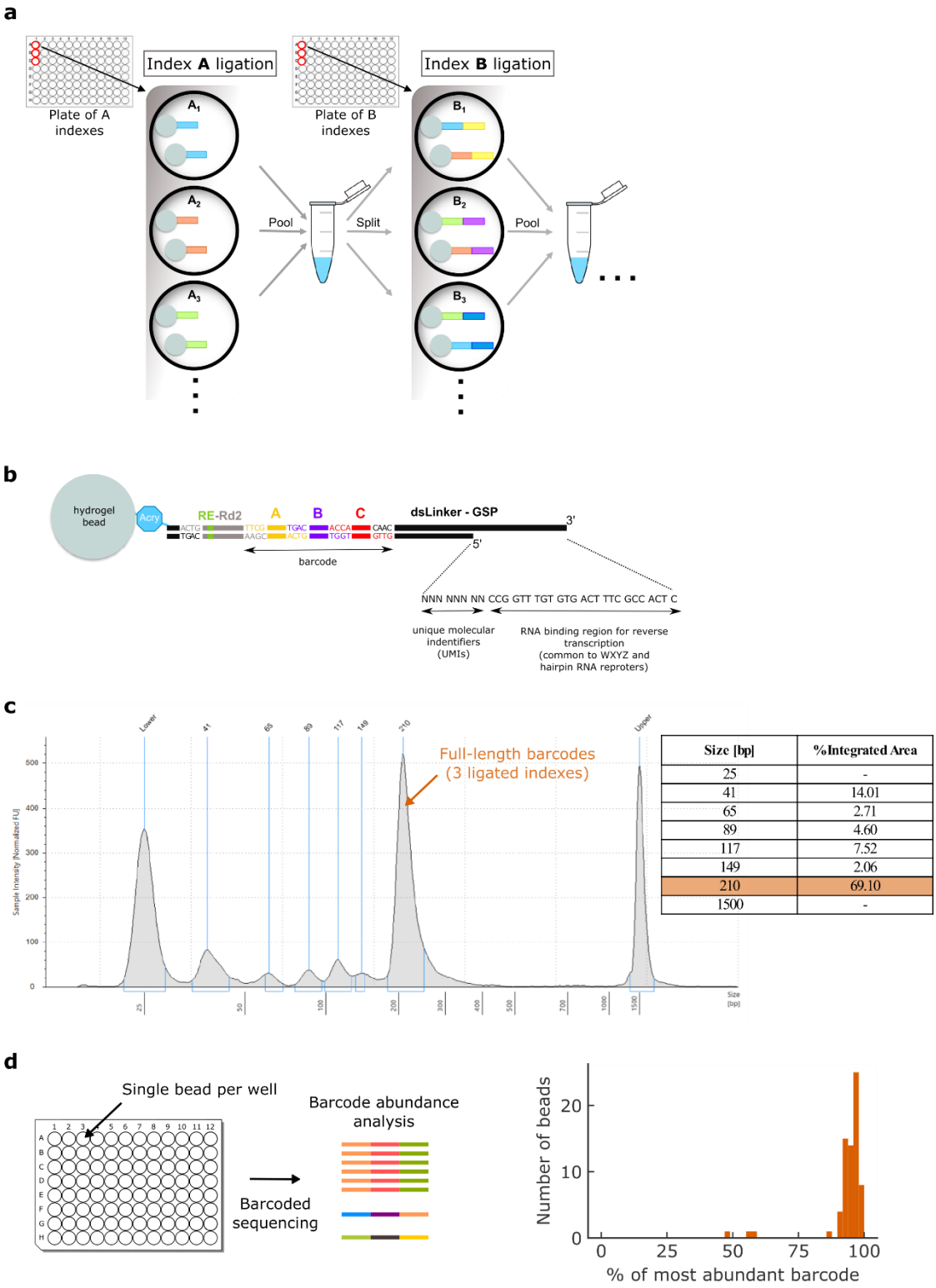
28. Sciambi, A. & Abate, A. R. Generating electric fields in PDMS microfluidic devices with salt water electrodes. *Lab Chip* **14**, 2605-2609 (2014).
29. Anna, S. L., Bontoux, N. & Stone, H. A. Formation of dispersions using "flow focusing" in microchannels. *Appl. Phys. Lett.* **82**, 364-366 (2003).
30. Link, D. R., Anna, S. L., Weitz, D. A. & Stone, H. A. Geometrically mediated breakup of drops in microfluidic devices. *Phys. Rev. Lett.* **92**, 054503 (2004).
31. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
32. Anna, S. L., Bontoux, N. & Stone, H. A. Formation of dispersions using "flow focusing" in microchannels. *Appl Phys Lett* **82**, 364-366 (2003).
33. Grosselin, K. *et al.* High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat. Genet.* **51**, 1060-1066 (2019).
34. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72-74 (2011).
35. von Kiedrowski, G. A Self-Replicating Hexadeoxynucleotide. *Angew. Chem. Int. Ed. Engl.* **25**, 932 (1986).

Extended Data



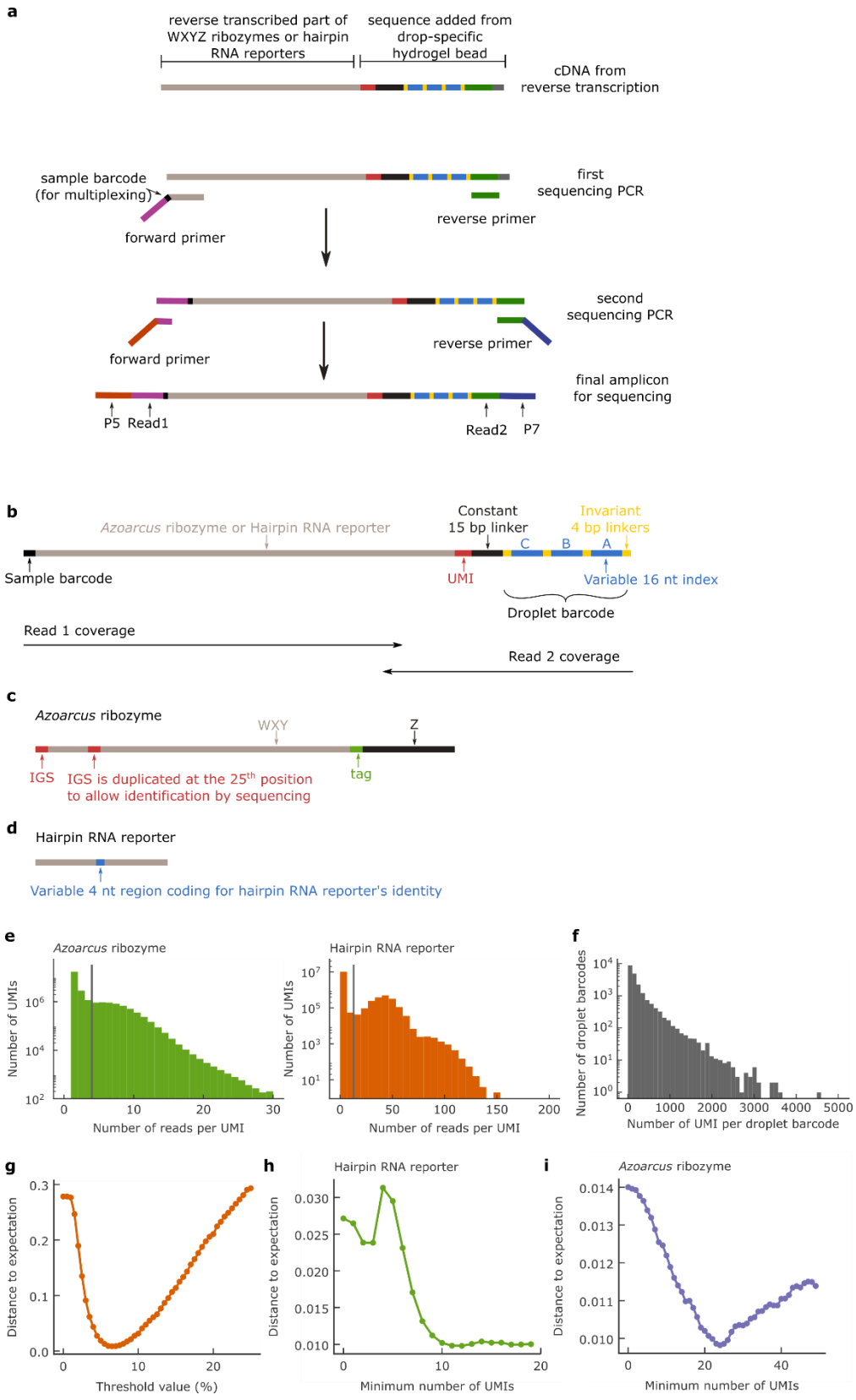
Extended Data Fig. 1. Schematic of the complete experimental set-up. The different steps involved in the experiment. See Methods for details. On a first microfluidic device, 24 different initial emulsions, comprising 5 pL droplets, were prepared, each containing a unique combination of WXY fragments (Extended Data Table 1), and a unique molecular hairpin RNA reporter. These emulsions were collected in a single tube (1.5 mL) and mixed thoroughly. These 5 pL droplets were then re-injected into a second microfluidic device, where they were fused by electrocoalescence with 50 pL droplets, produced on a separate chip, containing Z fragment and the reaction buffer. The pairing of 5 pL and 50 pL droplets was controlled such that one to five 5 pL droplets fuse with one 50 pL droplet. After fusion, the resulting emulsion, comprising 65 pL droplets, was incubated at 48°C for 1 h, and then split into 5 pL droplets on a third microfluidic device to reduce the volume. These droplets were re-injected into a fourth microfluidic device, where they were fused with 50 pL droplets, produced on the same chip, containing single hydrogel beads carrying barcoded cDNA primers, Z fragment and reverse transcription reaction reagents. The ratio of 5 pL to 50 pL droplets was kept low (1:10) such that only one 5 pL RNA droplet was fused with one 50 pL droplet. The resulting emulsion was incubated at 60°C for 1 h to allow

release of barcoded primers from the beads and cDNA synthesis. Then, the emulsion was broken, the barcoded cDNA was amplified by PCR to append sequencing adaptors, and sequenced.



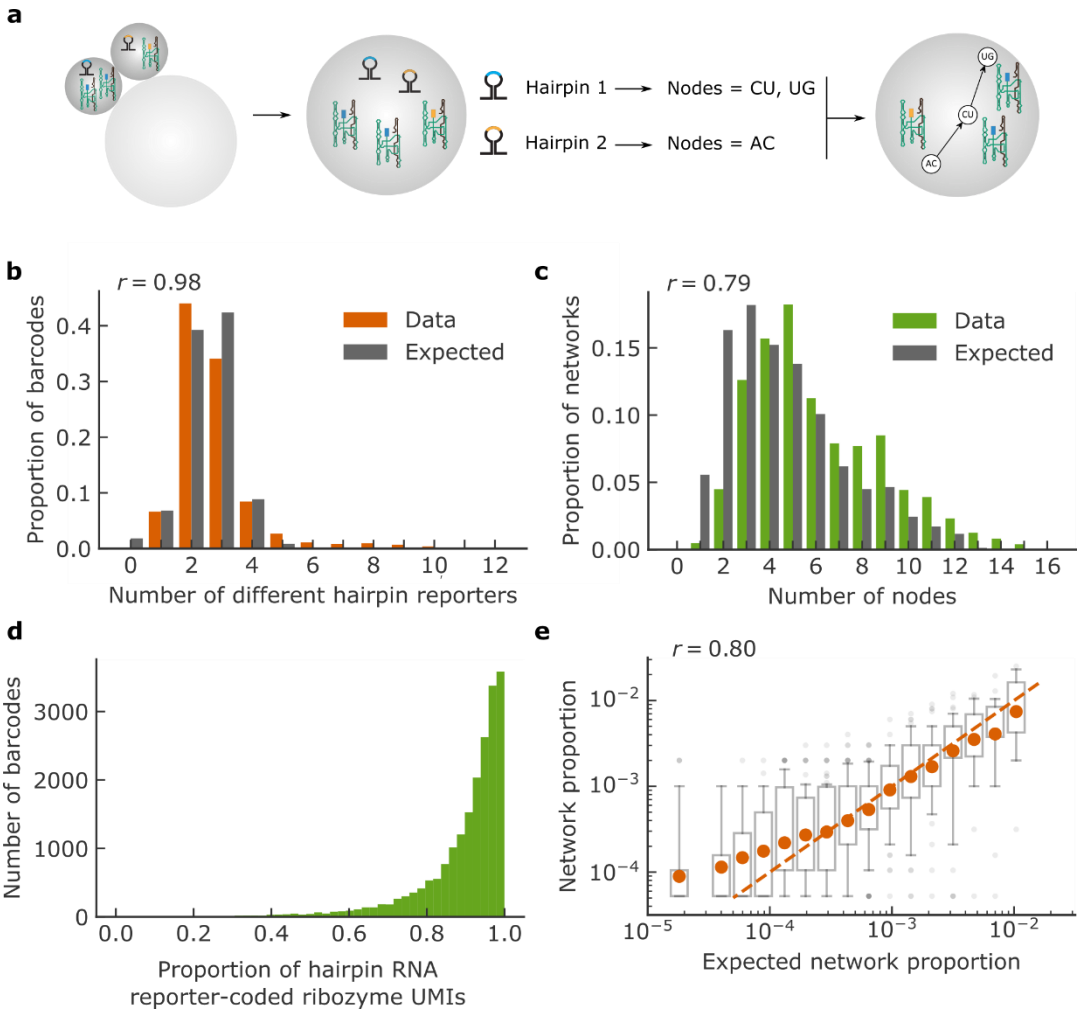
Extended Data Fig. 2. Barcoded hydrogel beads. a, Synthesis of barcoded hydrogel beads using a split and pool ligation strategy (details in Methods). On each bead, the clonal population of barcodes is built by successive ligation of indexes in one of the wells of 3 successive 96-well plates, in which each well contains a different index. Between each ligation step, beads are pooled before being distributed randomly in the 96 wells of the next plate. After the ligation of three indexes (A, B and C), a common oligo containing a linker (double stranded), UMIs (8 nt, single stranded), and the RNA binding region for

reverse transcription (25 nt, single stranded) is ligated. **b**, Final structure of a barcode: linker, restriction site (RE), *Read* 2 Illumina sequence (Rd2), 3 barcode indexes (A, B and C) separated by linker regions, UMI and cDNA primer. The sequences of the ligated 4 base sticky-ends in the linker regions are indicated. **c**, Capillary electrophoresis (TapeStation, Agilent) data of the barcoded primers released from beads by restriction digestion. The length at the main peak (~70%) corresponds to fully ligated barcodes with three indexes. Note that the actual length of full barcodes is 152 bp compared to the length indicated by TapeStation (210 bp) due to barcodes being partially single-stranded. See Supplementary File 1 for details. **d**, Distribution of percentage of the most abundant barcode sequence per bead is plotted for n=71 beads obtained by sequencing of barcodes released from single beads in microplate wells. See Supplementary File 1 for the details.



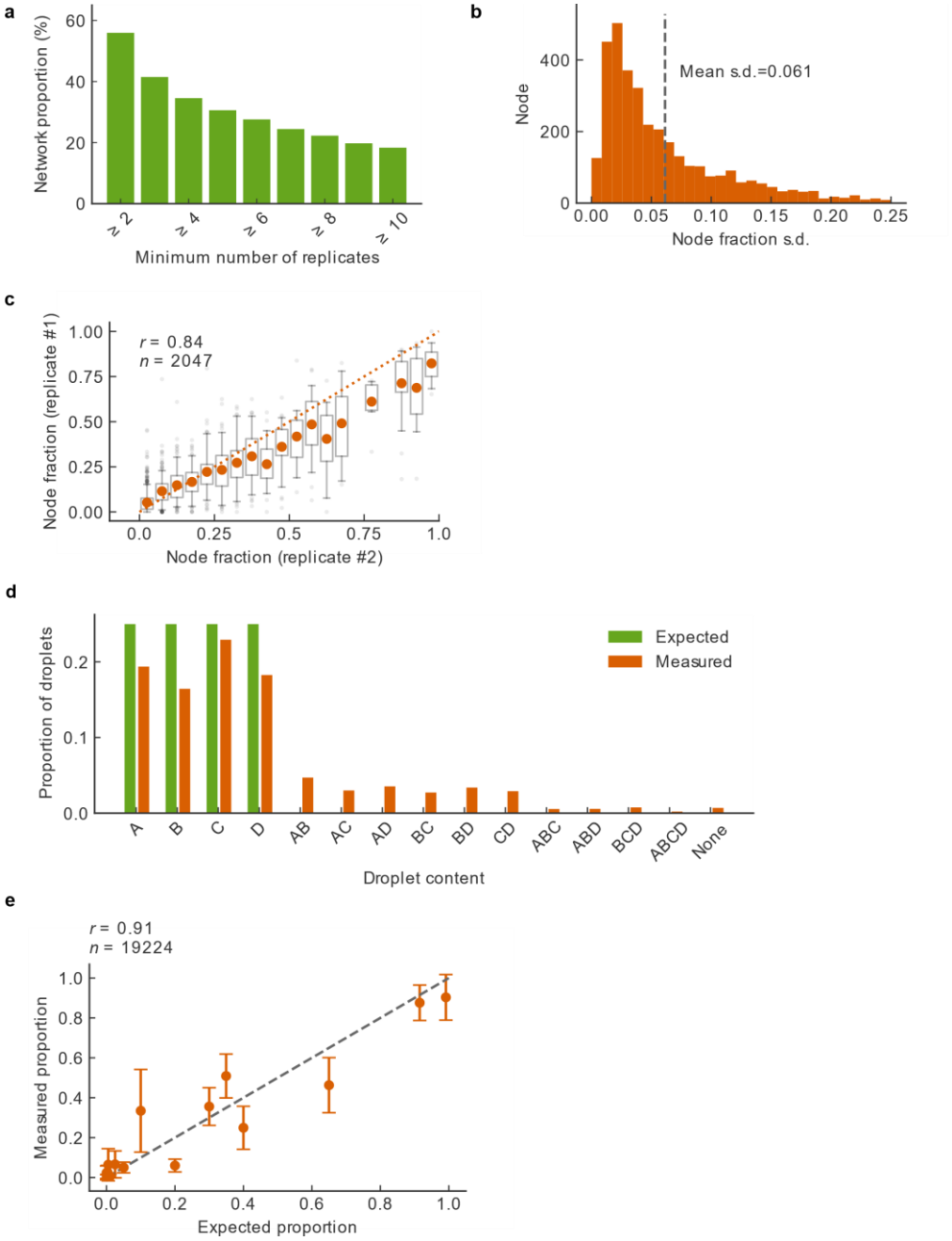
Extended Data Fig. 3. Sequencing and analysis. a, Schematic of PCR steps to append sequencing adaptors (see Methods). In the first step, barcoded cDNAs are amplified with a gene-specific forward primer, which also contains a barcode for multiplexing different samples for sequencing, and a common *read* 2 reverse primer. The PCR products from first PCR are then subjected to a second PCR to add adaptors compatible with Illumina sequencing platforms. **b-c**, Sequencing *read* structures **b**, Generic paired-end *read* structure. **c**, Detailed structure

of the *reads* for *Azoarcus* ribozymes and for hairpin RNA reporters. **d**, Structure of molecular hairpin RNA reporters, which share the same 5' and 3' regions as WXYZ to allow amplification. These hairpins have a variable loop region of 4 nucleotide serving as barcode. **e**, Distribution of number of UMIs as a function of the number of *reads* per UMI for *reads* associated with *Azoarcus* ribozymes or hairpin RNA reporters. The threshold, indicated by the gray line, is determined by the intersection of the the "noise" distribution with very low number of *reads* per UMI and the "signal" distribution with higher mean number of *reads* per UMI. Any UMIs with less *reads* than this threshold were discarded. Note that x-axis scale for the plots is different because of the more efficient amplification of hairpin reporters during PCR due to their smaller size. **f**, Distribution of number of UMIs (ribozyme and hairpin) per droplet barcode. **g-i**, Choosing the thresholds. The distance to expectation is the Euclidian distance between the distribution of the number of hairpin RNA reporters per droplet barcode once the three thresholds (see below) have been applied, and the distribution of droplets at the fusion step measured by video acquisition. The three thresholds are: **g**, the minimum percentage of total hairpin RNA reporter UMIs used to declare the corresponding hairpin reporter to be part of the coding set for a given droplet barcode (chosen value = 7.5%), **h**, the minimum number of hairpin RNA reporter UMIs for a droplet barcode to be included in the final dataset (chosen value = 10), and **i**, the minimum number *Azoarcus* ribozyme UMIs for a droplet barcode to be included in the final dataset (chosen value = 20).



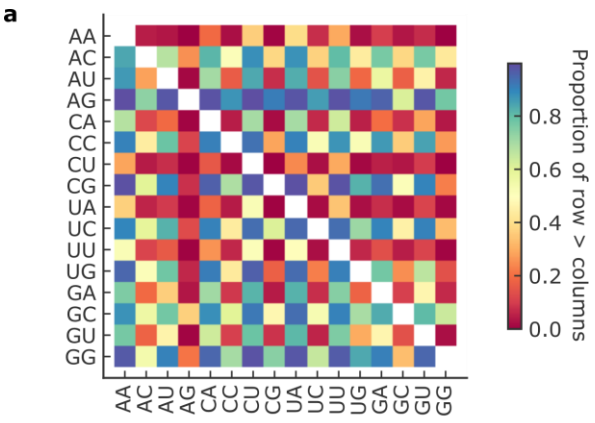
Extended Data Fig. 4. Identification of the network structure at the droplet level using hairpin RNA reporters. **a**, Schematic illustrating identification of a network structure using hairpin RNA reporters. The hairpin RNAs where the UMI counts make up ≥7.5% of the total number of UMI counts for hairpin RNAs are considered as part of the coding set. This can be used to derive network structure in addition to node identification using an internal mutation in WXYZ ribozyme (Extended Data Fig. 3c, Supplementary File 1). The number of UMIs for each node of the identified network structure gives the node fraction. **b**, In orange, measured distribution of the number of hairpin RNA reporters per droplet barcode in the final dataset. In dark grey, distribution of number of 5 pL droplets (containing initial WXY emulsions) fused with a 50 pL droplet (containing reaction buffer and Z) measured by video acquisition (Methods). Pearson's correlation coefficient is reported ($p<0.001$, $n=13$). **c**, In green, distribution of number of nodes per network per droplet barcode in the final dataset. In dark grey, expected distribution from simulating the library fusion step with the same number of droplets as in the data and with the fusion frequencies measured by video acquisition (Methods). Pearson's correlation coefficient is reported ($p<0.001$, $n=17$). **d**, Histogram of distribution of the proportion of hairpin RNA reporter-coded ribozyme UMIs, defined as the ratio between the total numbers of UMIs associated with nodes identified by the hairpin reporters and the total number of UMIs for *Azoarcus* ribozyme identified by the internal mutation (Supplementary File 1) for the same droplet barcode. A value of 1 is reached when all ribozyme UMIs predicted by the set of hairpin reporters are identified. **e**, Correlation between the measured and expected network proportions. Network proportion is calculated as the ratio between the number of replicates of a given network and the total number of replicates for all networks. Expected network proportion is obtained by simulating the library fusion step with 50,000 droplets mimicking the experimental set-up using the fusion frequencies measured by video acquisition. About 1/3 of the networks in the data do not have any replicate in the simulation and are excluded from the analysis. Bins with less than 10 points were discarded. The box extends from the lower

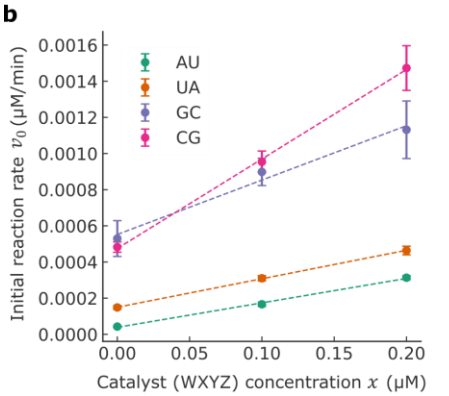
to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5^{th} percentile to the 95^{th}. Flier points are those past the end of the whiskers. The dotted orange line is the identity line. Pearson's correlation coefficient is reported ($p<0.001$, $n=2160$).



Extended Data Fig. 5. Measurement precision. a, Proportion of unique networks having replicates. 56.0% of the networks in the final dataset have at least duplicates and 18.3% of them have at least 10 replicates. **b**, Distribution of node fraction standard deviation where standard deviation is measured from the networks with at least 5 replicates. The dotted grey line is the mean of the distribution of standard deviation. The result indicates a mean precision of ~6% in species fractions. **c**, **c**,The complete experiment (Methods, Extended Data Fig. 1) was replicated and node fractions in both replicates were measured. Node fractions measured in one experimental replicate are plotted against the other. Data is restricted to the common set of networks present in the two replicates. Data points are binned in 20 linearly spaced bins according to their x-axis component. Bins with less than 10 points are discarded. The box extends from the lower to upper quartile values of the data, with a dot at the mean. The whiskers extend from the 5th percentile to the 95th. Flier points are those past the end of the whiskers. The dotted orange line is the identity line. Pearson's correlation coefficient is reported ($p<0.001$, $n=2047$). These technical replicates demonstrated high repeatability ($r=0.84$ between species fractions) **d-e**, Control for cross-talk and quantification bias during droplet barcoding and sequencing. Barcoded sequencing was performed on a mix of four emulsions (A, B, C and D) each containing a mixture of ribozymes in known

proportions as well as a pair of specific hairpin RNA reporters for identification (Methods). **d**, Measured proportions of each population of droplets (orange) and expectations (green). The result indicates 87% of sequenced droplet contained a single set of ribozymes which demonstrate a low cross-talk between droplets. **e**, Measured against expected ribozymes concentrations (normalized) aggregating data from the four emulsions show high correlation ($r = 0.91$) and rule out biases of the droplet level sequencing. Each dot corresponds to a ribozyme in one of the emulsions (4 ribozymes in each of the 4 emulsions). Error bars are standard deviations. Grey dotted line is the identity line. Pearson correlation coefficient $p<0.001$ in all cases.

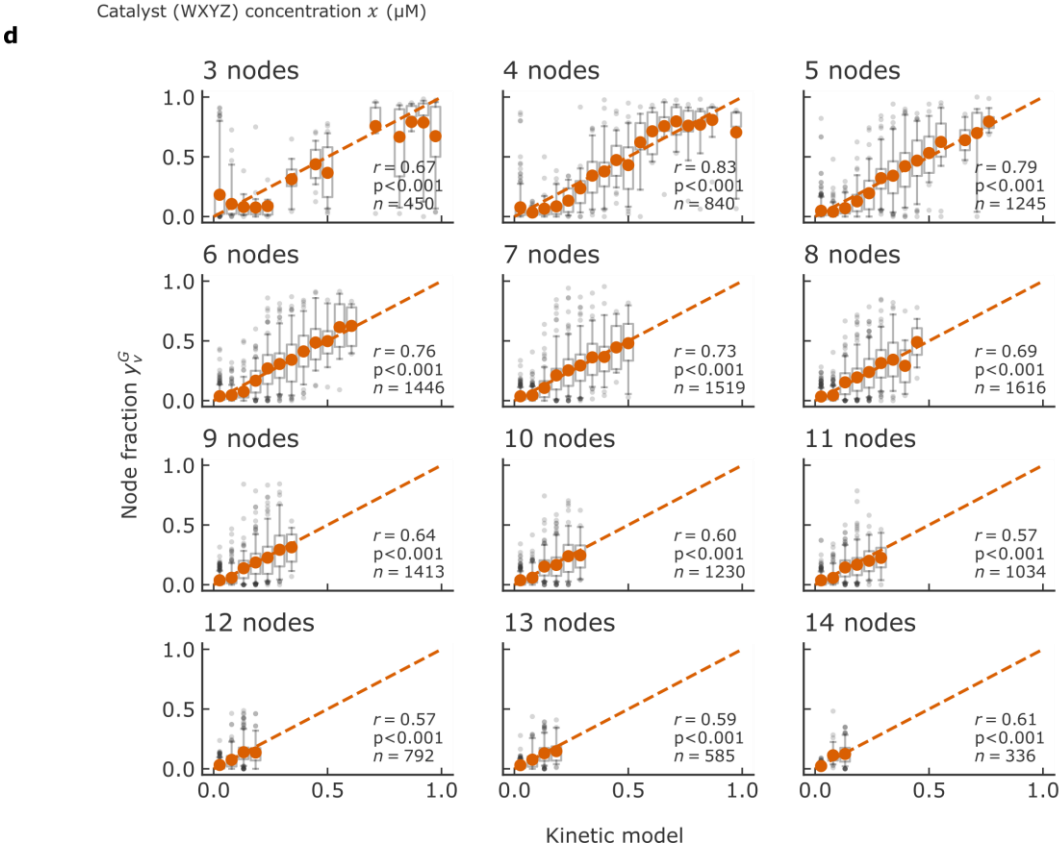




c

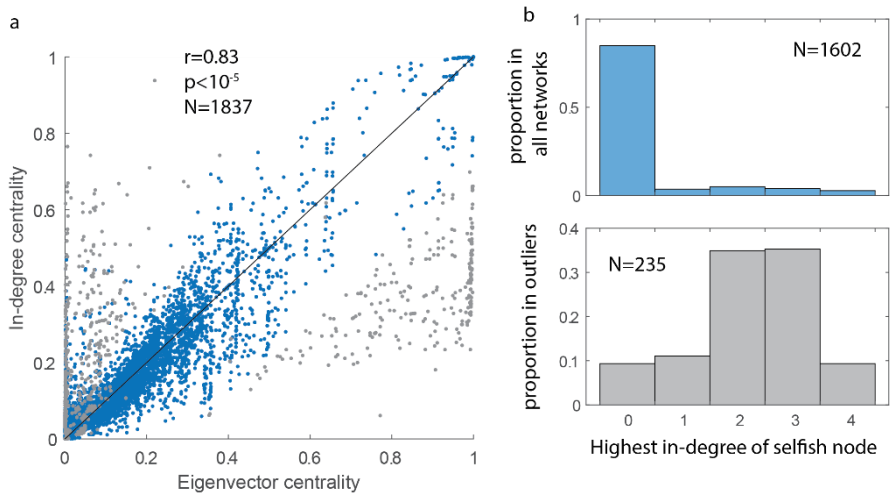
$$v_0 = \alpha x + \beta$$

Interaction $M \to N$	α (min^{-1})	β (μM.min^{-1})	r^2
$A \to U$	0.001352 ± 0.000041	0.000038 ± 0.000004	0.9975
$U \to A$	0.001575 ± 0.000097	0.000150 ± 0.000005	0.9998
$C \to G$	0.004950 ± 0.000283	0.000474 ± 0.000006	0.9993
$G \to C$	0.003008 ± 0.000335	0.000552 ± 0.000051	0.9833
$G \to U$	0.000160 ± 0.000007	0.000015 ± 0.000000	0.9997
$U \to G$	0.000066 ± 0.000001	0.000004 ± 0.000000	0.9986
$A \to G$	0.000018 ± 0.000000	0.000000 ± 0.000000	0.9279

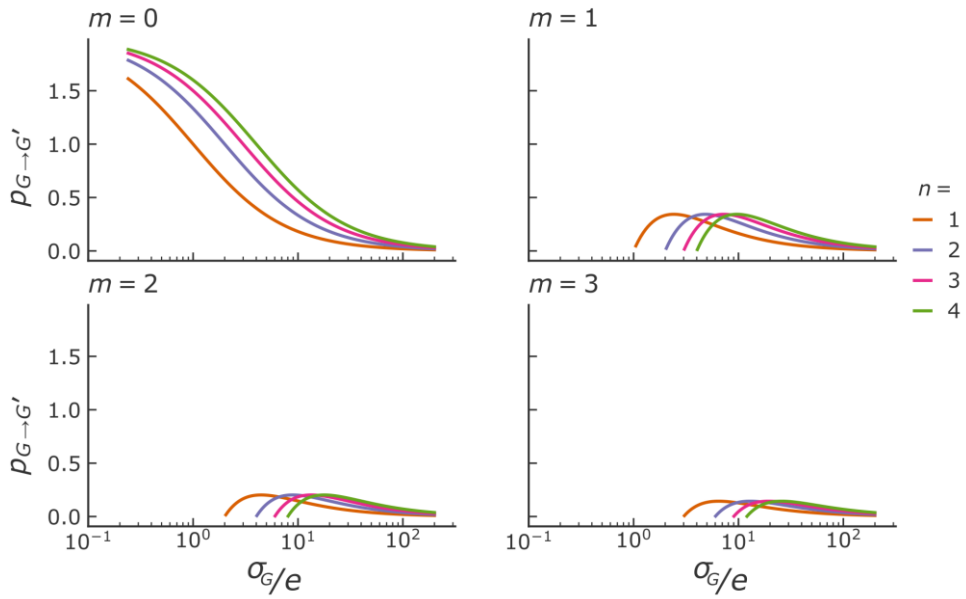


Extended Data Figure 6. Fitting of edge weights in *Azoarcus* networks across various models of node centrality and comparison of the best model with experimental measurements. a, For each pair of nodes, fractions are extracted in all networks that contain the two nodes. The color code indicates the proportion of cases where one node (row) has higher fraction compared to the other node (column). **b-c**, Experimental measurements to extract catalytic parameters used to describe the *Azoarcus* network

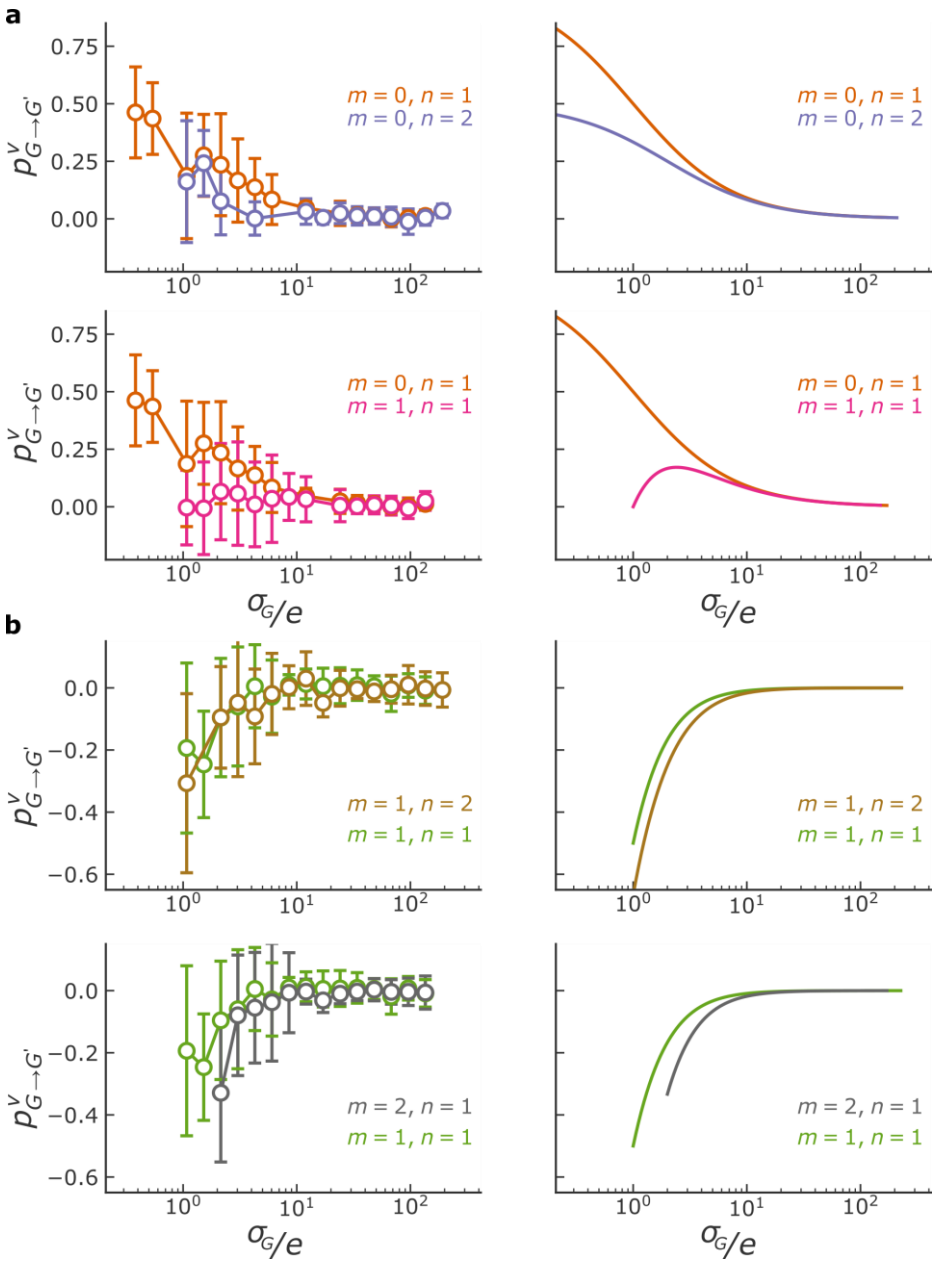
system[21,35] (Methods). **b**, Plot showing reaction kinetics for all four Watson-Crick IGS/tag interactions. Initial rates (v_0) of WXYZ formation from WXY and Z RNA fragments were measured by doping the reaction with covalent WXYZ ribozyme with the same IGS and tag as the WXY fragment, as reported by Yeates *et al*.[21]. **c**, Table showing the α and β parameters derived from fitting a linear relationship (top) between the initial rate of formation of WXYZ ribozymes (v_0) as a function of doped covalent WXYZ ribozyme concentration (x)[21,35]. Here α quantifies the synthesis of WXYZ ribozymes catalyzed by covalent ribozymes and, is derived from the slope of this relationship (plot in **b**). Whereas, β quantifies the synthesis of WXYZ ribozymes by non-covalent ribozymes and is derived from the intercept of this relationship (plot in **b**). **d**, Fraction obtained with the kinetic model versus measured fraction for networks with same number of nodes. Bins with less than 10 points are discarded. Dark grey dots are averages, quartile boxplots have 5th - 95th percentiles whiskers with flier points. The dotted orange line is the identity line.



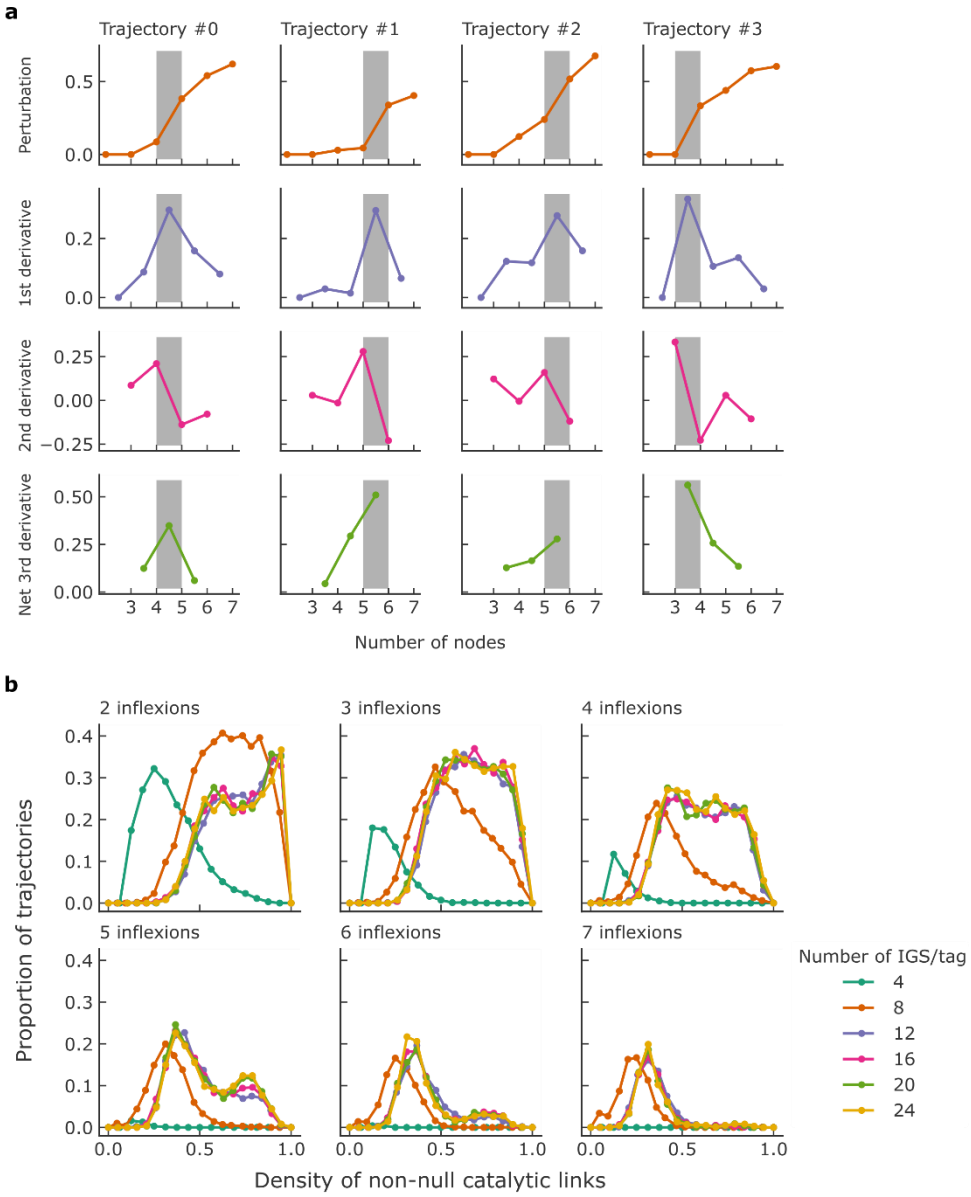
Extended Data Fig. 7. Network composition predictions using eigenvector centrality and in-degree centrality **a**, Covalent ribozyme fraction predicted by eigenvector centrality versus in-degree centrality for all networks of the experimental dataset (N=1837). In grey are species participating to networks containing outlier ribozymes (fraction underestimated by 20% or more by the in-degree centrality). **b**, Networks containing outliers tend to have selfish nodes (bottom) compared to others (top). Selfishness is quantified here as the in-degree of nodes with self-loops and no out-going edges.



Extended Data Fig. 8. Analytical predictions of Equation 1 with all possible values of the parameters n and m. Network perturbation $p_{G\to G'}$ upon addition of a new node is plotted against normalized background strength σ_G/e for all possible values of the catalytic novelty m (panels) and of perturbation b*read*th n (color coded).

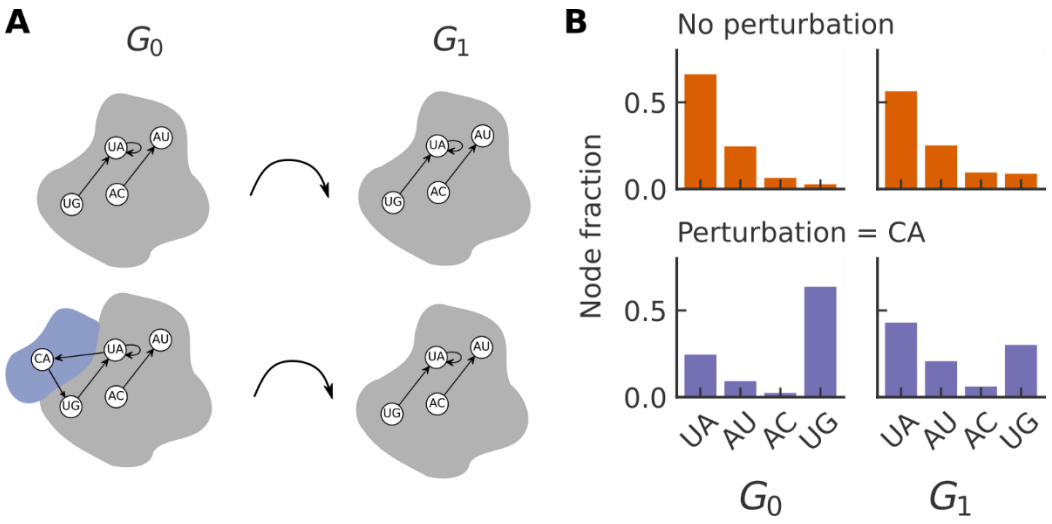


Extended Data Fig. 9. Node perturbation $p^v_{G\to G'}$ upon addition of a new node a to network G, comparing experimental data with analytical predictions. a, $p^v_{G\to G'}$ plotted against normalized background strength σ_G/e for different values of catalytic novelty *m* and perturbation b*read*th n in the case where v is a target for a and for both the experimental data (left) and the analytical results (right). For details of how analytical predictions were derived see Supplementary File 1. Data points were distributed along the x-axis in 20 log-spaced bins between $10^{-0.5}$ and $10^{2.5}$. Bins with less than 3 points were discarded and the mean and standard deviation for each bin are plotted. Networks with less than 3 nodes were discarded for the analysis. **b**, Same as **a** but the case where v is a not target for a.



Extended Data Fig. 10. Determination of strong inflexion points and influence of the number of different IGS/tag interactions on the number of inflexion points per trajectory. a, Cumulative perturbation for four representative trajectories from the experimental data are plotted (first row) as well as the first derivative (second row), the second derivative (third row) and the net third derivative (fourth row). The derivative is the difference between two successive points divided by the width of the interval (here 1). Inflexion points are first detected with the second derivative when it switches from positive to negative and the sharpness of these is measured with the net third derivative. To calculate the derivative we take the difference of values between n and n+1 nodes. Strong inflexion points are defined as being in the top 25% in sharpness and are indicated with dark grey rectangles. An extra-step with no perturbation is added at the beginning of the trajectory to detect starting strong inflexion points. **b**, The proportion of trajectories with a given number of inflexions is plotted against catalytic density for different numbers of IGS/tag combinations. Note that panels corresponding to 6 and 7 inflexions show no trajectory with inflexions when the number of IGS/tag pairs equals 4. Hence analysis of trends holds only when the number of IGS/tag pairs exceeds the number of inflexions, as mentioned in the main text. These results were obtained by simulating 20,000 trajectories as a function of catalytic density values

(defined as the proportion of the non-null specific interaction between an IGS and a tag over the complete set of possible interactions) for various number of IGS/tag. See Fig. 4gfor the details.



Extended Data Figure 11. Relaxation of chemical compositions to states due to non-covalent catalysts. a, Schematic of two serial transfer experiments with different initial conditions and two generations, G_0 and G_1, incubated for an hour at 48°C as in the droplet experiment. Top: control experiment (unperturbed network), where only the fragment for the AC, AU, UA and UG ribozymes are provided at each round, with concentrations 0.1 µM for WXY fragments and 1.6 µM for Z. Bottom: same as top, but the network is perturbed at G_0 by addition of CA fragments at the same concentration. **b**, Relative fraction of covalent ribozymes AC, AU, UA and UG produced during each generation of the serial transfer. Although introduction of fragments at G_0 has a significant impact on the final covalent ribozyme composition, these differences are not preserved at G_1 where the composition relaxes toward the one of the unperturbed network.

Extended Data Table 1. Different combination of WXY fragments in initial 5 pL emulsions.

Combination no.	$_{gMg}WXY_{cNu}$ fragments (MN)[#]	Hairpin RNA reporter variable region[¶]
1	UA	UUUU
2	CA	ACUG
3	GA	AAUU
4	AG	GAUA
5	CU	GUGG
6	GC	UAUG
7	UU	CCCA
8	CU, UG	AAGG
9	AC	GGAA
10	CA, GA, CG, UA, AU	AUAG
11	GG, GC	UAAC
12	AU	UGUA
13	CA, UC, GA, CC	UACA
14	AC, AG	CGUG
15	UA, GU, GG, CG, UG, AG, UU, UC	GCCG
16	GG	GAAU
17	AU, CA, AA, AG	GGGU
18	GC	CGCC
19	GU	UUGG
20	CC, GU, CA	GUCC
21	GG, GU, CC	UAGC
22	UU, CU	AAAA
23	CU, CA	CUAA
24	UC	AUGC

[#]All 16 possible $_{gMg}WXY_{cNu}$ RNA fragments were used where 'M' is the middle nucleotide of IGS and 'N' is the middle nucleotide of tag[18]. [¶]4 nucleotide variable region of the hairpin RNA reporter added to barcode the composition.

Extended Data Table 2. Variation in growth for different sizes of networks.

network size	minimum growth	maximum growth	number of networks	fold difference (max/min growth)
2	0.002	0.095	44	48.4
3	0.009	0.173	150	19.5
4	0.006	0.295	210	51.7
5	0.005	0.585	249	129.9
6	0.004	0.638	241	167.5
7	0.019	0.563	217	29.4
8	0.012	0.663	202	55.2
9	0.034	0.715	157	21.1
10	0.039	0.974	123	25.1
11	0.084	1.624	94	19.4
12	0.134	0.777	66	5.8
13	0.187	0.782	45	4.2
14	0.180	0.856	24	4.8
15	0.200	0.439	8	2.2

Supplementary Information File 1

for

Trade-offs between Darwinian properties in autocatalytic RNA networks

Sandeep Ameta[1,ǂ], Simon Arsène[1,ǂ], Sophie Foulon[1], Baptiste Saudemont[1], Bryce E. Clifton[2], Andrew D. Griffiths[1], Philippe Nghe[1]**

ǂSandeep Ameta and Simon Arsène contributed equally to the work.

[1]Laboratoire de Biochimie, CNRS UMR8231, Chimie Biologie Innovation, ESPCI Paris,

10 Rue Vauquelin, 75005, Paris, France.

[2]Department of Chemistry, Portland State University,

Portland, Oregon 97207, USA.

E-mail: andrew.griffiths@espci.fr, philippe.nghe@espci.fr

The Supplementary Information file contains the following sections:

1. Supplementary Methods
2. Supplementary Equations

Supplementary Methods

Quality control of hydrogel beads.

The quality of the barcoded beads was checked by analyzing the percentage of full-length barcodes in the final library as well as the diversity of barcodes on beads. Percentage of full-length barcodes was analyzed by subjecting a small portion of the beads to restriction endonuclease treatment (BclI, New England Biolabs, Product No.: R0160L). The barcodes beads have unique BclI site in the first common adaptor which is also used to release barcodes during reverse transcription in droplets (see Methods). The beads were centrifuged and supernatant was analyzed on TapeStation (Agilent 2200 TapeStation, using high sensitivity D1000 ScreenTape®, Product No.: 5067-5584). On average ~70% of the total oligonucleotide released contained full-length barcodes (Extended Data Fig. 2c).

To check the diversity of barcoded primers on beads, single beads were sorted into wells, containing 4 µL of water, of a 96-well plate using a fluorescence activated cell sorter (FACS; ARIA III, BD Biosciences, San Diego), at the Plate-forme de Cytometrie et d'Immuno-biologie, CYBIO, Institute Cochin, Paris). Then a reverse transcription reaction, in a total volume of 10 µL, was performed inside the plate using the same protocol and conditions as described in the Methods section for droplets, but using a row-specific hairpin RNA reporter (~0.1 µM) as template. This generate cDNAs with the bead barcode coupled to the row-specific hairpin reporter sequence. After 1 h incubation, 2 µl of the RT reaction from each well was transferred to a new plate and two sequential PCRs were performed to prepare samples for sequencing. The first PCR was performed in a volume of 20 µL using a column-specific forward primer (Oligo 16th sample barcodes 1 to 12, Supplementary File 2) and a common reverse primer (Oligo 17, Supplementary File 2). PCR products were purified separately from each well using 1.2 equivalent of AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) and re-suspended in 20 µL of water. 4 µL of purified products were used for second 40 µL PCR, using common forward and reverse primers (Oligo 18 and 19 respectively, Supplementary File 2) for all the wells. Thermocycling was for both PCRs was as described in Methods for the droplet experiments. After the second PCR, products were analysed on 1% agarose gels. 71/96 wells showed amplification. Then, 1 µL of PCR products from the positive wells were pooled,

purified using 1.2 equivalent of AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881), and sequenced on a HiSeq 4000 system (2*150 High Output mode at BGI Sequencing, Hong-Kong, China).

The sequencing data was processed as described in Methods. No filter based on the number of *reads* per UMIs was applied. UMI-normalized *reads* were associated to a bead using row and column-specific barcoding (rows coded by specific hairpin RNA reporter for each row; columns coded barcode introduced during the first PCR). For each bead (n=71), we computed the percentage of the most common barcode and found that all expect 3 beads had a mean value of 95.3 ±2.3% (Extended Data Fig 2d). The three beads with values significantly lower (around 50%), contained two dominant barcodes, each comprising ~50% of the *reads*, consistent with the presence of two beads in these wells.

DNA templates for *in vitro* transcription of ribozyme fragments

In order to sequence the 'IGS' at the 5' end of the ribozyme (WXYZ) RNA (without erasing it with the forward primer during PCR and to avoid biases from a ligation-based strategy), the central nucleotide of the IGS sequence was duplicated within all 16 WXY fragments. This was achieved by mutating the 25th nucleotide of the WXY fragment ('A' in the wild-type ribozyme, as used in the study of Vaidya *et al.*[18]) to either C, U, or G with complementary base-pairing mutations at the 7th nucleotide using site-directed mutagenesis. For this, wild-type DNA template (~0.25 pg/µL) was amplified by PCR using forward and backward M25 mutation primers (Oligo 1 to 4 and Oligo 5, respectively, Supplementary File 2). The respective PCR product was cloned in pJET2.1 vector using CloneJET PCR cloning kit (Thermo Scientific, Product No.: K1231) following the manufacturer's protocol. The cloned products were transformed in chemical competent *E.coli* (Top10 Chemical competent cells, Thermo Scientific, Product No.: C404010), incubated on LB agar-ampicillin plates, positive colonies were selected, and plasmids were isolated (using NucleoSpin® Plasmid kit from Macherey-Nagel, Product No.: 740588.10) and sequenced by Sanger sequencing (GATC Biotech). The clones with the correct sequence were used as template in PCR reactions to obtain dsDNA templates for the *in vitro* transcription in order to generate all 16 combinations of WXY RNA fragments ($_{gMg}$WXY$_{cNu}$). For PCRs, plasmids (at 25 pg/µL) were mixed with 0.5 µM of respective

primers (Oligo 1-4 and 6-9, Supplementary File 2) as forward and reverse primers in 1x PCR buffer (Thermo Scientific), 0.2 mM dNTPs, 0.01 U/µL of polymerase (Thermo Scientific Phusion Hot Start II, Product No.: F459) using the following protocol: initial denaturation 98°C/30sec, then 25 cycles of denaturing 98°C/10 s, annealing and extension 57°C/1 min, and a final extension of 72°C/5 min. PCR products were ethanol precipitated by adding 1/10th volume of 3M Na-Ac and 1.2 volume of 100% ethanol to the PCR reaction and centrifuging at 13.6 rcf for 60 min at 4°C. Pellets were vacuum dried, re-suspended in 20 µL of water and used for *in vitro* transcription reactions as described in methods. The Z fragment RNA used here was custom synthesized (PAGE purified) by IDT DNA technologies and used without further purification.

Checking identification of IGS sequences in WXY RNA fragments using the mutations at positions 7 and 25

All 16 WXY RNA fragments were poly(A) tailed, ligated with an RNA adaptor at their 5'end, converted to cDNA, appended with sequencing adaptors, and sequenced. For poly(A) tailing, 2.5 µM of each WXY RNA (in a separate reaction) was mixed with 1X reaction buffer (50 mM Tris-Hcl pH 7.9, 250 mM NaCl, 10 mM $MgCl_2$), 2 mM ATP, 50 U/µL of *E. coli* poly(A) polymerase (New England Biolabs, Product No.: M0276S) and incubated at 37°C for 40 min. After heat inactivation (70°C, 10 min) and isopropanol precipitation, the RNA were re-suspended in water, mixed with 1X reaction buffer (10 mM Tris-HCl pH 8.0, 5 mM $MgCl_2$, 100 mM KCl, 0.02% Triton X-100, 0.1 mg/mL BSA)), 0.1 U/µL of phosphatase enzyme (FastAP, Thermo Fisher Scientific, Product No.: EF0652) and incubated at 37°C for 1 h to remove the 5'-triphosphate RNA. After heat inactivation (70°C, 10 min), the dephosphorylated RNA reaction (20 µL) was subjected to phosphorylation (adding 5'-monophosphate) reaction (in 40µL volume) by adding 1X reaction buffer (50 mM Tris-HCl pH 7.6, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM spermidine), 2 mM ATP, 0.25 U/µL of T4 Polynucleotide Kinase (Thermo Fisher Scientific, Product No.: EK0031) and incubating at 37°C for 1 h. After heat inactivation (70°C, 10min), RNAs were purified on AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881). The mono-phosphorylated RNA (10 µL) were ligated (in 20 µL volume) to an RNA adaptor (Oligo 21, Supplementary File 2) by mixing with 1X reaction buffer (50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 mM DTT), 2 mM of ATP, 3% of PEG8000, 1 µM of adaptor RNA, 2.2 U/µL of T4 RNA ligase 1 (New England Biolabs, Product No.: M0204S) and incubating at 16°C overnight.

After heat inactivation, RNAs were purified on AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) and reverse transcribed. For reverse transcription (RT), purified ligated RNAs (10 µL) were mixed with 1X reaction buffer (50 mM Tris-HCl pH 8.3, 75 mM KCl, 3 mM $MgCl_2$), 0.5 mM of each dNTP, 5 mM of DTT, 2.5 µM of RT primer (Oligo 22, Supplementary File 2), 10 U/µL of Superscript III enzyme (Thermo Fisher Scientific, Product No.: 18080085) in reaction volume of 10µL and incubated at 55°C for 1 h. After RT, cDNAs were purified using AMPure XP magnetic beads (Beckman Coulter, Product No.: A63881) and amplified by PCR to append sequencing adaptors (see Methods, with Oligo 23, 17 as forward and reverse primer, respectively in PCR I and Oligo 18, 19 as forward and reverse primer, respectively in PCR II, Supplementary File 2). Final amplicons were analyzed by 2*150 bp pair-end sequencing (microMiSeq, Institute Curie High Throughput Sequencing Platform, Paris). For each WXY fragment, the corresponding *reads* were first analyzed to extract the sequence of the IGS and the two internal mutations (at positions 7 and 25) coding for the IGS. The results show that ~97% of the IGS correctly matched the expected nucleotides at positions 7 and 25.

Droplet barcode, UMI, IGS, tag and hairpin RNA reporter identification from sequencing data

Droplet barcode identification: The droplet barcode is composed of three 16 nucleotide long variable regions denoted as 'indexes' separated by 4 nucleotide long constant regions denoted as 'linkers'. First the linker positions are searched using a sliding window approach around their expected positions in a window of 5 bp and allowing a maximum Hamming distance of 1. The sequence in between two identified linkers is then extracted and aligned against the 96 possible variants (Supplementary File 2) for the corresponding index using Bowtie 2 version 2.2.9 with the following parameters: --very-sensitive --reorder -t. Alignments with a mapping quality lower than 20 were discarded. This gives the identity of the 3 indexes composing the droplet barcode for each *read*. *UMI identification*: To identify the UMI, first the position of the constant 15 bp sequence before the UMI is searched around its expected position in a window of 20 bp, allowing a maximum Hamming distance of 2. If a position is found and, if the last 3 bp exactly match the end of the expected constant 15 bp sequence, then the following 8 bp are extracted as the UMI for the *read*. *IGS identification*: Position 25 in the *Azoarcus* ribozyme, which corresponds to position 31 in *read* 1, codes for the IGS. To identify the base at position

31 of *read* 1, an upstream constant 7 base region (which starts at position 23) is searched in a window of 10 bp allowing a maximum Hamming distance of 2. If a position is found, and if the 3 bases after this stretch are in the following format: 'CMA' then 'M' is extracted as the IGS for the *read*. *Tag identification*: Tag identification is similar to IGS identification. The constant region 10 bases upstream of the tag ('CNU') in *read* 1 is searched around its expected position in a window of 10 bp allowing a maximum Hamming distance of 2. If a position is found and if the 3 bases, located 10 bp after this position, are in the following format: 'CNU', then N is extracted as the tag for the *read*. *Hairpin RNA reporter identification*: The hairpin RNA reporter sequence contains a variable 4 nucleotide region in the loop region which codes for its identity. To test whether a *read* corresponds to a hairpin reporter, the constant 10 bases region before and after the variable region are searched around their expected positions in windows of 10 bases and allowing a maximum Hamming distance of 2. If the two sequences are found the *read* is declared to be a hairpin RNA reporter *read*. The variable 4 nucleotides between these two constant regions are then used to assign the hairpin reporter identity. This sequence is compared to the list of possible variable region and the closest one in terms of Hamming distance (but not more distant than 1) is taken as the hairpin RNA reporter identity.

Determination of network composition in each droplet

To determine which *Azoarcus* network is present in each droplet, the ribozyme composition (obtained with the WXYZ *reads* where 'IGS' and 'tag' are identified, see sections above) is compared with the expected network structure from molecular hairpin RNA reporters. For this, the 24 initial 5 pL droplet emulsions are additionally barcoded with these hairpin RNA reporters as an independent measurement of expected network composition inside droplets after fusion. This allows thresholds to be chosen in sequencing data analysis, to correct compositional data, and to assign the network composition in each droplet. These molecular RNA reporters are small hairpin RNAs which contain a 4 nucleotide long unique barcode in the loop region and share the same primer binding site as *Azoarcus* ribozyme in the stem part so that they can be specifically barcoded per droplet and sequenced together with WXYZ ribozymes.

All the 24 hairpin RNA reporters were produced by *in vitro* transcription as described in[25]. Templates for transcription were generated by PCR using 0.0025 µM of synthetic ssDNA as template and 0.5 µM of each forward and reverse primer (Oligo 20, and 5, respectively Supplementary File 2). To check the effect of these molecular reporters, *Azoarcus* recombination reactions were performed in the presence or absence of hairpin RNAs. For this, 0.5 µM each of WXY (with 'g**A**g' as IGS and 'c**U**u' as tag) and Z RNA fragments were mixed with 1x reaction buffer (30 mM EPPS pH 7.4, 20 mM $MgCl_2$). This reaction was carried out with or without 36 nM of hairpin RNA reporter and incubated for 6 h. Sample were withdrawn at regular time intervals and analysed on a 12% denaturing polyacrylamide gel. The formation of WXYZ ribozymes were calculated from the band intensity using ImageJ software (https://imagej.nih.gov/ij/). The results show that there is no impact of these RNA reporters on the *Azoarcus* recombination reactions as the time-courses are indistinguishable.

Supplementary Equations

Analytical expressions for node-level perturbation

We consider the addition of a node a to the network G resulting in network $G' = (V', E')$ as $V' = V + \{a\}$. Fraction of node v in G', $y_v^{G'}$ can be taken relative only to the nodes in V (i.e. excluding a giving $y_v^{G'}|_V = \frac{y_v^{G'}}{\sum_{u \in V} y_u^{G'}}$). Node v response to the perturbation induced by the addition of node a to network G can be measured by $p_v^{G \to G'} = y_v^{G'}|_V - y_v^G$. Under the assumption that $y_v^G \sim \frac{D_v^{in,G}}{\sigma_G}$, $p_v^{G \to G'}$ can be first reformulated as:

$$p_v^{G \to G'} = \frac{\frac{D_v^{in,G'}}{\sigma_{G'}}}{\sum_{u \in V} \frac{D_u^{in,G'}}{\sigma_{G'}}} - \frac{D_v^{in,G}}{\sigma_G}$$

$$p_v^{G \to G'} = \frac{D_v^{in,G} + e_{a \to v}}{\sigma_G + D_a^{out,G}} - \frac{D_v^{in,G}}{\sigma_G}$$

$$p_v^{G \to G'} = \frac{e_{a \to v} - \frac{D_v^{in,G} D_a^{out,G}}{\sigma_G}}{\sigma_G + D_a^{out,G}}$$

Here, $e_{a \to v}$ is the weight of the directed edge between the new node a and node v. We can now distinguish two cases depending on whether or not the node v is a target for the new node a. If v is a target for a, then $e_{a \to v} = e$, and we can introduce the normalized in-degree $m_v = \frac{D_v^{in,G}}{e}$, the normalized sum of all the weights in the network $\sigma_e = \frac{\sigma_G}{e}$ and $n = \frac{D_a^{out,G}}{e}$, which is the number of targets of a in the network. Then dividing the obtained $p_v^{G \to G'}$ expression by e gives:

$$p_v^{G \to G'} = \frac{1 - n \frac{m_v}{\sigma_e}}{\sigma_e + n} \quad (i)$$

Because $\sigma_G \geq n m_v e$, in this case $p_v^{G \to G'} \geq 0$.

If v is not a target for a, then $e_{a \to v} = 0$ but we can still divide by e which is the typical weight of an outgoing edge from a. With the same parameters introduced before, it gives in this case:

$$p_v^{G \to G'} = \frac{-n \frac{m_v}{\sigma_e}}{\sigma_e + n} \ (ii)$$

In this case, it is clear that $p_v^{G \to G'} \leq 0$.

Analytical expression for network-level perturbation

Let $V_a = \{v \in V, e_{a \to v} > 0\}$ be the set of nodes in G for which a is a catalyst and $V_a^* = \{v \in V, e_{a \to v} = 0\}$ the set of nodes for which it is not. We assume that a is a catalyst for at least one node ($n > 0$) and that there is at least one link in G ($\sigma_G > 0$). The network's total response is:

$$p_{G \to G'} = \sum_{v \in V} |p_v^{G \to G'}|$$

Using expressions (i) and (ii) derived above, we obtain:

$$p_{G \to G'} = \sum_{v \in V_a} \left(\frac{1 - n \frac{m_v}{\sigma_e}}{\sigma_e + n}\right) + \sum_{v \in V_a^*} \frac{n \frac{m_v}{\sigma_e}}{\sigma_e + n}$$

$$p_{G \to G'} = \frac{n}{\sigma_e + n}\left(1 + \frac{1}{\sigma_e}\left(-\sum_{v \in V_a} m_v + \sum_{v \in V_a^*} m_v\right)\right)$$

Let m be the number of nodes in G with the same IGS as a, in other words, these m nodes and a are similar catalysts. It follows that $m = \sum_{v \in V_a} m_v / n$ and that $\sum_{v \in V_a^*} m_v = \sigma_e - nm$. This gives us the expression presented in the main text:

$$p_{G \to G'} = 2n \frac{1 - n \frac{m}{\sigma_e}}{\sigma_e + n} \ (1)$$

Bibliographie

Abate, A. R., Chen, C.-H., Agresti, J. J., & Weitz, D. A. (2009). Beating Poisson encapsulation statistics using close-packed ordering. *Lab on a Chip*, *9*(18), 2628. http://doi.org/10.1039/b909386a

Abate, A. R., Hung, T., Mary, P., Agresti, J. J., & Weitz, D. A. (2010). High-throughput injection with microfluidics using picoinjectors. *Proceedings of the National Academy of Sciences*, *107*(45), 19163–19166. http://doi.org/10.1073/pnas.1006888107

Abate, A. R., Poitzsch, A., Hwang, Y., Lee, J., Czerwinska, J., & Weitz, D. A. (2009). Impact of inlet channel geometry on microfluidic drop formation. *Physical Review E*, *80*(2), 026310. http://doi.org/10.1103/PhysRevE.80.026310

Agresti, J. J., Antipov, E., Abate, A. R., Ahn, K., Rowat, A. C., Baret, J.-C., … Weitz, D. A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences*, *107*(9), 4004–4009. http://doi.org/10.1073/pnas.0910781107

Ahrberg, C. D., Lee, J. M., & Chung, B. G. (2018). Poisson statistics-mediated particle/cell counting in microwell arrays. *Scientific Reports*, *8*(1), 2438. http://doi.org/10.1038/s41598-018-20913-0

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular Biology of the Cell. 4th edition* (Garland Sc). New York. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK26885/

Andrews, T. S., & Hemberg, M. (2018). Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, *59*, 114–122. http://doi.org/10.1016/j.mam.2017.07.002

Baker, S. C., Bauer, S. R., Beyer, R. P., Brenton, J. D., Bromley, B., Burrill, J., … Zadro, R. (2005). The external RNA controls consortium: A progress report. *Nature Methods*, *2*(10), 731–734. http://doi.org/10.1038/nmeth1005-731

Baret, J. C. (2012). Surfactants in droplet-based microfluidics. *Lab on a Chip*, *12*(3), 422–433. http://doi.org/10.1039/c1lc20582j

Bernstein, B. E., Meissner, A., & Lander, E. S. (2007). The Mammalian Epigenome. *Cell*, *128*(4), 669–681. http://doi.org/10.1016/j.cell.2007.01.033

Bollenbach, T., Quan, S., Chait, R., & Kishony, R. (2009). Nonoptimal Microbial Response to Antibiotics Underlies Suppressive Drug Interactions. *Cell*, *139*(4), 707–718. http://doi.org/10.1016/j.cell.2009.10.025

Brennecke, P., Anders, S., Kim, J. K., Kołodziejczyk, A. A., Zhang, X., Proserpio, V., … Heisler, M. G. (2013). Accounting for technical noise in single-cell RNA-seq experiments. *Nature Methods*, *10*(11), 1093–1098. http://doi.org/10.1038/nmeth.2645

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., … Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490. http://doi.org/10.1038/nature14590

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., … Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nature Biotechnology*, *33*(2), 155–160. http://doi.org/10.1038/nbt.3102

Butovsky, O., Jedrychowski, M. P., Moore, C. S., Cialic, R., Lanser, A. J., Gabriely, G., … Weiner, H. L. (2014). Identification of a unique TGF-β-dependent molecular and functional signature in microglia. *Nature Neuroscience*. http://doi.org/10.1038/nn.3599

Campbell, P. J., Stephens, P. J., Pleasance, E. D., O'Meara, S., Li, H., Santarius, T., … Futreal, P. A. (2008). Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, *40*(6), 722–729. http://doi.org/10.1038/ng.128

Cao, J., Packer, J. S., Ramani, V., Cusanovich, D. A., Huynh, C., Daza, R., … Shendure, J. (2017). Comprehensive single-cell transcriptional profiling of a multicellular organism. *Science*, *357*(6352), 661–667. http://doi.org/10.1126/science.aam8940

Chabert, M., Dorfman, K. D., & Viovy, J.-L. (2005). Droplet fusion by alternating current (AC) field electrocoalescence in microchannels. *ELECTROPHORESIS*, *26*(19),

3706–3715. http://doi.org/10.1002/elps.200500109

Chabert, M., & Viovy, J.-L. (2008). Microfluidic high-throughput encapsulation and hydrodynamic self-sorting of single cells. *Proceedings of the National Academy of Sciences*, *105*(9), 3191–3196. http://doi.org/10.1073/pnas.0708321105

Chattopadhyay, P. K., Gierahn, T. M., Roederer, M., & Love, J. C. (2014). Single-cell technologies for monitoring immune systems. *Nature Immunology*, *15*(2), 128–35. http://doi.org/10.1038/ni.2796

Chen, J., Zhou, Q., Wang, Y., & Ning, K. (2016). Single-cell SNP analyses and interpretations based on RNA-Seq data for colon cancer research. *Scientific Reports*, *6*(1), 34420. http://doi.org/10.1038/srep34420

Chhor, V., Le Charpentier, T., Lebon, S., Oré, M. V., Celador, I. L., Josserand, J., … Fleiss, B. (2013). Characterization of phenotype markers and neuronotoxic potential of polarised primary microglia In vitro. *Brain, Behavior, and Immunity*, *32*, 70–85. http://doi.org/10.1016/j.bbi.2013.02.005

Church, G. M., Vigneault, F., Laserson, U., & Bachelet, I. (2012). High-Throughput Immune Sequencing. US. Retrieved from https://patentimages.storage.googleapis.com/f1/f2/19/d7a12e9e739f8b/WO2012048340A2.pdf

Clark, M. B., Mercer, T. R., Bussotti, G., Leonardi, T., Haynes, K. R., Crawford, J., … Dinger, M. E. (2015). Quantitative gene profiling of long noncoding RNAs with targeted RNA sequencing. *Nature Methods*, *12*(4), 339–342. http://doi.org/10.1038/nmeth.3321

Clausell-Tormos, J., Lieber, D., Baret, J. C., El-Harrak, A., Miller, O. J., Frenz, L., … Merten, C. A. (2008). Droplet-Based Microfluidic Platforms for the Encapsulation and Screening of Mammalian Cells and Multicellular Organisms. *Chemistry and Biology*, *15*(5), 427–437. http://doi.org/10.1016/j.chembiol.2008.04.004

Cusanovich, D. A., Reddington, J. P., Garfield, D. A., Daza, R. M., Aghamirzaie, D., Marco-Ferreres, R., … Furlong, E. E. M. (2018). The cis-regulatory dynamics of embryonic development at single-cell resolution. *Nature Publishing Group*, *555*. http://doi.org/10.1038/nature25981

Das, A., Chai, J. C., Kim, S. H., Park, K. S., Lee, Y. S., Jung, K. H., & Chai, Y. G. (2015). Dual RNA Sequencing Reveals the Expression of Unique Transcriptomic Signatures in Lipopolysaccharide-Induced BV-2 Microglial Cells. *PLOS ONE*, *10*(3), e0121117. http://doi.org/10.1371/journal.pone.0121117

Das, A., Kim, S. H., Arifuzzaman, S., Yoon, T., Chai, J. C., Lee, Y. S., … Chai, Y. G. (2016). Transcriptome sequencing reveals that LPS-triggered transcriptional responses in established microglia BV2 cell lines are poorly representative of primary microglia. *Journal of Neuroinflammation*, *13*(1). http://doi.org/10.1186/s12974-016-0644-1

Deczkowska, A., Keren-Shaul, H., Weiner, A., Colonna, M., Schwartz, M., & Amit, I. (2018). Disease-Associated Microglia: A Universal Immune Sensor of Neurodegeneration. *Cell*, *173*(5), 1073–1081. http://doi.org/10.1016/j.cell.2018.05.003

Eberwine, J., Yeh, H., Miyashiro, K., Cao, Y., Nair, S., Finnell, R., … Coleman, P. (1992). Analysis of gene expression in single live neurons. *Proceedings of the National Academy of Sciences*, *89*(7), 3010–3014. http://doi.org/10.1073/pnas.89.7.3010

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic Gene Expression in a Single Cell. *Science*, *297*(5584), 1183–1186. http://doi.org/10.1126/science.1070919

Eyer, K., Doineau, R. C. L., Castrillon, C. E., Briseño-Roa, L., Menrath, V., Mottet, G., … Baudry, J. (2017). Single-cell deep phenotyping of IgG-secreting cells for high-resolution immune monitoring. *Nature Biotechnology*, *35*(10), 977–982. http://doi.org/10.1038/nbt.3964

Fan, H. C., Fu, G. K., & Fodor, S. P. A. (2015). Combinatorial labeling of single cells for gene expression cytometry. *Science*, *347*(6222). http://doi.org/10.1126/science.1258367

Fan, X., Zhang, X., Wu, X., Guo, H., Hu, Y., Tang, F., & Huang, Y. (2015). Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biology*, *16*(1). http://doi.org/10.1186/s13059-015-0706-1

Femino, A. M., Fay, F. S., Fogarty, K., & Singer, R. H. (1998). Visualization of single RNA transcripts in situ. *Science (New York, N.Y.)*, *280*(5363), 585–90. Retrieved from https://www.jstor.org/stable/2895287 RE

Franke, T., Abate, A. R., Weitz, D. A., & Wixforth, A. (2009). Surface acoustic wave (SAW) directed droplet flow in microfluidics for PDMS devices. *Lab on a Chip*, *9*(18), 2625. http://doi.org/10.1039/b906819h

Frenz, L., Blank, K., Brouzes, E., & Griffiths, A. D. (2009). Reliable microfluidic on-chip incubation of droplets in delay-lines. *Lab Chip*, *9*(10), 1344–1348. http://doi.org/10.1039/B816049J

Georgiou, G., Ippolito, G. C., Beausang, J., Busse, C. E., Wardemann, H., & Quake, S. R. (2014, February). The promise and challenge of high-throughput sequencing of the antibody repertoire. *Nature Biotechnology*. http://doi.org/10.1038/nbt.2782

Gerber, M., Isel, C., Moules, V., & Marquet, R. (2014). Selective packaging of the influenza A genome and consequences for genetic reassortment. *Trends in Microbiology*, *22*(8), 446–455. http://doi.org/10.1016/j.tim.2014.04.001

Gertig, U., & Hanisch, U.-K. (2014). Microglial diversity by responses and responders. *Frontiers in Cellular Neuroscience*, *8*, 101. http://doi.org/10.3389/fncel.2014.00101

Ginhoux, F., Greter, M., Leboeuf, M., Nandi, S., See, P., Solen, G., … & Merad, M. (2010). Fate Mapping Analysis Reveals That Adult Microglia Derive from Primitive Macrophages. *Science*, *330*(6005), 841–845. http://doi.org/10.1126/science.1194637

Ginhoux, F., & Prinz, M. (2015). Origin of microglia: Current concepts and past controversies. *Cold Spring Harbor Perspectives in Biology*, *7*(8). http://doi.org/10.1101/cshperspect.a020537

Grosselin, K., Durand, A., Marsolier, J., Poitou, A., Marangoni, E., Nemati, F., … Gérard, A. (2019). High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nature Genetics*, *51*(6), 1060–1066. http://doi.org/10.1038/s41588-019-0424-9

Grün, D., Kester, L., & Van Oudenaarden, A. (2014). Validation of noise models for

single-cell transcriptomics. *Nature Methods*, *11*(6), 637–640. http://doi.org/10.1038/nmeth.2930

Gubler, U. (1987). Second-Strand cDNA Synthesis: MRNA Fragments as Primers. *Methods in Enzymology*, *152*(C), 330–335. http://doi.org/10.1016/0076-6879(87)52038-9

Hagberg, H., Gressens, P., & Mallard, C. (2012). Inflammation during fetal and neonatal life: Implications for neurologic and neuropsychiatric disease in children and adults. *Annals of Neurology*, *71*(4), 444–457. http://doi.org/10.1002/ana.22620

Hagberg, H., Mallard, C., Ferriero, D. M., Vannucci, S. J., Levison, S. W., Vexler, Z. S., & Gressens, P. (2015). The role of inflammation in perinatal brain injury. *Nature Reviews. Neurology*, *11*(4), 192–208. http://doi.org/10.1038/nrneurol.2015.13

Hammond, T. R., Dufort, C., Dissing-Olesen, L., Giera, S., Young, A., Wysoker, A., … Stevens, B. (2018). Single-Cell RNA Sequencing of Microglia throughout the Mouse Lifespan and in the Injured Brain Reveals Complex Cell-State Changes. *Immunity*, 1–19. http://doi.org/10.1016/j.immuni.2018.11.004

Han, Q., Bagheri, N., Bradshaw, E. M., Hafler, D. A., Lauffenburger, D. A., & Love, J. C. (2012). Polyfunctional responses by human T cells result from sequential release of cytokines. *Proceedings of the National Academy of Sciences*, *109*(5), 1607–1612. http://doi.org/10.1073/pnas.1117194109

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine*. http://doi.org/10.1186/s13073-017-0467-4

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Reports*, *2*(3), 666–673. http://doi.org/10.1016/j.celrep.2012.08.003

Hatch, A. C., Fisher, J. S., Tovar, A. R., Hsieh, A. T., Lin, R., Pentoney, S. L., … Lee, A. P. (2011). 1-Million droplet array with wide-field fluorescence imaging for digital PCR. *Lab on a Chip*, *11*(22), 3838. http://doi.org/10.1039/c1lc20561g

Henn, A., Lund, S., Hedtjärn, M., Schrattenholz, A., Pörzgen, P., & Leist, M. (2009).

The suitability of BV2 cells as alternative model system for primary microglia cultures or for animal experiments examining brain inflammation. *ALTEX*, *26*(2), 83–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19565166

Heppner, F. L., Ransohoff, R. M., & Becher, B. (2015). Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience*, *16*(6), 358–372. http://doi.org/10.1038/nrn3880

Hu, Y., An, Q., Sheu, K., Trejo, B., Fan, S., & Guo, Y. (2018). Single Cell Multi-Omics Technology: Methodology and Application. *Frontiers in Cell and Developmental Biology*, *6*(April). http://doi.org/10.3389/fcell.2018.00028

Hutchinson, E. C., von Kirchbach, J. C., Gog, J. R., & Digard, P. (2010). Genome packaging in influenza A virus. *Journal of General Virology*, *91*(2), 313–328. http://doi.org/10.1099/vir.0.017608-0

Hwang, B., Lee, J. H., & Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine*, *50*(8), 96. http://doi.org/10.1038/s12276-018-0071-8

Isel, C., Munier, S., & Naffakh, N. (2016). Experimental approaches to study genome packaging of influenza a viruses. *Viruses*, *8*(8), 1–15. http://doi.org/10.3390/v8080218

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Research*, *21*(7), 1160–1167. http://doi.org/10.1101/gr.110882.110

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., … Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. http://doi.org/10.1038/nmeth.2772

Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., … Quake, S. R. (2014). Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science (New York, N.Y.)*, *343*(6172), 776–9. http://doi.org/10.1126/science.1247651

Kadhim, H., Tabarki, B., De Prez, C., Rona, A.-M., & Sebire, G. (2002). Interleukin-2

in the pathogenesis of perinatal white matter damage. *Neurology*, *58*(7), 1125–1128. http://doi.org/10.1212/WNL.58.7.1125

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., … Zinzen, R. P. (2017). The Drosophila Embryo at Single Cell Transcriptome Resolution. *Science*, *3235*(August), 117382. http://doi.org/10.1101/117382

Kay, A. W., Strauss-Albee, D. M., & Blish, C. A. (2016). Application of Mass Cytometry (CyTOF) for Functional and Phenotypic Analysis of Natural Killer Cells. *Methods in Molecular Biology (Clifton, N.J.)*, *1441*, 13–26. http://doi.org/10.1007/978-1-4939-3684-7_2

Keren-Shaul, H., Spinrad, A., Weiner, A., Matcovitch-Natan, O., Dvir-Szternfeld, R., Ulland, T. K., … Amit, I. (2017). A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell*, *169*(7), 1276–1290.e17. http://doi.org/10.1016/j.cell.2017.05.018

Kettenmann, H., Hanisch, U.-K., Noda, M., & Verkhratsky, A. (2011). Physiology of Microglia. *Physiological Reviews*, *91*(2), 461–553. http://doi.org/10.1152/physrev.00011.2010

Kintses, B., van Vliet, L. D., Devenish, S. R., & Hollfelder, F. (2010). Microfluidic droplets: new integrated workflows for biological experiments. *Current Opinion in Chemical Biology*, *14*(5), 548–555. http://doi.org/10.1016/j.cbpa.2010.08.013

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. http://doi.org/10.1038/nmeth.1778

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., … Kirschner, M. W. (2015). Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell*, *161*(5), 1187–1201. http://doi.org/10.1016/j.cell.2015.04.044

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015a). Molecular Cell The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, *58*, 610–620. http://doi.org/10.1016/j.molcel.2015.04.005

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015b). The Technology and Biology of Single-Cell RNA Sequencing. *Molecular Cell*, *58*(4), 610–620. http://doi.org/10.1016/j.molcel.2015.04.005

Kolodziejczyk, A. A., & Lönnberg, T. (2017). Global and targeted approaches to single-cell transcriptome characterization. *Briefings in Functional Genomics*. http://doi.org/10.1093/bfgp/elx025

Kornberg, R. D., & Thonmas, J. O. (1974). Chromatin Structure: Oligomers of the Histones. *Science*, *184*(4139), 865–868. http://doi.org/10.1126/science.184.4139.865

Köster, S., Angilè, F. E., Duan, H., Agresti, J. J., Wintner, A., Schmitz, C., … Weitz, D. A. (2008). Drop-based microfluidic devices for encapsulation of single cells. *Lab on a Chip*, *8*(7), 1110. http://doi.org/10.1039/b802941e

Krammer, F., Smith, G. J. D., Fouchier, R. A. M., Peiris, M., Kedzierska, K., Doherty, P. C., … García-Sastre, A. (2018). Influenza. *Nature Reviews Disease Primers*, *4*(1), 3. http://doi.org/10.1038/s41572-018-0002-y

Krishnan, M. L., Van Steenwinckel, J., Schang, A.-L., Yan, J., Arnadottir, J., Le Charpentier, T., … Gressens, P. (2017). Integrative genomics of microglia implicates DLG4 (PSD95) in the white matter development of preterm infants. *Nature Communications*, *8*(1), 428. http://doi.org/10.1038/s41467-017-00422-w

Kurimoto, K., Yabuta, Y., Ohinata, Y., Ono, Y., Uno, K. D., Yamada, R. G., … Saitou, M. (2006). An improved single-cell cDNA amplification method for efficient high-density oligonucleotide microarray analysis. *Nucleic Acids Research*, *34*(5). http://doi.org/10.1093/nar/gkl050

Levin, J. Z., Berger, M. F., Adiconis, X., Rogov, P., Melnikov, A., Fennell, T., … Gnirke, A. (2009). Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biology*, *10*(10), 115. http://doi.org/10.1186/gb-2009-10-10-r115

Link, D. R., Anna, S. L., Weitz, D. A., & Stone, H. A. (2004). Geometrically Mediated Breakup of Drops in Microfluidic Devices. *Physical Review Letters*, *92*(5), 054503. http://doi.org/10.1103/PhysRevLett.92.054503

Link, D. R., Grasland-Mongrain, E., Duri, A., Sarrazin, F., Cheng, Z., Cristobal, G., … Weitz, D. A. (2006). Electric Control of Droplets in Microfluidic Devices. *Angewandte Chemie International Edition*, *45*(16), 2556–2560. http://doi.org/10.1002/anie.200503540

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., … Black, R. E. (2012). Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *The Lancet*, *379*(9832), 2151–2161. http://doi.org/10.1016/S0140-6736(12)60560-1

Loane, D. J., & Kumar, A. (2015). Microglia in the TBI brain: The good, the bad, and the dysregulated. *Experimental Neurology*, *275*, 316–327. http://doi.org/10.1016/j.expneurol.2015.08.018

Lohr, J. G., Kim, S., Gould, J., Knoechel, B., Drier, Y., Cotton, M. J., … Golub, T. R. (2016). Genetic interrogation of circulating multiple myeloma cells at single-cell resolution. *Science Translational Medicine*, *8*(363), 363ra147-363ra147. http://doi.org/10.1126/scitranslmed.aac7037

Lowe, K. C. (2002). Perfluorochemical respiratory gas carriers: Benefits to cell culture systems. *Journal of Fluorine Chemistry*, *118*(1–2), 19–26. http://doi.org/10.1016/S0022-1139(02)00200-2

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., & Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, *13*(5). http://doi.org/10.1371/journal.pcbi.1005457

Lowen, A. C. (2017). Constraints, Drivers, and Implications of Influenza A Virus Reassortment. *Annual Review of Virology*, *4*(1), 105–121. http://doi.org/10.1146/annurev-virology-101416-041726

Lu, Y., Shen, Y., Warren, W., & Walter, R. (2016). Next Generation Sequencing in Aquatic Models. In *Next Generation Sequencing - Advances, Applications and Challenges* (Vol. i, p. 13). InTech. http://doi.org/10.5772/61657

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. http://doi.org/10.1016/j.cell.2015.05.002

Marcy, Y., Ishoey, T., Lasken, R. S., Stockwell, T. B., Walenz, B. P., Halpern, A. L., … Quake, S. R. (2007). Nanoliter Reactors Improve Multiple Displacement Amplification of Genomes from Single Cells. *PLoS Genetics*, *3*(9), e155. http://doi.org/10.1371/journal.pgen.0030155

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., & Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: Stochasticity in gene expression and RNA splicing. *Genome Research*, *24*(3), 496–510. http://doi.org/10.1101/gr.161034.113

Matcovitch-Natan, O., Winter, D. R., Giladi, A., Vargas Aguilar, S., Spinrad, A., Sarrazin, S., … Amit, I. (2016). Microglia development follows a stepwise program to regulate brain homeostasis. *Science*.

Matcovitch-Natan, O., Winter, D. R., Giladi, A., Vargas Aguilar, S., Spinrad, A., Sarrazin, S., … Amit, I. (2016). Microglia development follows a stepwise program to regulate brain homeostasis. *Science*, *353*(6301), aad8670-aad8670. http://doi.org/10.1126/science.aad8670

Mathys, H., Adaikkan, C., Gao, F., Young, J. Z., Manet, E., Hemberg, M., … Tsai, L.-H. (2017). Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Reports*, *21*(2), 366–380. http://doi.org/10.1016/j.celrep.2017.09.039

Mazutis, L., Baret, J.-C., & Griffiths, A. D. (2009). A fast and efficient microfluidic system for highly selective one-to-one droplet fusion. *Lab on a Chip*, *9*(18), 2665. http://doi.org/10.1039/b903608c

Mazutis, L., Gilbert, J., Ung, W. L., Weitz, D. A., Griffiths, A. D., & Heyman, J. A. (2013). Single-cell analysis and sorting using droplet-based microfluidics. *Nature Protocols*, *8*(5), 870–91. http://doi.org/10.1038/nprot.2013.046

McDaniel, J. R., DeKosky, B. J., Tanno, H., Ellington, A. D., & Georgiou, G. (2016). Ultra-high-throughput sequencing of the immune receptor repertoire from millions of lymphocytes. *Nature Protocols*, *11*. http://doi.org/10.1038/nprot.2016.024

Mercer, T. R., Clark, M. B., Crawford, J., Brunck, M. E., Gerhardt, D. J., Taft, R. J., … Mattick, J. S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature Protocols*, *9*(5), 989–1009.

http://doi.org/10.1038/nprot.2014.058

Metzker, M. L. (2010). Sequencing technologies — the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. http://doi.org/10.1038/nrg2626

Miron, V. E., Boyd, A., Zhao, J.-W., Yuen, T. J., Ruckh, J. M., Shadrach, J. L., … Ffrench-Constant, C. (2013). M2 microglia and macrophages drive oligodendrocyte differentiation during CNS remyelination. *Nature Neuroscience*, *16*(9), 1211–1218. http://doi.org/10.1038/nn.3469

Moreillon, P., & Majcherczyk, P. A. (2003). Proinflammatory Activity of Cell-wall Constituents from Gram-positive Bacteria. *Scandinavian Journal of Infectious Diseases*, *35*(9), 632–641. http://doi.org/10.1080/00365540310016259

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. http://doi.org/10.1038/nmeth.1226

Mrdjen, D., Pavlovic, A., Hartmann, F. J., Schreiner, B., Utz, S. G., Leung, B. P., … Becher, B. (2018). High-Dimensional Single-Cell Mapping of Central Nervous System Immune Cells Reveals Distinct Myeloid Subsets in Health, Aging, and Disease. *Immunity*, *48*(2), 380–395.e6. http://doi.org/10.1016/j.immuni.2018.01.011

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, *320*(5881), 1344–1349. http://doi.org/10.1126/science.1158441

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., … Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*(7341), 90–94. http://doi.org/10.1038/nature09807

Noda, T., Sagara, H., Yen, A., Takada, A., Kida, H., Cheng, R. H., & Kawaoka, Y. (2006). Architecture of ribonucleoprotein complexes in influenza A virus particles. *Nature*, *439*(7075), 490–492. http://doi.org/10.1038/nature04378

Olsen, T. K., & Baryawno, N. (2018). Introduction to Single-Cell RNA Sequencing. *Current Protocols in Molecular Biology*, *122*(1), e57.

http://doi.org/10.1002/cpmb.57

Paolicelli, R. C., Bolasco, G., Pagani, F., Maggi, L., Scianni, M., Panzanelli, P., … Gross, C. T. (2011). Synaptic Pruning by Microglia Is Necessary for Normal Brain Development. *Science*, *333*(6048), 1456–1458. http://doi.org/10.1126/science.1202529

Pasteur, I. (2014). Grippe : fiche maladie. Retrieved from https://www.pasteur.fr/fr/centre-medical/fiches-maladies/grippe

Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., … Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, *344*(6190), 1396–1401. http://doi.org/10.1126/science.1254257

Paul, L., Kubala, P., Horner, G., Ante, M., Hollaender, I., Alexander, S., & Reda, T. (2016). SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing. *BioRxiv*, *28*(1), 080747. http://doi.org/10.1101/080747

Peterson, V. M., Zhang, K. X., Kumar, N., Wong, J., Li, L., Wilson, D. C., … Klappenbach, J. A. (2017). Multiplexed quantification of proteins and transcripts in single cells. *Nature Biotechnology*, *35*(10), 936–939. http://doi.org/10.1038/nbt.3973

Picelli, S., Björklund, Å. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature Methods*, *10*(11), 1096–1100. http://doi.org/10.1038/nmeth.2639

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. *Nature Protocols*, *9*(1), 171–181. http://doi.org/10.1038/nprot.2014.006

Pollen, A. A., Nowakowski, T. J., Shuga, J., Wang, X., Leyrat, A. A., Lui, J. H., … West, J. A. A. (2014). Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nature Biotechnology*, *32*(10), 1053–1058. http://doi.org/10.1038/nbt.2967

Popiolek-Barczyk, K., & Mika, J. (2016). Targeting the microglial signaling pathways: new insights in the modulation of neuropathic pain. *Current Medicinal Chemistry*,

23, 2908–2928. http://doi.org/10.2174/09298673236661606071

Poulin, J. F., Tasic, B., Hjerling-Leffler, J., Trimarchi, J. M., & Awatramani, R. (2016). Disentangling neural cell diversity using single-cell transcriptomics. *Nature Neuroscience*, *19*(9), 1131–1141. http://doi.org/10.1038/nn.4366

Prakadan, S. M., Shalek, A. K., & Weitz, D. A. (2017). Scaling by shrinking: Empowering single-cell “omics” with microfluidic devices. *Nature Reviews Genetics*, *18*(6), 345–361. http://doi.org/10.1038/nrg.2017.15

Proserpio, V., & Mahata, B. (2016). Single-cell technologies to study the immune system. *Immunology*, *147*(2), 133–140. http://doi.org/10.1111/imm.12553

Raap, A. K., Van De Corput, M. P. C., Vervenne, R. A. W., Van Gijlswijk, R. P. M., Tanke, H. J., & Wiegant, J. (1995). *Ultra-sensitive FISH using peroxidase-mediated deposition of biotin-or fluorochrome tyramides*. *Human Molecular Genetics* (Vol. 4). Retrieved from https://api-istex-fr.insb.bib.cnrs.fr/ark:/67375/HXZ-FKRR2Q0D-Q/fulltext.pdf?auth=ip,fede&sid=ebsco,istex-view

Raj, A., van den Bogaard, P., Rifkin, S. A., van Oudenaarden, A., & Tyagi, S. (2008). Imaging individual mRNA molecules using multiple singly labeled probes. *Nature Methods*, *5*(10), 877–879. http://doi.org/10.1038/nmeth.1253

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., … Sandberg, R. (2012). Full-length mrNA-seq from single-cell levels of rNA and individual circulating tumor cells. *Nature Biotechnology*, *30*. http://doi.org/10.1038/nbt.2282

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., … Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182. http://doi.org/10.1126/science.aam8999

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology*, *33*(11), 1165–1172. http://doi.org/10.1038/nbt.3383

Saikia, M., Parker, J. S. L., Wang, M. F. Z., Moral-Lopez, P., De Vlaminck, I., Danko,

C. G., … Keshavjee, S. H. (2018). Simultaneous multiplexed amplicon sequencing and transcriptome profiling in single cells. *Nature Methods*, *16*(1), 59–62. http://doi.org/10.1038/s41592-018-0259-9

Salter, M. W., & Beggs, S. (2014). Sublime Microglia: Expanding Roles for the Guardians of the CNS. *Cell*, *158*(1), 15–24. http://doi.org/10.1016/j.cell.2014.06.008

Salter, M. W., & Stevens, B. (2017). Microglia emerge as central players in brain disease. *Nature Medicine*. http://doi.org/10.1038/nm.4397

Samji, T. (2009). Influenza A: understanding the viral life cycle. *The Yale Journal of Biology and Medicine*, *82*(4), 153–9. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20027280

Saunders, A., Macosko, E. Z., Wysoker, A., Goldman, M., Krienen, F. M., de Rivera, H., … McCarroll, S. A. (2018). Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell*, *174*(4), 1015–1030.e16. http://doi.org/10.1016/j.cell.2018.07.028

Schwabe, A., & Bruggeman, F. J. (2014). Single yeast cells vary in transcription activity not in delay time after a metabolic shift. *Nature Communications*, *5*. http://doi.org/10.1038/ncomms5798

Shaffer, S. M., Dunagin, M. C., Torborg, S. R., Torre, E. A., Emert, B., Krepler, C., … Raj, A. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature Publishing Group*, *546*. http://doi.org/10.1038/nature22794

Shahi, P., Kim, S. C., Haliburton, J. R., Gartner, Z. J., & Abate, A. R. (2017). Abseq: Ultrahigh-throughput single cell protein profiling with droplet microfluidic barcoding. http://doi.org/10.1038/srep44447

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., … Regev, A. (2014). Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature*, *510*(7505), 363–369. http://doi.org/10.1038/nature13437

Shin, D., Lee, W., Lee, J. H., & Bang, D. (2019). Multiplexed single-cell RNA-seq via transient barcoding for simultaneous expression profiling of various drug

perturbations. *Science Advances*, *5*(5), eaav2249. http://doi.org/10.1126/sciadv.aav2249

Shum, E. Y., Walczak, E. M., Chang, C., & Christina Fan, H. (2019). Quantitation of mRNA Transcripts and Proteins Using the BD Rhapsody™ Single-Cell Analysis System. In *Advances in experimental medicine and biology* (Suzuki Y, Vol. 1129, pp. 63–79). Springer, Singapore. http://doi.org/10.1007/978-981-13-6037-4_5

Simons, J. H., & Linevsky, M. J. (1952). The Solubility of Organic Solids in Fluorocarbon Derivatives. *Journal of the American Chemical Society*, *74*(19), 4750–4751. http://doi.org/10.1021/ja01139a007

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499. http://doi.org/10.1101/gr.209601.116

Sousa, C., Golebiewska, A., Poovathingal, S. K., Kaoma, T., Pires-Afonso, Y., Martina, S., … Michelucci, A. (2018). Single-cell transcriptomics reveals distinct inflammation-induced microglia signatures. *EMBO Reports*, *19*(11), e46171. http://doi.org/10.15252/embr.201846171

Stansley, B., Post, J., & Hensley, K. (2012). A comparative review of cell culture systems for the study of microglial biology in Alzheimer's disease. *Journal of Neuroinflammation*, *9*(1), 115. http://doi.org/10.1186/1742-2094-9-115

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, *16*(3), 133–145. http://doi.org/10.1038/nrg3833

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., … Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868. http://doi.org/10.1038/nmeth.4380

Stoeckius, M., Zheng, S., Houck-Loomis, B., Hao, S., Yeung, B. Z., Mauck, W. M., … Satija, R. (2018). Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biology*, *19*(1), 224. http://doi.org/10.1186/s13059-018-1603-1

Streets, A. M., Zhang, X., Cao, C., Pang, Y., Wu, X., Xiong, L., … Huang, Y. (2014). Microfluidic single-cell whole-transcriptome sequencing. *Proceedings of the National Academy of Sciences*, *111*(19), 7048–7053. http://doi.org/10.1073/pnas.1402030111

Stuart, T., & Satija, R. (2019). Integrative single-cell analysis. *Nature Reviews Genetics*, *20*(5), 257–272. http://doi.org/10.1038/s41576-019-0093-7

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., … Yaspo, M.-L. (2008). A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome. *Science*, *321*(5891), 956–960. http://doi.org/10.1126/science.1160342

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., … Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. *Nature Methods*, *14*(4), 381–387. http://doi.org/10.1038/nmeth.4220

Svensson, V., Vento-Tormo, R., & Teichmann, S. A. (2018). Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, *13*(4), 599–604. http://doi.org/10.1038/nprot.2017.149

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., … Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods*, *6*(5), 377–382. http://doi.org/10.1038/nmeth.1315

Tawfik, D. S., & Griffiths, A. D. (1998). Man-made cell-like compartments for molecular evolution. *Nature Biotechnology*, *16*(7), 652–656. http://doi.org/10.1038/nbt0798-652

Theberge, A. B., Courtois, F., Schaerli, Y., Fischlechner, M., Abell, C., Hollfelder, F., & Huck, W. T. S. (2010). Microdroplets in microfluidics: An evolving platform for discoveries in chemistry and biology. *Angewandte Chemie - International Edition*, *49*(34), 5846–5868. http://doi.org/10.1002/anie.200906653

Torre, E., Dueck, H., Shaffer, S., Kim, J., Murray, J., & Correspondence, A. R. (2018). Rare Cell Detection by Single-Cell RNA Sequencing as Guided by Single-Molecule RNA FISH. http://doi.org/10.1016/j.cels.2018.01.014

Tremblay, M.-E., Stevens, B., Sierra, A., Wake, H., Bessis, A., & Nimmerjahn, A.

(2011). The Role of Microglia in the Healthy Brain. *Journal of Neuroscience*, *31*(45), 16064–16069. http://doi.org/10.1523/JNEUROSCI.4158-11.2011

Unger, M. A., Chou, H. P., Thorsen, T., Scherer, A., & Quake, S. R. (2000). Monolithic Microfabricated Valves and Pumps by Multilayer Soft Lithography. *Science*, *288*(5463), 113–116. http://doi.org/10.1126/science.288.5463.113

van Dijk, D., Nainys, J., Sharma, R., Kaithail, P., Carr, A. J., Moon, K. R., … Pe'er, D. (2017). MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *BioRxiv*, 1–61. http://doi.org/10.1101/111591

Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., & Hieter, P. (1997). *Characterization of the Yeast Transcriptome*. *Cell* (Vol. 88). Retrieved from https://api-istex-fr.insb.bib.cnrs.fr/ark:/67375/6H6-0NCWVS4C-V/fulltext.pdf?auth=ip,fede&sid=ebsco,istex-view

VERA, J. C., WHEAT, C. W., FESCEMYER, H. W., FRILANDER, M. J., CRAWFORD, D. L., HANSKI, I., & MARDEN, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, *17*(7), 1636–1647. http://doi.org/10.1111/j.1365-294X.2008.03666.x

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. http://doi.org/10.1038/nrg2484

White, M. C., & Lowen, A. C. (2018). Implications of segment mismatch for influenza A virus evolution. *Journal of General Virology*, *99*(1), 3–16. http://doi.org/10.1099/jgv.0.000989

Whitesides, G. M. (2006). The origins and the future of microfluidics. *Nature*, *442*(7101), 368–73. http://doi.org/10.1038/nature05058

Wu, A. R., Neff, N. F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M. E., … Quake, S. R. (2014). Quantitative assessment of single-cell RNA-sequencing methods. *Nature Methods*, *11*(1), 1–8. http://doi.org/10.1038/nmeth.2694

Wu, A. R., Wang, J., Streets, A. M., & Huang, Y. (2017). Single-Cell Transcriptional Analysis. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)*, *10*(1), 439–

462. http://doi.org/10.1146/annurev-anchem-061516-045228

Zhang, X., Li, T., Liu, F., Chen, Y., Yao, J., Li, Z., … Wang, J. (2019). Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems. *Molecular Cell*, *73*(1), 130–142.e5. http://doi.org/10.1016/j.molcel.2018.10.020

Zhang, Y., Gao, J., Huang, Y., & Wang, J. (2018). Recent Developments in Single-Cell RNA-Seq of Microorganisms. *Biophysical Journal*, *115*(2), 173–180. http://doi.org/10.1016/j.bpj.2018.06.008

Zhao, S., Fung-Leung, W. P., Bittner, A., Ngo, K., & Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*, *9*(1). http://doi.org/10.1371/journal.pone.0078644

Zhao, Y., Wang, K., Wang, W.-L., Yin, T.-T., Dong, W.-Q., & Xu, C.-J. (2019). A high-throughput SNP discovery strategy for RNA-seq data. *BMC Genomics*, *20*(1), 160. http://doi.org/10.1186/s12864-019-5533-4

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., … Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, *8*. http://doi.org/10.1038/ncomms14049

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., Siebert, P. D., Laboratories, C., & Alto, P. (2001). Reverse Transcriptase Template Switching : A SMART ™ Approach for Full-Length cDNA Library Construction, 892–897.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., … Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. *Molecular Cell*, *65*(4), 631–643.e4. http://doi.org/10.1016/j.molcel.2017.01.023

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M., & Mazutis, L. (2017). Single-cell barcoding and sequencing using droplet microfluidics. *Nature Protocols*, *12*(1), 44–73. http://doi.org/10.1038/nprot.2016.154

RÉSUMÉ

Les technologies d'analyse à l'échelle de la cellule unique ont vu le jour il y a quelques années et sont depuis en constante évolution. Ces technologies permettent de mieux comprendre le fonctionnement d'ensemble de cellules très hétérogènes. Elles permettent par exemple de découvrir et suivre des sous types cellulaires, avec des applications en cancérologie ou encore en neurobiologie. Nous avons développé une technologie pour étudier le profil d'expression de gènes d'intérêt au niveau d'une cellule unique, en utilisant la microfluidique en gouttes. En limitant le nombre de gènes étudiés comparé aux technologies commerciales de transcriptome entier, l'approche ciblée a plusieurs avantages potentiels : gagner en profondeur de séquençage, augmenter le nombre de cellules étudiées, optimiser la détection pour les bas niveaux d'expression, tout en réduisant la complexité des données et des coûts. Le ciblage est parfois indispensable, notamment lorsque les ARNs ne portent pas de séquence d'amorce générique, comme dans le cas des ARNs viraux. Deux applications sont présentées : l'analyse de l'inflammation des cellules immunitaires du cerveau aux premières étapes du développement, ainsi que l'étude de la recombinaison génétique chez le virus.

MOTS CLÉS

Technologies cellule unique, séquençage ciblé de l'ARN, microfluidique en gouttes, virus, microglie

ABSTRACT

Single cells technologies were introduced a few years ago and have been dramatically evolving ever since. These technologies have revolutionized biology, making it possible to better understand how heterogeneous cell systems works. For example, they permit to discover and follow cell subtypes, with applications in oncology or neurobiology. We have developed a technology to study the expression profile of genes of interest at the level of a single cell, using droplet-based microfluidics. By limiting the number of genes studied compared to commercial whole-transcriptome technologies, the targeted approach has several potential benefits: gaining deeper sequencing, increasing the number of cells studied, optimizing detection for low levels of expression, while reducing the complexity of data and costs. Targeting is sometimes essential, especially when the RNAs do not carry a generic primer sequence, as in the case of viral RNAs. Two applications are presented: the analysis of inflammation of the immune cells of the brain in the early stages of development, as well as the study of genetic recombination in the virus.

KEYWORDS

Single-cell technologies, Targeted RNAseq, droplet-based microfluidics, microglia, virus