Development of new engineering methodologies for cell sequencing landscape : unbiased mRNA sampling of living cells by TRanscriptomic Analysis Captured in Extracellular vesicles (TRACE)

Francois Cherbonneau

Université de Paris

ED 561 – Hématologie, Oncogenèse, et Biothérapies

Development of new engineering methodologies for cell sequencing landscape

Unbiased mRNA sampling of living cells by TRanscriptomic Analysis Captured in Extracellular vesicles (TRACE).

Par François Cherbonneau

Thèse de doctorat de Biothérapies et Biotechnologies

Dirigée par le Pr Jérôme Larghero
Et le Dr Ibrahim Domian

Présentée et soutenue publiquement le 10/02/2021

Devant un jury composé de :

Dr Mireille Betermier, PhD, Directrice de recherche, Institut de biologie intégrative de la cellule/ Université Paris Saclay, Présidente, Rapporteur
Pr Jérôme Larghero, PharmD, PhD, Professeur d'université, Hôpital Saint Louis/ AP-HP, Paris, Directeur de thèse
Pr Louis Casteilla, PhD, Professeur d'université, établissement Français du Sang (EFS)/ Hôpital Rangeuil Toulouse, Rapporteur
Dr Ibrahim Domian, MD, PhD, Principal Investigator, Massachusetts General Hospital/Harvard Medical School, Boston, Co-Directeur de thèse
Dr Saumya Das, MD, PhD, Principal Investigator, Massachusetts General Hospital/Harvard Medical School, Boston

Titre : Développement de nouvelles méthodes dans le domaine du séquençage cellulaire.

Résumé : L'Hétérogénéité cellulaire et les expressions génétiques fluctuantes dans un microenvironnement spécifique restent mal comprises. Ainsi, afin d'apporter un début de réponse à toutes ces questions, beaucoup de paradigmes scientifiques ont été développés, permettant de toujours repousser plus loin les limites du possible. Ainsi, l'objectif du premier projet de cette thèse fut de développer une méthode innovante pour l'analyse épigénétique multiplexée de cellules à une résolution cellulaire unique. En reliant la transposase Tn5 à des anticorps ciblant des facteurs épigénétiques clés, il pourrait être possible d'identifier le site de liaison de facteurs de transcription spécifiques à l'échelle du génome. Néanmoins, en raison de la relative concurrence dans le développement d'une nouvelle technologie dans ce domaine, cet outil très prometteur fut breveté par une autre entreprise et ce projet de thèse a donc été interrompu au profit d'un autre projet dans cette même thématique. Ainsi, beaucoup de progrès importants en biologie sont fortement corrélés avec de nouvelles méthodologies toujours plus innovantes et qui permettent de définir le destin cellulaire au niveau moléculaire. Mais une grande majorité d'entre elles nécessite l'utilisation de procédures destructrices. Pour ces raisons, nous avons développé une nouvelle technologie permettant une analyse transcriptomique dans le temps sans aucune destruction cellulaire. Nommée TRACE pour l'analyse TRanscriptomique par capture dans des vésicules extracellulaires, elle est caractérisée par l'expression d'un transgène fournissant une translation d'une partie représentative du transcriptome cellulaire à l'intérieur des vésicules extracellulaires. Ainsi, ce «translatome» des cellules qui expriment TRACE peut être suivi dans le temps de manière non destructrice *in vitro* et *in vivo*, ce qui est un outil puissant pour de nombreux domaines de recherches fondamentale et translationnelle.

Mots clefs : Technologie de séquençage, Transcriptome, Vésicules Extracellulaires.

Title: Development of new engineering methodologies for cell sequencing landscape

Abstract: Cell heterogeneity and fluctuant genetic expression in specific microenvironments remain poorly understood. Thus, to provide the beginning of an answer to all of these general questions, a lot of new scientific paradigms were developed and enable to push the limits of the possible. Thus, the goal of the first thesis project was to develop a highly innovative method for multiplexed epigenetic analysis of cells at a single cell resolution. By linking the Tn5 transposase with antibodies targeting key epigenetic factors, it could be possible to identify the binding site of specific transcription factors at a genome wide level. Nevertheless, due to the relative competition to develop a new technology in the field, this very promising tool has been patented by another company, thus the decision was taken to abort this project and focus on another one. A lot of progress and discovery in Biology is strongly correlated with new methodologies that provide the ability to define cell fate at molecular level, but a large majority of them require the use of destructive procedures. For these reasons, the second research project was to develop a new technology allowing transcriptomic analysis over time without any cell destruction. Named TRACE for "TRanscriptomic" Analysis Captured in Extracellular vesicles, it is characterized by a cell-type specific transgene expression providing a translation of a representative part of the cell transcriptome inside Extracellular vesicles. Thus, "Translatome" of cells which express TRACE can be followed over time by non-destructive manner *in vitro* as well as *in vivo*, which is a powerful tool for many fields of fundamental and translational research.

Keywords: Sequencing technology, Transcriptome, Extracellular vesicles.

"*Life is and will ever remain an equation incapable of solution, but it contains certain known factors.*"

— Nikola Tesla

I dedicate this thesis:

To my parents who have instilled in me courage, discipline, stubbornness and persistence. I am proud to be your son.

Without forgetting

My dear twin brother Pierre, for who I always decided to continue my journey against the wind and the tide. I am proud to be your brother.

And

To my future wife, Aurore who, with a lot of patience, tenderness and encouragement helped me to realize my dream to make science. I am proud to be your future husband.

Acknowledgment

"*Be alone, that is the secret of invention; be alone, that is when ideas are born*."

Nikola Tesla

This is what Nikola Tesla liked to say about his relationship with creativity and, to my knowledge, it corresponds in every way to my own perception of it! The quest was hard, but *"the most important is not the destination but the journey travele*d"- Ralph Waldo Emerson.

First, I would like to express a very warm thank you to my two thesis directors, the Pr Jérôme Larghero and the Dr Ibrahim Domian, who, from both side of the Atlantic, trusted me, my capacities and my work. Spending my time in your respective Labs was for me a moment which will be memorable in my scientist carrier. As everybody knows, an important part of our responsibilities as researcher is to share our knowledge, it is thanks to both of you that I truly know, now, the meaning of this sentence. It was arduous to be under your responsibility but was very instructive and I am sure it will bring me a lot for my future professional accomplishment. Again, thank you to the both of you.

Secondly, I would like to present my friendship and acknowledgment to the Dr Saumya Das who accepted me in his Lab after the departure of the Dr Domian. I think I can really consider you as my « third » thesis director. Thank you, a lot, for your commitment in my thesis project. A non-negligible part of my research would have never been done without you. Thank you also for your patience and your help to finish this project.

And last but not least, I wish a very warm thank you to my fiancée Aurore Prunevieille, who was also a PhD Candidate in the MGH/ Harvard Medical School. It is due to your support and help during all this research Marathon that my thesis is finally done. Thank you, a lot, dear, for your professionalism and attendance. Without forgetting the American co-thesis director of Aurore, the Dr Gilles Benichou, who always had excellent advices. Thank you, Gilles, for your time and help.

Disclaimers

This document is the result of a long work process made in both sides of the Atlantic with a Co-thesis agreement between the Saint Louis hospital Paris/University of Paris and the Massachusetts General Hospital/Harvard Medical School.

The two organizations equally participated to the production of this work and none of them can be removed or substituted.

Summary

ACKNOWLEDGMENT **6**

DISCLAIMERS **8**

SUMMARY **9**

LIST OF THE PRINCIPAL ABBREVIATIONS **18**

LIST OF THE PUBLICATIONS AND COMMUNICATIONS **19**

GENERAL INTRODUCTION **20**

GENERAL BIBLIOGRAPHY AND CONTEXT **22**

I. THE HISTORY OF THE HEART REGENERATION **22**

1. THE POOR REGENERATION CAPABILITY OF CARDIOMYOCYTES 23

A. The proliferation capability in adult 23

2. IMPROVEMENT OF THE HEART REGENERATION 24

A. Direct adult cardiomyocytes regeneration by restoring the CM progenitors' pool 26

B. Another adult stem cell as a source of CMs 26

C. Human pluripotent stem cells in cardiac regeneration 27

D. Other possible therapeutic studies to compensate the loss of CMs 29

II. THE MYOCARDIAL DIFFERENTIATION PATHWAY IN HUMAN PLURIPOTENT STEM CELLS **30**

1. OVERVIEW OF THE DIFFERENTIATION PATHWAY FOR HPSCS-CMS 30

A. Epithelia-Mesenchymal transition 32
B. Mesoderm progenitor 32
C. Precardiac mesoderm 33
D. Cardiac mesoderm 33
E. Heart specific progenitors 34
2. STUDY FOCUS IN THE DOMIAN LAB AND LARGHERO LAB 35
3. THE PROTOCOL OF DIFFERENTIATION AND SIDE PROJECTS IN THE DOMIAN LABORATORY 36
A. The Domian's Laboratory protocol 36
B. A possible technology for Multiplex live single-cell transcriptional analysis 37
C. The urgent need of a personalized sequencing technology 38
III. SEQUENCING TECHNIQUES IN THE HEART REGENERATION AND CONTEXT 39
1. HISTORY OF SEQUENCING 39
A. Sanger and the revolution of the PCR 40
B. The human genome project and NGS 42
2. MOST ACTUAL ADVANCED SEQUENCING TECHNIQUES 44
A. Single cell sequencing 44
B. Epigenetic and DNA methylation analysis 46
a. Define epigenome and DNA methylation 46
i. The impact of the epigenome 46
ii. Histone and DNA methylation 48
Histone methylation 48
DNA methylation 50

b. Sequencing methods in epigenome and histone/DNA methylation context 50

c. Other sources of Methylation, the RNA regulation 52

i. RNA methylation 52

ii. Sequencing techniques related to RNA methylation 54

C. Revolution of CRISPR-Cas9 and the sequencing 54

3. SEQUENCING TECHNIQUE IN THE HEART REGENERATION 55

4. ACTUAL REMAINING GAP AND THESIS OBJECTIVES 56

FIRST PART: A NOVEL METHOD FOR HIGHLY MULTIPLEXED EPIGENETIC ANALYSIS AT SINGLE CELL RESOLUTION 60

I. INTRODUCTION 60

II. BIBLIOGRAPHY CONTEXT 63

1. THE TRANSPOSON PROTEIN FAMILY AND THEIR ROLE IN SEQUENCING 63

A. The transposon superfamily 63

B. Transposons and Tn5 protein 64

a. The Transposon class of proteins 64

b. The Wild-type Tn5 65

C. Applicability of the transposon in biology 66

a. EZ: Tn5® and cloning strategy improvement 66

2. THE STREPTAVIDIN CLASS OF PROTEIN AND THE MONOMERIC STREPTAVIDIN MSA 69

A. The streptavidin family 69

B. The monomeric streptavidin mSA 70

III. MATERIAL AND METHOD 71

1. CLONING STRATEGY 71

2. TN5 PRODUCTION 71

A. Protein Expression 71

B. Protein Purification 72

3. TN5 VALIDATION 74

A. Annealing Oligos 74

B. Tn5 oligo preparation 74

C. Tagmentation reaction 75

D. PCR after tagmentation 75

E. Tn5 Streptavidin complex formation 77

IV. RESULTS AND DISCUSSION 78

1. EXPERIMENTAL RESULTS 78

A. Design and theorical cloning strategy 78

a. In silico studies for the two-fusion proteins 78

b. Tn5-mSA and mSA-Tn5 prediction 79

i. In silico Tn5-mSA 79

ii. In silico mSA-Tn5 80

c. Theorical cloning design for the two-fusion protein plasmids: Tn5-mSA and mSA-Tn5 and the control 81

B. Production and purification of the two isoforms Tn5/mSA fusions proteins complex 81

a. Theorical design protocol and justification of the production/purification strategy 82

i. Quick resume of the different classes of protein purification 82
ii. Justification of the Neb IMPACT® strategy of purification 83

b. Optimization and validation of the protocol of purification and conservation. 85

i. Introduction and conceptual protocol 85
ii. Results and optimization 85

C. Validation of the Tn5/mSA fusion protein 90

a. Verification of Tn5 tagmentation by DNA and plasmids fragmentation 90

i. Tn5 fusion and focus on the tagmentation reaction 91
Synaptic complex formation 91
Tagmentation reaction 92

b. Verification of mSA/biotin binding with EMSA Tn5-mSA/Biotin-Dna probe 94

i. mSA binding reaction 94
ii. Validation of the two-fusion protein focus on the mSA function 94
Introduction and conceptual propose 94
Results and optimization 95

c. Direct binding by chemical reaction on the Tn5 protein itself 96

i. Maleimide complexification test 96
The maleimide reaction 97
Results and optimization 97
ii. The Mal-PEG, Cys-PEG and NHS-PEG complexification 100

D. New design and validation strategy based on oligo customization 102

a. Oligo customization design and strategy 102

i. Return to the Tn5 production and more careful validation 103
Tagmentation test on regular oligo design 103
After the Tagmentation, the PCR amplification step 105
ii. Verification of tagmentation on different designs 106
Introduction and conceptual propose 106
Results and optimization 107
b. Definitive design and validation process of the new Oligo-Tn5 custom 111
i. The final design chosen 111
ii. Verification of Streptavidin binding on Chic-loop biotin oligo 113
iii. Verification of tagmentation whole complex 114
2. DISCUSSION 116
V. CONCLUSION 118
SECOND PART: ENGINEERED EXTRACELLULAR VESICLES FOR OVER TIME UNBIASED MRNA SAMPLING OF LIVING CELLS 120
I. INTRODUCTION 120
II. BIBLIOGRAPHY CONTEXT 123
1. EVS LANDSCAPE, A BIG FAMILY. 123
2. RNA CONTENT IN VESICLES, A WORLD WHERE EVERYONE DOES NOT AGREE 125
3. BINDING PROTEIN SYSTEMS: A LINK BETWEEN MRNA CATCHER AND THE EV IMPORT PROTEIN 126
4. RNA "CATCHER" IN A JUNGLE OF POTENTIAL CANDIDATES 129
A. The whole RBP landscape 129
B. YTHDF candidates and the m^6A mRNA methylation modification 130
5. CURRENT SEQUENCING LIBRARY GENERATION PROTOCOL COMPATIBLE WITH AN EVS CONTEXT 133
A. Purification of nucleic acid in EVs 133
B. Sequencing library compatible with EVs low content 133

III. MATERIAL AND METHOD 136

1. CELLS AND CLONING PROTOCOL 136

A. Cells 136

B. Plasmid and cloning strategy 136

C. Transfection, Transduction and virus Production 138

D. Stable cell lines generation 140

2. EXTRACELLULAR VESICLES AND RNA PURIFICATION 141

A. Extracellular vesicles purification 141

B. Cell purification 142

C. RNA purification 142

3. RNA MODIFICATION AND ANALYSIS 143

A. Reverse transcription and PCR preamplification 143

B. Tagmentation reaction and amplification of adapter-ligated fragments 145

C. cDNA sequencing 147

IV. RESULTS AND DISCUSSION 150

1. EXPERIMENTAL RESULTS 150

A. Design and theorical cloning strategy 150

a. Process of production and validation of the basic roles to validate a final design 150

i. A complex cloning strategy 150

ii. Validation of the basic role of both parts of the constructs 151

B. Basic proof of principle 154

a. Final design and cloning strategy....154

i. Cloning overview....154

ii. Basic validation....156

C. Validation and robustness of the technique....160

a. TRACE-seq, an mRNA translation methodology carried by Extracellular Vesicles....160

b. Detection of the mRNA inside the Extracellular vesicles and reverse transcription....162

c. The transcriptome in TRACE-Seq MVs is representative of the cellular transcriptome....168

d. Following specific gene expression from H_2O_2 stressed cells over time with TRACE....172

2. DISCUSSION....175

V. CONCLUSION....176

GENERAL CONCLUSION....178

PERSPECTIVES....181

ANNEXES....183

I. ANNEXES TN5 PROJECT....183

II. ANNEXES TRACE PROJECT....187

III. ANNEXES MANUSCRIPTS....203

1. THE MAGIC MANUSCRIPT....203

A. Abstract....203

B. Introduction....204

2. THE TRACE-SEQ MANUSCRIPT....211

A. Abstract....211

B. Introduction....211

BIBLIOGRAPHY....213

ILLUSTRATIONS TABLE....226

FIGURES....226

TABLES....232

RESUME EN FRANÇAIS....233

List of the principal abbreviations

ATAC-seq	Transposase-accessible chromatin using sequencing
BM	Bone Marrow
BMMNC	Bone Marrow Mononuclear Cells
BMP	Bone Morphogenetic Proteins
BSA	Bovine Serum Albumin
CAGE	Cap Analysis of Gene Expression
CDCs	Cardiac progenitor Derived Cells
Chic-seq	Chromatin Histone Immune Coding coupled to whole genome sequencing
Chip-Seq	Chromatin Immunoprecipitation Sequencing
CM	Cardiomyocytes
DNA	Desoxyribonucleic Acid
EMSA	Electro mobile shift assay
ESC	Embryonic Stem cells
Evs	Extracellular Vesicles
G&T-seq	Genome and Transcriptome Sequencing
GBP	GFP Binding Protein
GO	Gene Ontology
hPSCs	Human Pluripotent Stem Cells
IP	Immunoprecipitation
IPS	Induce Pluripotent Stem cells
MI	Myocardial Infarction
mRNA	Messenger Ribonucleic Acid
mSA	monomeric Streptavidin
MSCs	Mesenchimal Stem Cells
MVs	Microvesicles
NGS	Next Generation Sequencing
RBP	RNA Binding Protein
RNA	Ribonucleic Acid
scBS-seq	Single Cell genome-wide BiSulfite Sequencing
scM&T-seq	Single Cell genome-wide Methylome and Transcriptome Sequencing
TFs	Transcription Factors

List of the publications and communications

Papers:

Papers in reviews, iScience : TRACE-seq: A transgenic system for unbiased loading of mRNA into extracellular vesicles allows non-invasive transcriptome profiling of living cells.

François Cherbonneau, Aurore Prunevieille, Robert Kitchen, Gilles Benichou, Jerome Larghero, Saumya Das*, Ibrahim Domian*

Multiplex Live Single-Cell Transcriptional Analysis Demarcates Cellular Functional Heterogeneity

doi: 10.7554/eLife.49599

Ayhan Atmanli; Dongjian Hu; Frederik Ernst Deiman; Annebel Marjolein van de Vrugt; **Francois Cherbonneau;** Lauren Black; Ibrahim Domian

Prices:

Reached the first place in the annual AHA conference and the Vascular research initiative Conference: Autophagy is impaired in thoracic aortic aneurysm

Ahajournals.org

Elizabeth L Chou, MGH Vascular Surgery, CVRC, Boston, MA; **Francois Cherbonneau**, MGH CVRC, Boston, MA; Mark F Conrad, MGH Vascular Surgery, Boston, MA; Christian L Lino Cardenas, Mark E Lindsay, MGH CVRC, Boston, MA

General introduction

Cell therapy and tissue regeneration represents a major therapeutic challenge for modern medicine. Thus, millions of patients are pinning their hopes on the use of stem/progenitor cells, especially in the management of a major public health problem such as heart failure. The first bone marrow transplantation was performed by the team of Professor George Mathé [1] (in Paul Brousse hospital, Villejuif Paris in 1958). From this discovery, the cellular biology had an important contribution in the future regenerative medicine. At the beginning of the century, a Japanese team headed by Shinya Yamanaka managed to reprogram somatic cells into pluripotent stem cells, named Induced Pluripotent Stem cells (IPS) [2, 3]. Cell therapy made big progresses in a short period of time and this growth of knowledge raised a lot of interest because it could be the key in solving the incomprehension around the cellular regeneration as well as a big hope for patients.

Nowadays, thanks to the advances in stem cell research, scientific techniques which appeared as Science Fiction in the past, gradually began to take their place in the global medical landscape. The two major properties that define a stem cell signature are self-renewal and multi-potential differentiation abilities. These cells can be isolated from the patient himself (to avoid immune response) and opened the scientist community to various possibilities like neural graft or organ regeneration.

Nevertheless, all these clinical applications will be possible right after fundamental research projects made to better understand the mechanisms and their consequences. Achieving a better understanding of the differentiation and proliferation of cardiac stem/progenitor cell is a challenge for the future research.

"*By all means keep your enthusiasm, but let verification be its constant companion*."

Louis Pasteur

Thus, the understanding of genome-wide epigenomic in cell population is a key challenge for the future in particular in stem cells population [4]. Also, sequencing technologies have become a powerful investigation tool in the last decade. Some doors have been opened with the apparition of epigenetic/transcriptomic analysis by sequencing in the field of heart regeneration [5]. Two major goals remain poorly understood, and the sequencing technology could bring responses: first, to better understand the capacities of stem/progenitor cell to increase the cardiomyocytes proliferation; the second one, better targeting the remodeling and dedifferentiation of cardiomyocyte during the regeneration process.

In this context, my research project at the Cardiovascular Research center of the Massachusetts General Hospital/ Saint Louis Hospital Paris is to create a new methodology of sequencing to determine the key epigenetic/transcriptomic factors during the commitment of stem cell in the myocardial pathway.

General bibliography and context

I. The history of the heart regeneration

Over the last century, many scientists thought that the myocardium was deprived of regenerative capacities. This paradigm has been challenged by a lot of studies in rodent and human demonstrating a low potential of regeneration insured by the heart muscle with or without stress conditions [6-8]. Thus, the heart is one of the least regenerative organs in the body and represents an issue for patients developing chronic heart failure, being associated with 50% survival rate over 5 years [9]. One of the major causes of heart failure is myocardial infarction (MI). Indeed, 1 billion cardiomyocytes (CMs) are lost during MI [10]. The damage caused by MI is hopelessly irreversible considering that the differentiation of progenitors into CMs and the proliferative potential of adult CMs is very limited. The only treatments usually available are limited to transplant a new organ: from human allo-graft [11] or artificial heart [12].

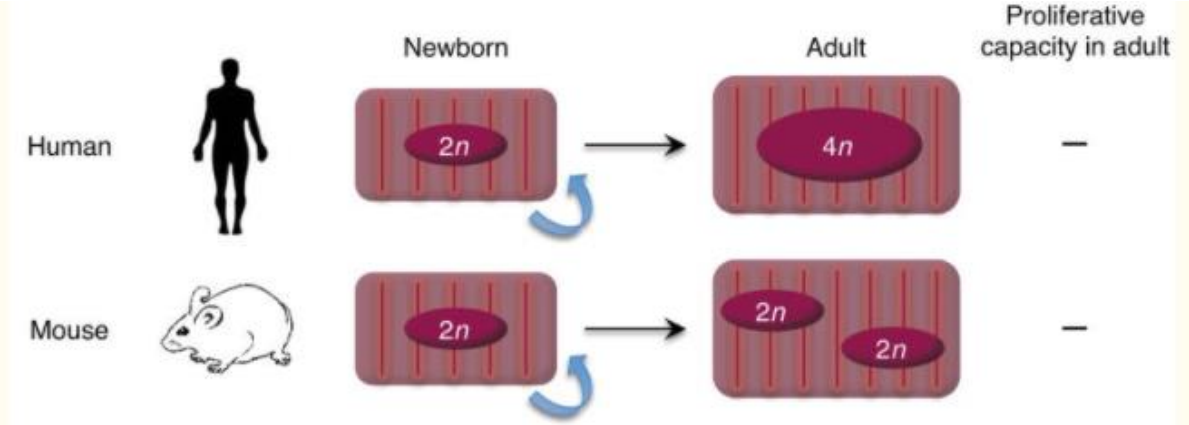
Thanks to the progress of Biomedicine, new treatments are in development. Although lesions in the cardiac tissue seem to persist indefinitely, regeneration techniques have been used for almost 10 years involving cell therapy in order to restock new cardiomyocytes in the infarct myocardial area [13].

1. The poor regeneration capability of cardiomyocytes

A. The proliferation capability in adult

The heart muscle has a tremendous role in maintaining the blood flow and the circulation through the entire body. Its task is critical for a body in perfect health. Thus, we could imagine that the proliferative profile of adult CMs is important but in fact it is not true (Figure1).

Figure 1. Nuclear dynamics and proliferative capacity of cardiomyocytes during growth.



Kazu Kikuchi and Kenneth D. Poss, Annu Rev Cell Dev Bioll, 2013 [14]

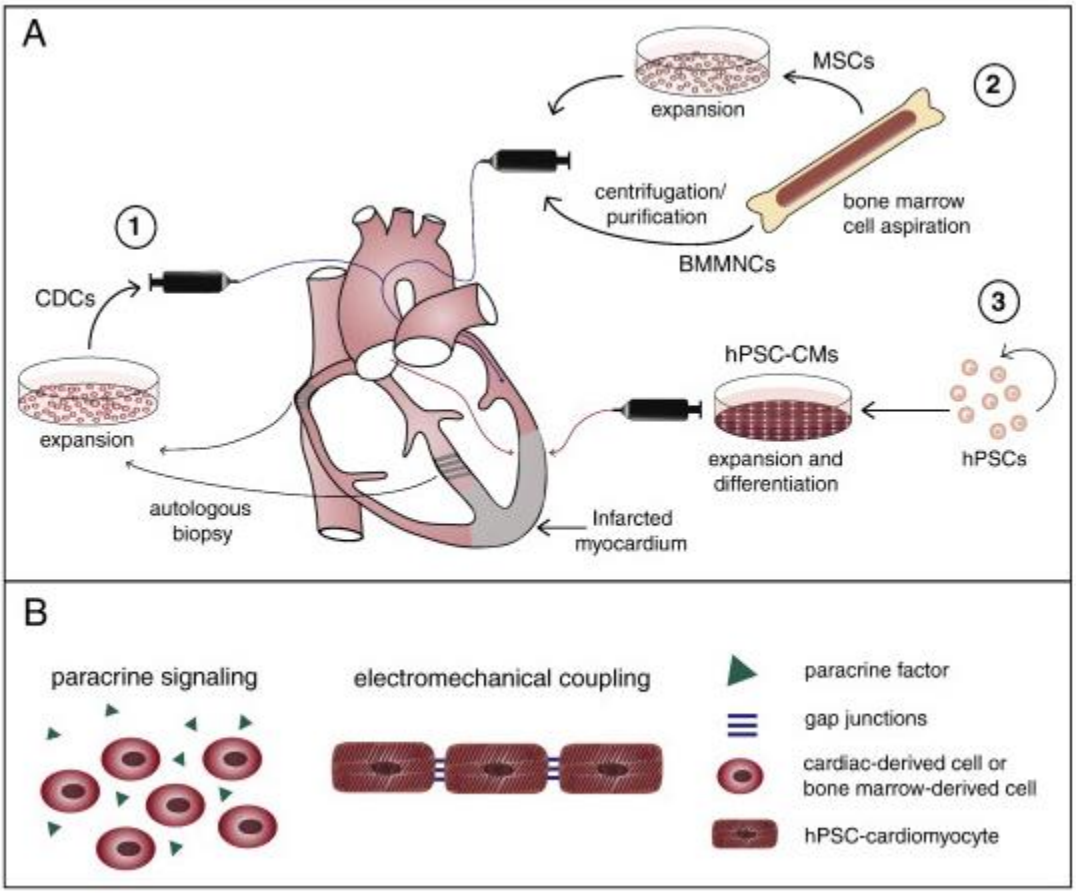
Many researchers disagree to say that the poor capability of regeneration/proliferation of CMs in mammals is due to their inability to enter the regular cell cycle division [14].

As shown in Figure 1, CMs from newborns possess a regeneration capability due to their mononucleated phenotype (DNA replication ability). This cell conformation is then lost in adults and become binucleated with diploid nuclei in mouse and mononucleated with polyploid nuclei in human [10]. This configuration does not allow a possible cell/cell division in adult mammals and blocks CMs into a non-proliferative state.

2. Improvement of the heart regeneration

The heart is one of the least adaptative tissue to brutal cell loss and least regenerative organs in the body, which represents an issue for patients developing chronic heart failure. It is known that CMs loss after MI is directly corelated with a very severe fibrosis which affects the beating and flexibility capabilities of the heart. Usually, the unique treatment is to give the patient a new functional heart with an allogenic graft. But performing a highly invasive surgery presents a non-negligible risk for patients and is not compatible with all clinical cases. In this context, developing a new therapeutic by boosting the proliferation capability of CMs or infusing new *in vitro* generated CMs from different sources (Figure2) become a very enticing research area.

Figure 2. Cell transplantation techniques and proposed mechanisms of cell therapy for heart regeneration.



Kaytlyn A.Gerbin,aCharles E.Murry, Cardiovascular Pathology,2015 [15]

a *Cell transplantation after MI. 1 Cardiac-derived cells (CDCs) isolated from the atrium or the interventricular septum. 2 Bone marrow mononuclear cells (BMMNCs) and mesenchymal stem cells (MSCs) harvested from the bone marrow. 3 Human cardiomyocytes derived from human pluripotent stem cells (hPSCs). All cells are infused by pre-aortic injection except for the hPSCs which are directly transfused to the scar.* ***b*** *Possible mechanism of action occurring after the cell transplantation. All cell types secrete (at different level) paracrine factors to the surrounding scar, but only the hPSCs are capable to direct electromechanical integration.*

A. Direct adult cardiomyocytes regeneration by restoring the CM progenitors' pool

Even though heart tissue was relegated by the scientist community as one of the least regenerative organs, some described a possible source of CMs by reinfusing Cardiac-derived cells: CDCs (Figure 2 A 1). These cardiac stem cells need to be derived from a myocardial biopsy, grown in culture *ex-vivo* and reinfused by aortic injection. Studies suggest that progenitor cells are capable of differentiating into CMs and be more effective than BMMNCs or MSCs [16]. These CDCs regroup a vast batch of studies with variable Cardiac progenitor Derived Cells from different sources as c-kit progenitors (SCIPIO) [17], Cardiosphere-derived cells (CADUCEUS) [18] and Autologous Human cardiac-derived Stem Cell (ALCADIA) [19].

However, a majority of groups agreed to assume that cells work preferentially through paracrine effects (Figure 2 B) and do not bring a lot of CM renewal which is illustrated by a minimal long-term engraftment and a poor CM differentiation [15]. Thus, direct regeneration of the adult CMs from cardiac progenitors in Mammals seems to still suffer from insurmountable challenges.

B. Another adult stem cell as a source of CMs

Hopefully for patients, the CMs differentiation from cardiac stem cells are not the unique source of potential CMs renewal. Many groups of researchers explored different sources of stem cells as CMs "producers". For this aim, different populations of cells other than cardiac derived cells have been used in clinical trials or are under study in research

labs: Bone marrow derived cells such as Bone marrow mononuclear cells (BMMNCs) [20] and mesenchymal stem cells (MSCs) [21]. But these two developmental therapeutics suffer from their own problematic. From the side of the BMMNCs, the true amount of stem cells capable of cardiac differentiation is very low (less than 0.1% of the entire "BMMNCs" isolated population [22]) which remains the main issue for a clinical scale up in a therapeutic point of view. On the other side, the MSCs also appear to suffer from the same issue as cells capable of differentiation represent less than 0.01% of the isolated cells from bone marrow [21, 23]. Moreover, lot of studies in animals suggest that the Bone marrow derived cells only induce a short-term paracrine effect on the remaining CMs (short term "booster" of the myocardial tissue which compensates -in a functional aspect- the loss of CMs) after an MI [24, 25]. Like the CDCs, these cells do not bring a long term CMs renewal and functional engraftment of new cells.

Although a large emerging panel of cell types are proposed to restore the CMs pool after MI, they all suffer from a poor yield of functional CMs.

C. Human pluripotent stem cells in cardiac regeneration

Pluripotent stem cells like embryonic stem cells (ESCs) [26] and induced pluripotent stem cells (IPS) [27] are currently under study in several labs and are under clinical trials (Figure 2 A 3). Despite a developmental issue especially in culture conditions and control of differentiation into viable CMs, it exists now several efficient protocols [28]. Thus, hPSCs could virtually be differentiated and expended into CMs without any limit which bring hope for the future of cell therapeutics despite the time-consuming aspect. Studies have shown a

functional (although light) improvement of the left ventricular parameters (LVEF) [29, 30], which may justify the benefit of cell therapy towards long term engraftment [31]. Moreover, reported hPSCs-CMs electrical integration improvement permitted to reduce the incidence of arrhythmias in rodents' ischemia/reperfusion injured hearts and cryoinjured guinea pig hearts [30]. In fact, studies showed that with the concourse of GAP junction hPSCs-CMs, cells are electrically integrated (Figure 2, B) and participate to the beating force generation [32, 33]. In the opposite, the same group reported some difficulties of acceptability of transplanted CMs derived hPSCs in terms of integration and electrical modulation in chronic MI animal (guinea pig) models which result in a non-improvement of the myocardial function [34]. In fact, it exists many contradictory researches on the subject (especially for the electrical integration in their microenvironmental context) and to date it is complicated to judge the benefit post MI of the infusion of the CMs derived from hPSCs. Moreover, it is also very problematic to differentiate specific cell types from hPSCs which reflect the cardiac leaflets phenotypical (atrial, ventricular, and pacemaker cells) and functional heterogeneity. Knowing if the benefit brought by the hPSCs-CMs is mainly due to a direct cell differentiation or a paracrine effect is still on debate [35].

Despite challenges, the hPSCs-CMs infused cells remain the most promising therapies to avoid heart transplantation and improve CM regeneration, but still needs to be improved for long-term benefits. Moreover, the cells heterogeneity in the heart must be taken into account. The heart tissue should be apprehended as a global landscape (cardia syncytium) but composed of different cell populations with distinct power of contractibility and electrically integration/transfer.

Thus, stem cell therapeutic needs to have a specific protocol of differentiation to better consider the critical key points and adapt the neo hPSCs-CMs produced to their future goal (microenvironmental context and functional situation). Moreover, developing new technologies especially in sequencing to better characterize cell by cell each heart area is a major goal for the current research.

D. Other possible therapeutic studies to compensate the loss of CMs

Recent studies in zebrafish and newts suggest that the dedifferentiation of cardiomyocytes is possible [36, 37]. We know very little thing about dedifferentiation mechanisms of mature cardiomyocytes in mammals, many pathways are proposed, especially the role of the oncostatin M to increase the dedifferentiation of cardiomyocytes and have a protective effect against MI [38-40]. This mechanism opens doors to new methodologies to dedifferentiate CMs, proliferate and redifferentiate into CMs. Even if this new methodology suffers from a default of new CMs obtention after dedifferentiation and redifferentiation (7% maximum) [41], the protocol needs to be improved and there is no doubt that the technic will become very interesting to replace CMs after an MI.

II. The myocardial differentiation pathway in human pluripotent stem cells

As previously mentioned, the differentiation of hPSCs into CMs is a critical point to get enough material and infuse new hPSCs derived CMs to animals to improve heart regeneration after MI. Nevertheless, it exists a lot of protocols for the differentiation of the hPSCs into CMs. We will first focus on the theorical steps for a successful differentiation and secondly, we will present the protocol used in my unit at the Massachusetts General Hospital.

1. Overview of the differentiation pathway for hPSCs-CMs

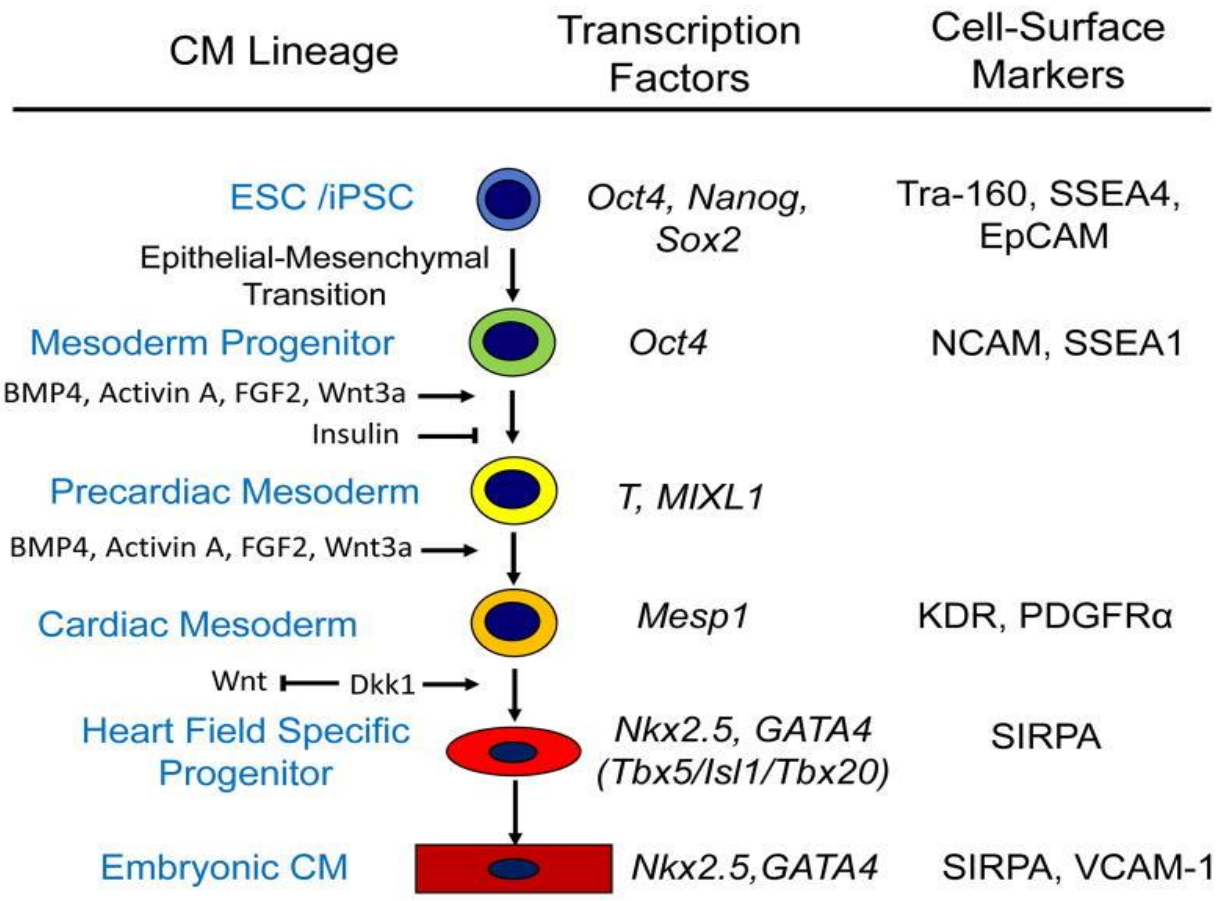
The first successful differentiation into viable beating CMs was performed with isolated hECSs, then recently followed by induced pluripotent stem cells (hiPSCs).

Moreover, the ESCs-CMs were reported to be used in the first clinical trial made by my research unit in Saint Louis hospital Paris [26]. Both cell populations gave a large opportunity as a future therapeutic [42, 43].

The theorical differentiation and the actual protocols to generate CMs from hPSCs represent several decades of studies on embryonic cardiac development [44, 45].

The heart embryonic tissue is one of the earliest tissues formed soon after the gastrulation. It is formed by migrating mesodermal cells and sandwiched between the ectoderm and endoderm.

Figure 3. Model of Differentiation of Human PSC via Sequential Progenitors to Cardiomyocytes regeneration.



C. Mummery, et al. Circ Res,2012 [46]

A. Epithelia-Mesenchymal transition

Before any differentiation, stem cells are engaged. Embryonic stem cells are characterized by a cocktail of transcription factors such as Nanog, Sox2 and Oct4 which correspond to self-renewal regulatory factors (Figure 3). In fact, if these loops of TFs are broken, cells will lose their self-renewing totipotency and commitment into the mesodermal way of differentiation.

B. Mesoderm progenitor

To engage stem cells into the precardiac mesoderm and make them quit the mesodermal layer, three principal families of growth factors are critical: the bone morphogenetic proteins (BMPs) essentially BMP4, the members of the growth factor β superfamily, the fibroblast growth factors (FGFs) and the Wingless/INT proteins (WNTs) family, principally the Wnt 3a. If their factors are disturbed during the cardiac differentiation pathway initiation, it could have a dramatic effect on the heart tissue development [45].

All these factors are expressed into the mesoderm and their up expression/inhibition promote the cardiac differentiation. To give more details, when the Wnt pathway is activated at a late stage, it inhibits the cardiogenesis in synergy with the Wnt5A and BMP4 downregulation [47], and the strict opposite is necessary to promote a cardiogenesis activation. Moreover, the inhibition of the Wnt3A and Wnt8 (canonical Wnts) but not Wnt5A nor Wnt11 (noncanonical Wnts) allow a move to the precardiac mesoderm formation from the mesoderm dorsal [48]. Indeed, the inhibition of the Wnt/β-catenin signaling promotes the cardiac differentiation commitment [49] (Figure 3). To resume, in human ESCs, the

upregulation of BMP4 and the inhibition of Wnt3a (canonical Wnts) are generally a key process to engage mesodermal stem cells into the precardiac mesoderm.

C. Precardiac mesoderm

The precardiac mesoderm phase is very similar to the previous one with the same induction BMP4 and Activin A and inhibition of the Wnt canonical TFs with Wnt3a. Also, FGF promotes the mesoderm differentiation and cardiogenesis at a later stage and not an early stage [50]. Thus, an activation of the FGF2 in synergy with the previous induction of BMP4 and Activin A engages permanently cells into the cardiac differentiation layer. The control of the insulin level inside the cell culture media also plays a critical role during the precardiac shift. Indeed, using an insulin free media during the first differentiation step and then switch to one with insulin helps to move from the mesodermal layer to a precardiac mesoderm. During this critical phase, insulin inhibits very strongly the cardiac differentiation and needs to be removed [51].

D. Cardiac mesoderm

When cells are completely engaged into the cardiac mesoderm, they are fully considered as precardiac tissue with no possible comeback. At the first stage of differentiation, the mesoderm expresses TFs like the T-box factor Brachyury (T) and the homeodomain protein, (Mixl1), Figure 3. After receiving the cocktail of growth factors of differentiation BMP4, Activin A [52], FGF2 and inhibition of Wnt3a, both TFs are inhibited and engage the mesoderm into the much more mature precardiac tissue. Moreover, when cells are guided into the cardiac lineage, they express the basic helix-loop-helix transcription factor

mesoderm posterior 1 (Mesp1) [53]. Thus, cells are ready for a final cardiac differentiation stage with the Nkx 2.5 and GATA4 induction pathway.

E. Heart specific progenitors

When cells start to express the Mesp1, they also express in low quantity the transcription factor Nkx2.5. The TF Nkx2.5 is an essential factor for the formation of a functional heart tissue with a contractile capability. Indeed, during the primordial heart formation, the TF Nkx2.5 works in synergy with the Tbx5 and both play a critical role in the formation of the atrial and left ventricular compartment [46]. The zinc finger transcription factor (GATA4) is also a critical factor at this step because it activates the cardiac genes like the couple actin/myosin chain and the troponin. Moreover, GATA4 plays a critical role during the heart tube 3D formation and the proepicardium formation [54, 55]. At the end of this stage, the primordial heart is getting his final shape and receives his electrical and beating function.

2. Study Focus in the Domian Lab and Larghero Lab

This is in this context of myocardial regeneration using hPSCs that my work takes place. First, in the Largero laboratory, a lot of efforts are made to design a complete protocol of ESCs differentiation into CMs in order to translate it to clinical studies. The Larghero's team was the first to graft ESC-derived CMs (ESC-CMs) to patients by a myocardial patch put into the infracted area [26]. Nevertheless, a constant work is necessary to improve the beating function and electro-transfer of the differentiated CMs.

To that extent, many efforts were made *in vitro*, which is the main study focus of the Domian Lab. Thus, they try to improve the ESCs differentiation into CMs (point developed after). Many studies were made in the lab to determine a better condition to improve the myocardial differentiation such as the maturation of ESCs into CMs by the inhibition of HIF1α and LDHA [56] or with specific 3D aggregate culture [57]. Also, constant efforts were put on the characterization of the early steps of the ESCs myocardial commitment. The MAGIC methodology development in which I participated, was made "*to determine how individual cells with varied gene expression profiles and diverse functional characteristics contribute to development, physiology, and disease*" Atmanli *et al* elife 2019 [58]. Many efforts were put to improve the differentiation protocol to get more hPSCs-CMs with the best beating/electro-transfer capabilities. After one decade of work, we do have a very robust differentiation protocol. Nevertheless, continue to adapt it and ameliorate it is a permanent goal for the lab.

3. The protocol of differentiation and side projects in the Domian laboratory

A. The Domian's Laboratory protocol

The generation of hPSCs-CMs cells is a very promising source of viable CMs with beating function and a long-term survival of grafts. Nevertheless, it exists a lot of different protocols, with various rates of success, even if all of them respect the major steps we described above. Thus, I will refer essentially to the protocol we used in my research unit at MGH. The following protocol from *Atmanli et al., elife 2019* is the most recent protocol used in my unit:

"Cardiac differentiation of hPSCs was induced using small molecules as previously described [59]. Briefly, when hPSCs achieved confluency, cells were treated with CHIR99021 (Stemgent) in RPMI (Thermo Fisher) supplemented with Gem21 NeuroPlex without insulin (Gemini Bio Products) for 24 hr (from day 0 to day 1). The medium was replaced with RPMI/G21-insulin at day 1. The cells were then treated with IWP4 (Stemgent) in RPMI/G21-insulin at day 3 and the medium was refreshed on day 5 with RPMI/G21-insulin. Cells were maintained in RPMI supplemented with Gem21 NeuroPlex (Gemini Bio Products) starting from day 7, with the medium changed every 3 days. Ascorbic acid at 50 µg/ml was added to media until onset of beating. Beating clusters were seen starting day 6 of differentiation. Metabolic selection of cardiac myocytes was performed for 3 days by incubating cells in media without glucose but supplemented with 5 mM sodium DL-lactate. Cardiac myocytes were harvested by treating the cells with Collagenase A and B (Roche) for 5 min first and then TrypLe Express (Thermo Fisher) for another 5 min. Cells were plated onto Matrigel-

coated 96-well plates with a No.1.5 glass bottom or polydimethylsiloxane-coated glass dishes" *Atmanli et al. eLife,2019*

After decades of research, this protocol is the most efficient so far in the Domian lab and we use it routinely. As seen above, we use the CHIR99021 and the IWP4 as inhibitors of the Wnt pathway. Cell culture media does not contain insulin for the early stage of differentiation and insulin is added for the next step after the critical phase. To help for the shift to the precardiac mesodermal commitment, the media is supplemented with Gem21 NeuroPlex which is known to help for the neural differentiation and can be used for the cardiac differentiation as well. Finally, Ascorbic acid is added at day 7 of differentiation to maintain cells in a cardiac commitment state.

B. A possible technology for Multiplex live single-cell transcriptional analysis

To always improve and increase the yield of hPSCs-CMs, a particular attention was accorded to develop a better technology of cell maturation and characterization. Indeed, during my work in the Domian laboratory, I participated to the development of a new technique: **Multiplex live single-cell transcriptional analysis demarcates cellular functional heterogeneity.** The Main goal of this project was to better characterize the gene expression and cell physiology at a single cell level by "*utilizing fluorescently labeled mRNA-specific anti-sense RNA probes and dsRNA-binding protein to identify the expression of specific genes in real-time at single-cell resolution via FRET" Atmanli et al. elife 2019* [58]. Very concomitant to the actual smFISH technique, this MAGIC (*M*ultiplex *A*nalysis of Gene Expression in *I*ndividual living *C*ells) technology was able to visualize the cytoplasmic nature of the β-actin mRNA in

U2OS cells. Moreover, at the end of this work, we were able to characterize the cell physiology during early commitment from the myocardial lineage (differentiation from hPSCs). Thus, this new technology could be used to delineate the process of functional maturation in human cardiac myocytes. It could open doors to visualize the physiology heterogeneity from single cells with a very simple methodology using just specific fluorescents probes to target critical genes. The abstract and the introduction of this work can be found in the annexes. Thereby, I had the opportunity to participate to this work which gave me a better understanding of the strategic manner of new methods creation and critical steps required for the development of a new technology in the biotechnology field. It also allowed me to envision and select different techniques used to distinguish cell fate markers corroborating the differentiation states.

C. The urgent need of a personalized sequencing technology

The process of stem cell differentiation in a myocardial context is a very complicated operation with a lot a remaining unknowns and gaps. Better understand the overall and detailed part of this differentiation continuum will allow an improvement of the hPSCs-CMs yields and beating/electro-transfer capabilities. To do so, traditional characterization techniques (e.g. staining, Microscopy, gene quantification...) do not offer a complete visualization of the global/detailed process of cell maturation, which regroups a large amount of TFs at the same time. Having the opportunity to identify the global pool of TFs which play a critical role during the differentiation would be very powerful. Thus, with the growing “boom” of the sequencing technology this last decade, developing and adapting new sequencing methodologies to better characterize the stem cell differentiation in a cardiac context is important.

III. Sequencing techniques in the heart regeneration and context

In this context of differentiation with protocols in constant improvement, the cardiac differentiation pathway from pluripotent stem cells represents a real challenge with a lot of critical steps not so understood even after decades of research on it. The perfect one does not exist and research groups who work on cardiac stem cells differentiation need to have tools to better characterize stem cells epigenome/transcriptome during their key point of commitment on the road of the cardiac mesodermal formation. To do so and thanks to the recent advances in sequencing technologies development, a large panel of new methods are currently developed and give the opportunity to researchers to respond to some major critical questions. Nevertheless, heart tissue itself is a very complicated syncytium with a mix of cells with specific functions especially in term of beating forces, electrical integration and electro-transfer. But this is also a major problematic that sequencing can resolve. Indeed, recent development in the field brought new hope to characterize tissue with the new methodology of the single cell sequencing [60]. This heterogeneity and diversity of cells genotype within the same tissue should also be considered. To bring an answer, very sensitive and efficient methods such as Chromatin Immunoprecipitation Sequencing (Chip-seq) and single cell isolation were merged [61].

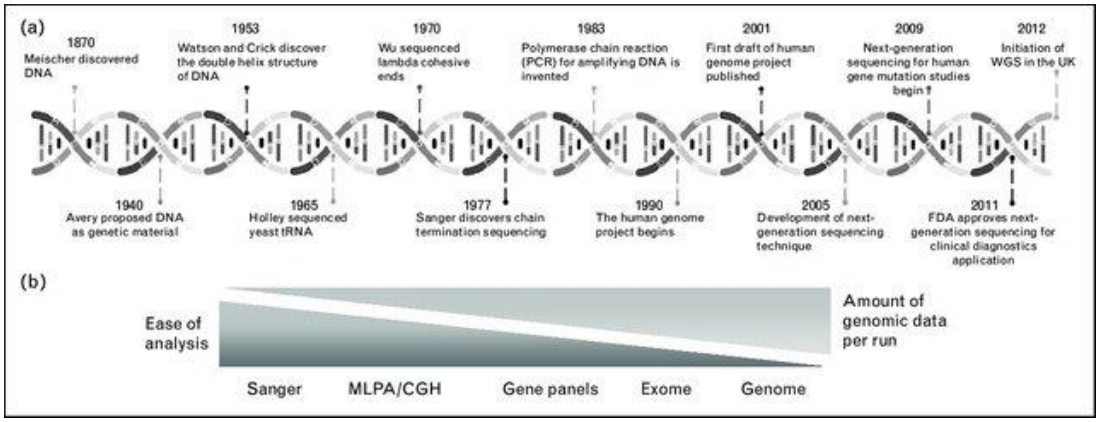
1. History of sequencing

Before developing the most recent sequencing technologies, it is important to make a review of the evolution of this kind of methods to better understand their ins and outs.

A. Sanger and the revolution of the PCR

This is in 1975 that Frederick Sanger created and developed his "plus and minus" technique for DNA sequencing (Figure 4).

Figure 4. Timeline and history of development of genomics and data impact.



a *A timeline of the genomics history* ***b*** *Ease of analysis /amount of genomic data generated per run.*

S. Efthymiou et al. Current Opinion in Neurology, 2016 [62]

Sanger sequencing is the first methodology of DNA sequencing, consisting of a specific incorporation of color marked deoxyribonucleotides (A, T, C and G) by DNA polymerase. All the reaction mimics the real DNA replication in the cell nucleus but is completely made *in vitro* (Figure 5).

Figure 5. How to sequence DNA.

a *DNA polymerase binds to a single-stranded DNA (blue) and generate a neo synthetize DNA strand (red).* ***b*** *When the DNA polymerase incorporate (during the strand synthesis) a fluorescently labeled ddNTP (5' to 3') into the neo DNA strand, the strand synthesis stops. This process produces a pool of different strands length with fluorescently labeled ddNTP at their 3'ends.* ***c*** *Fluorescent DNA fragment separated on electrophoresis gel (Maxam-Gilbert sequencing).* ***d*** *Chromatogram of the sequencing result each color of each peak represent a specific fluorescent labeled ddNTP.*

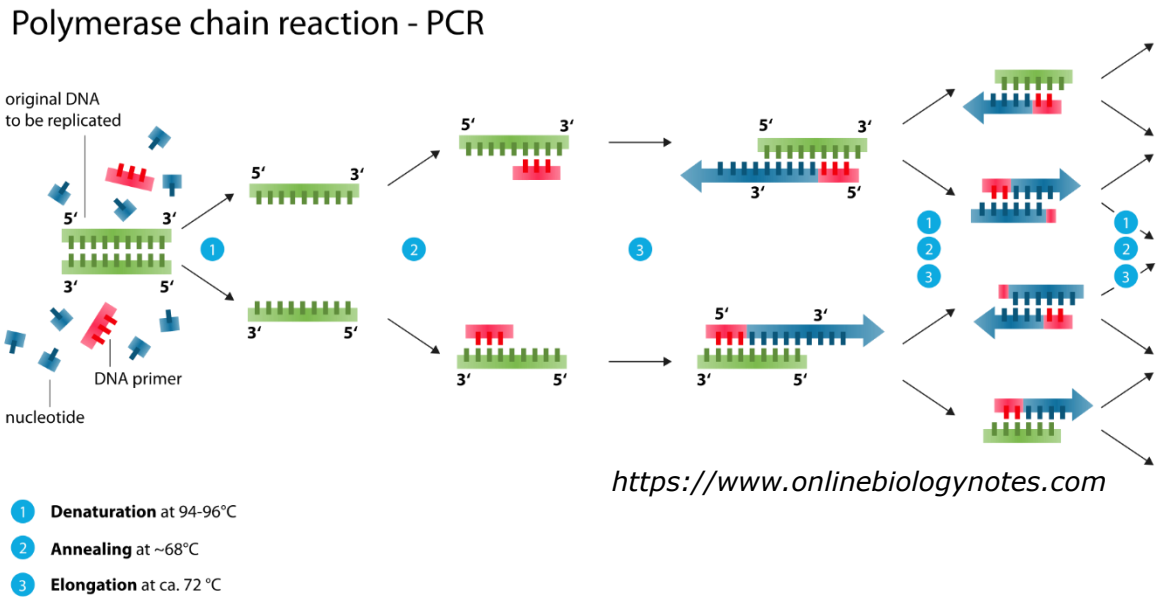
© 2003 Macmillan Publishers, Ltd. Dennis, C. & Gallagher, R. (eds) The Human Genome (Palgrave, Basingstoke, 2001)

The development of the Polymerase Chain Reaction (PCR) appeared after the Sanger technique in 1983 (Figure 4). Created by Kary Mullis which corresponds to an amplification by

million times of a specific fragment of DNA by specific DNA polymerase enzymes (thermostable polymerase like Taq polymerase) with a couple of predesigned primers. Thus, as resumed in Figure 6, this technique is characterized by a repetition of three phases: DNA denaturation, annealing step and amplification step, all depending on a specific scale of temperatures (95-98^{0}C for DNA denaturation, between 50 and 65^{0}C for annealing of primers, depending of their sequence, and 68-72^{0}C for the amplification).

Moreover, a lot of add-ons were made based on the PCR reaction essentially with the addition of fluorescing primers, probes and intercalant fluorophores (Taqman probs, SYBR green etc.) which allow to quantify the number of amplifications and can give to the researcher an idea of the quantity of DNA material present inside the cells [63]. Also, the PCR technique is generally coupled with the reverse transcriptase to generate cDNA from a whole cell RNA population with the enzyme which has the same name.

Figure 6. Polymerase Chain Reaction steps.



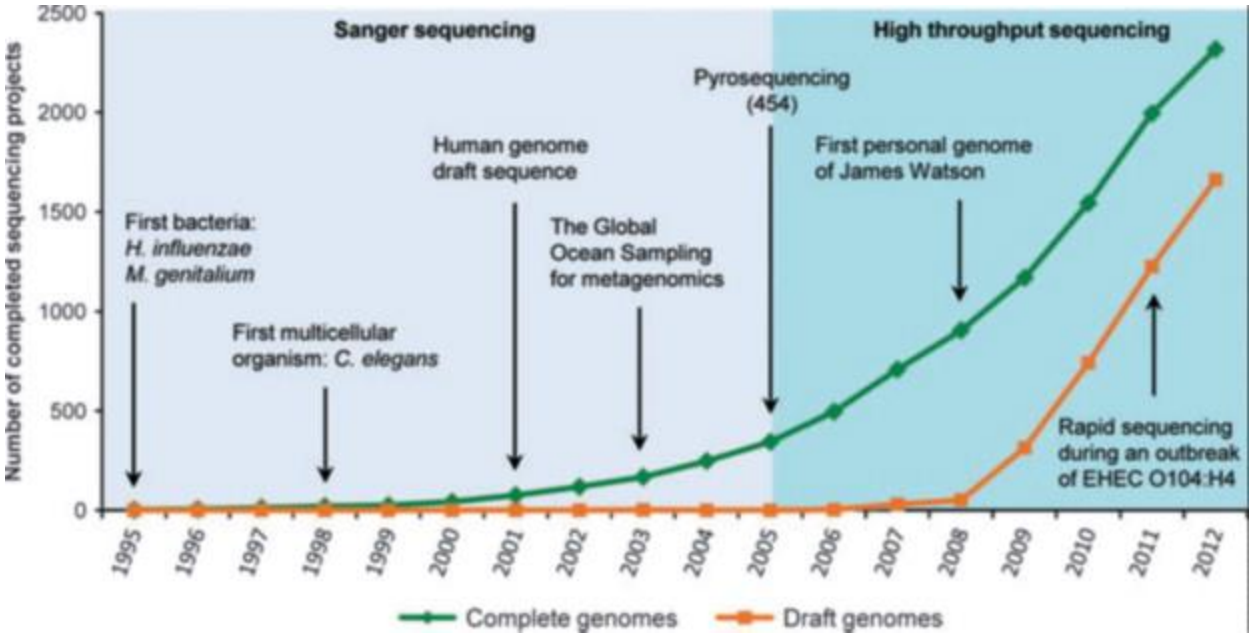
After these two techniques appeared, research groups had the idea to merge both to design a new technology. Indeed, to sequence the global genome from cells, the amount of starting material does not allow to use the sanger technique. We need to cut the reverse transcriptome (cDNA) in small pieces and amplify them through the PCR technique and after that, analyze all the specific primer amplified fragments with a complete "automated sanger sequencing technique" (high-throughput sequencing). This new method corresponds to a recent advanced technique called "Next Generation Sequencing (NGS)" or second generation of sequencing [64, 65].

B. The human genome project and NGS

This is from the impulse of the human Genome project that a new generation of sequencing emerged. Indeed, this is during the development of the high throughput sequencing in the 90's that the human genome project was launched. The new sequencing technology opened doors via a very old idea got after the discovery of the DNA structure in the 60's: completely sequence the human genome [66]. Started in October 1990 and finished in April 2003, this project was the result of an international collaboration to sequence the human genome in its globality [67] (more or less 20,500 human genes). Thus, capitalizing on the advanced technique of sequencing, this ambitious project allowed the development of sequencing methods and materials always more powerful [68].

This context supported the development of new NGS technologies to continuously increase the sequencing effectiveness of DNA. Thus, this is during the first decade of the 21st century that we saw an exploding development of sequencing techniques around the world with the impulse of the human genome project (Figure 7).

Figure 7. Milestones in whole genome sequencing.



C.Bertelli and G.Greub, Clinical Microbiology and Infection, 2013 [69]

These new technologies of high throughput sequencing are characterized by a rapid and cost-effective manner to sequence the genome. Adapted to the type of tissue/organism from which the RNA is extracted, a large panel of protocols were developed to first, increased the possibilities of micro-organisms/cells analyzed, and secondly, adapt sequencing to various contexts, microenvironments and experimental designs [70-72].

Indeed, the application of sequencing always requires deeper techniques, reducing as much as possible the number of starting materials (generally the number of cells used) [73-75].

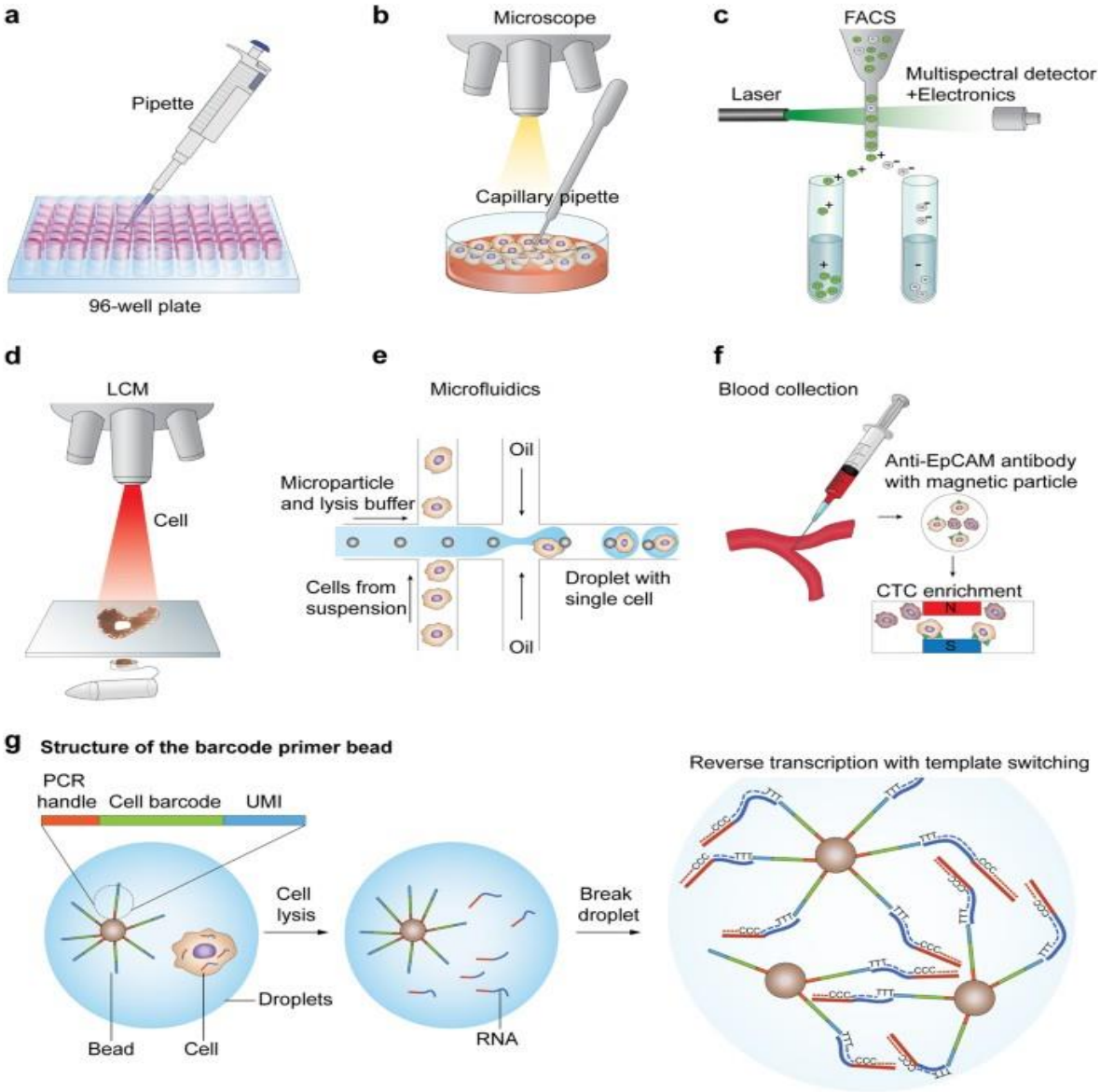
2. Most actual advanced sequencing techniques

From the last couple of years, the sequencing technology became personalized as the library generation is adapted to specific contexts and needs. It is hard to resume all these different technologies, but we will focus on the most advanced ones so far.

A. Single cell sequencing

As previously mentioned, reducing the number of cells/RNA starting material was a critical point for many different applications like in Chromatin immunoprecipitation coupled with high-throughput DNA sequencing (ChIP-seq) [76]. This requirement reached its paroxysm with the emergence of the single cell sequencing [77]. To summarize this technique, it is a complex add-on made on the regular RNA-seq techniques. Indeed, with the possibility offered by the single cell flow sorting FACS (fluorescence activated cell sorting) coupled with a multispectral detector, specific fluorescent cells population can be isolated. Then, with the concourse of nanodroplets microfluidic cell to cell separation, each cell receives a specific fluorescent barcoded primer. Thus, inside each “single cell oil droplet” a specific barcode is added by RT reaction with a specific couple barcoded primer, corresponding to one cell. After the droplets break, neo generated cDNA from each cell can be mixed to start the library for the PCR amplification and add paired end primers compatible with sequencing analysis (all the steps are presented in Figure 8). The real advantage of this technique is to analyze the transcriptome of each cell, thus permitting to compare them.

Figure 8. Single-cell isolation and library preparation.



Byungjin Hwang et al. Experimental & Molecular Medicine, 2018 [60]

Possible methodologies for single cells isolation and scRNA-seq libraries preparation. ***a*** *Limiting dilution method to isolate individual cells.* ***b*** *Single cell isolation using microscope-guided capillary pipettes.* ***c*** *FACS isolation technique which allows a high purification of fluorescent single cells.* ***d*** *Laser capture microdissection (LCM).* ***e*** *Microfluidic technology for single-cell isolation.* ***f*** *The CellSearch system which enumerates CTCs from patient blood samples (magnet + CTC binding antibodies).* ***g*** *Droplet-based library generation which is commonly used for scRNA-seq methodology.*

After isolating multiple interesting cell populations, it is also possible to establish the transcriptomic map of a complete organ/tissue in the resolution of a single cell in a key time moment [78, 79]. Moreover, this technique also allows to understand the *in vitro* microenvironment in a specific culture condition [80] or characterize stem cell differentiation [81].

B. Epigenetic and DNA methylation analysis

Another big move from the sequencing technology is to better define TFs pathway and global epigenome landscape during genetic expression. Moreover, these factors are strongly associated with DNA methylation (histone methylation which allow to relaxing dsDNA strand for gene regulation).

a. Define epigenome and DNA methylation

i. The impact of the epigenome

The epigenome corresponds to all epigenetic modifications occurring inside the cells. It corresponds in priority to the histone and DNA methylation (DNA or histone modification) which allows (for histone methylation) chromatin compaction. Moreover, the epigenome participates in the genetic expression with the concourse of the genetic TFs. They are also able to modify the chromatin state by recruitment of methylation and action on histone complexes and can be classified into two categories. First, it exists repressors of a DNA function by repressing the expression state which downregulates the gene expression level. Generally, the TFs (which allow generally cell expression) can be blocked by diverse recruitment mechanisms notably during the transcription process. Indeed, it exists many different possibilities of cell expression downregulation which involves DNA methylation

and TFs association. Also, interaction between Methylated DNA can inhibit the binding of TFs to DNA. Moreover, TFs can occupy a methylated CpG islet and thereby inhibit transcription, or methylase can play on TF which regulates its activity and will thereby block transcription. Finally, TFs could also directly recruit DNA methyltransferases to form a complex and repress transcription.

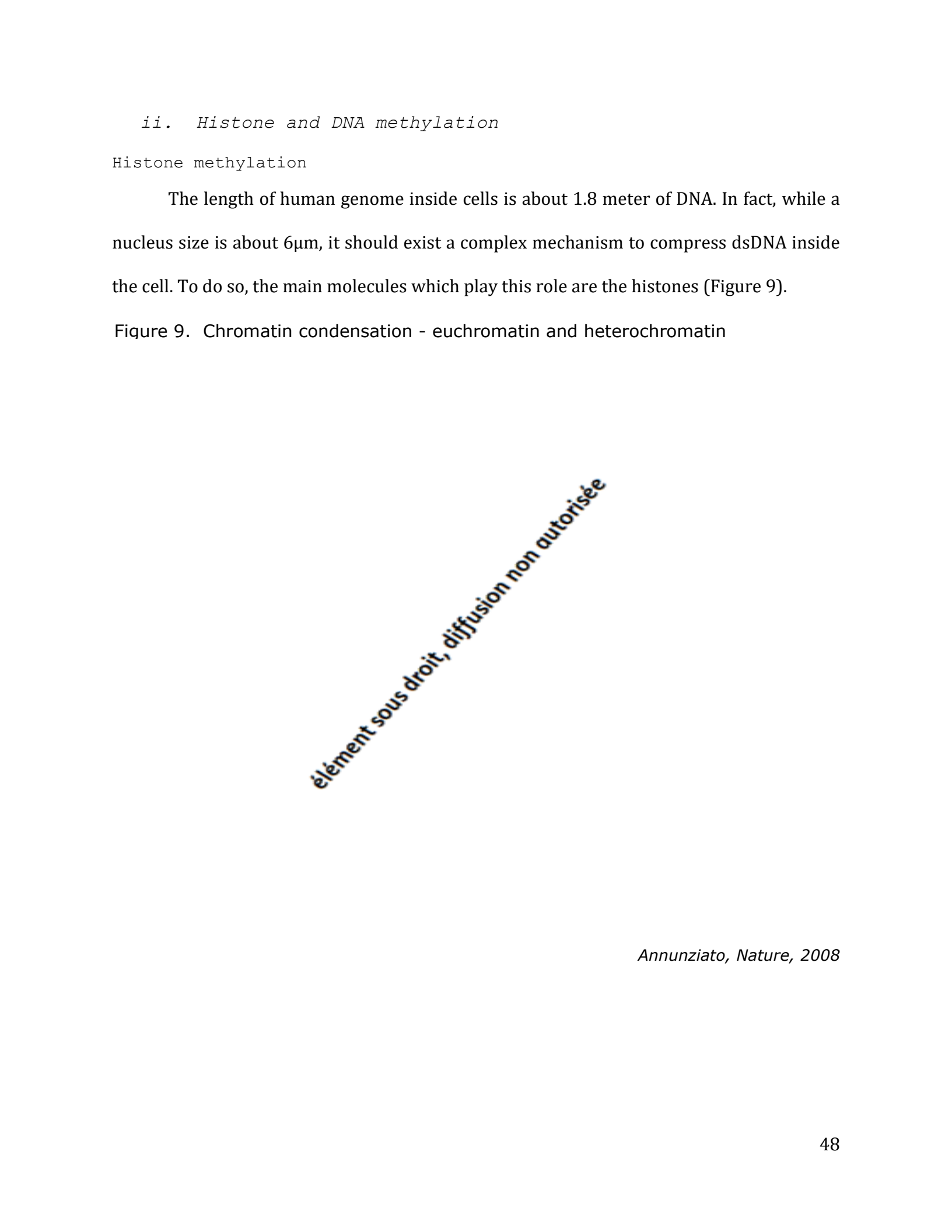
Secondly, they can upregulate genes expression level by a series of activation cascades which could help to enhance transcription of proteins. Thus, this second group of interactions between methylated DNA and TFs creates binding sites for some specific TFs and activates transcription.

ii. Histone and DNA methylation

Histone methylation

The length of human genome inside cells is about 1.8 meter of DNA. In fact, while a nucleus size is about 6µm, it should exist a complex mechanism to compress dsDNA inside the cell. To do so, the main molecules which play this role are the histones (Figure 9).

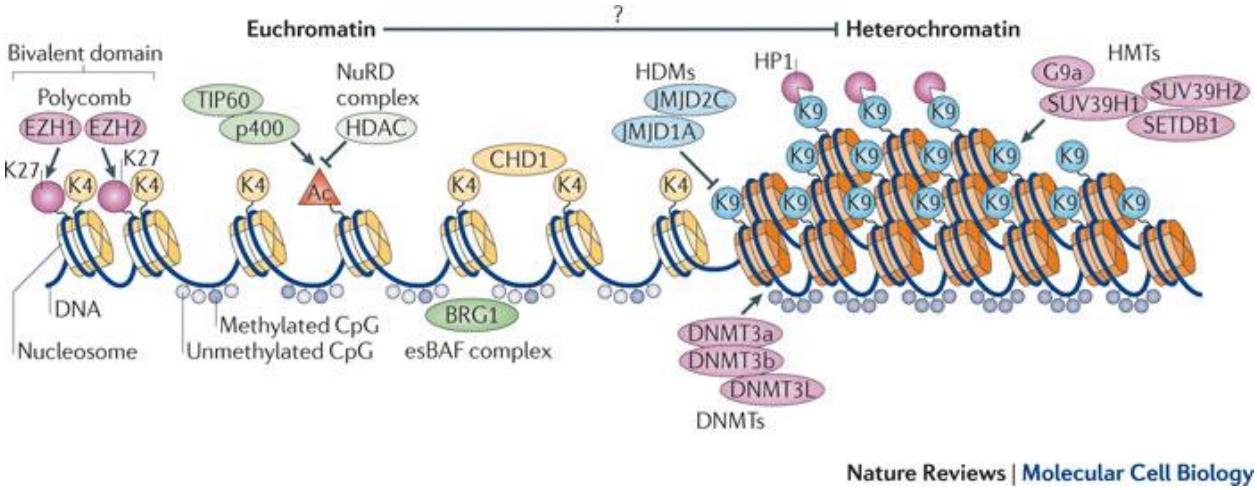
Figure 9. Chromatin condensation - euchromatin and heterochromatin



Annunziato, Nature, 2008

They participate to a complex methylation/acetylation process directly in the peripherical protein region. In order to go from an uncondensed DNA region (euchromatin) to a very condensed region (heterochromatin) the mechanism of histone methylation plays a preponderant role (Figure 10).

Figure 10. Euchromatin and heterochromatin condensation process



Gaspar-Maia et al, Nature Reviews, 2011 [82]

Thus, the heterochromatin presents a condensed structure which remains blocked for the mechanism of transcription and metabolically inactive for gene expression. It is important to know that the mono-methylation of lysine 9 of histone H3 (H3K9me), the repressive tri-methylations H3K9me3 and H3K27me3 and the recruitment of the HP1 protein are essential steps in the formation of heterochromatin. In the opposite, the euchromatin is little or no methylated, usually high in genes with a less condensed structure which allows the mechanism of transcription.

To give more details about the predominant factors of chromatin expression state, as shown in Figure 10, the histone-lysine N-methyltransferase (which methylate the K4, K9 and K27) and also the G9a and the Groups of SUV (Figure 10: SUV39H1, SUV39H2…) play an essential role for the upregulation of DNA transcription.

On the opposite, the histone demethylase proteins JMJD1A and JMJD2C have the role of demethylating the lysin site and through this process, downregulate the gene expression and replication. Moreover, other classes of proteins like the histone deacetylase HDAC1 and HDAC2 play the role of deacetylation by removing the acetyl group on the N-terminal tail of certain histone.

DNA methylation

The DNA methylation is very related to the cellular cycles, differentiation state and the genetic expression. Indeed, the surface of the DNA is sprinkled of methylation sites. When these sites are regrouped in very dense aeras of methylation sites, the area is called CpG Island. These CpG rich sequences play a very preponderant role during transcription and DNA expression. Indeed, around 60-70% of human genes have a CpG island in their promoter region [83, 84]. So, these sequences present strong DNA methylated sites and are corelated with a high transcription area which drives the vast majority of the metabolic gene expression.

b. Sequencing methods in epigenome and histone/DNA methylation context

To come back to our context of transcriptomic analysis technique, many research groups developed sequencing techniques to identify epigenetic markers or quantify the methylation state. Indeed, multiple techniques are available to analyze the epigenetic signature of the genome-wide, e.g. ChIP-seq [85]. This methodology includes non-histone ChIP or histone ChIP analysis. Other technological advances like scBS-seq (single cell genome-wide bisulfite sequencing) [86] have enabled the analysis of epigenetic signature in

genome-wide at single-cell level. In this method, post-bisulfite adaptor tagging is used to detect methylated cytosines in genomic DNA from single cells. The critical parameter is the genomic DNA treatment with sodium bisulfite, which allows DNA fragmentation and conversion (all cytosines are converted into uracil except for the methylated cytosines which are protected by a methyl group). Then, a random priming PCR amplification is engaged followed by a deep sequencing which provides a single nucleotide resolution of methylated cytosines from single cells. The most recent method called scM&T-seq (single cell genome-wide methylome and transcriptome sequencing) is based on the G&T-seq (genome and transcriptome sequencing) method, previously described by the same group, which brought more information about the level of expression of pluripotent genes such as *Esrrb*, negatively associated with DNA methylation [87]. Thus, determining with high efficiency the transcriptional variability in heterogeneous cell populations under the control of epigenetic events (DNA methylation) can also give information about different epigenetic states/expressions of key genes of pluripotency (*Esrrb*, *Nanog*, *Oct4*) possibly associated with the commitment of stem cells. Furthermore, parallel profiling of methylome and transcriptome can make a link between genomic and transcriptional heterogeneity by a thorough analysis of epigenetic events at a single cell genome-wide level. The understanding of genome-wide epigenomic in cell population becomes a key challenge for the future in particular in stem cell population [4]. The gene expression of the stem/progenitor cells is temporally regulated by transcription factors (TFs) which bind the DNA, thus upregulating or downregulating critical protein pathways involved in the commitment of cells during differentiation process. TFs are also able to modify the chromatin state by recruiting methylation and action on histone complexes [88]. The modification by methylation and

acetylation of particular histone families due to transcription factors (TFs) can control and regulate the potential differentiation and commitment of stem/progenitor cells [89, 90]. Focusing on the expression impacted by the TFs regulation: the transcriptomic expression is also a very big piece of investigation especially with the very recent development of the whole genome sequencing technology [5]. Thereby, new techniques quickly emerged in the field of epigenetic study like the Cap Analysis of Gene Expression (CAGE) [91, 92] or the Chromatin Immunoprecipitation sequencing (Chip-seq) [93, 94], which can give a better comprehension of the specific epigenetic profile/differentiation of cell populations.

Theses epigenetic modifications are involved in the commitment of the stem/progenitor cells in differentiation pathways [95, 96]. Thus, developing a sequencing tool to identify TFs for stem cell differentiation would be very useful.

Multiple potential applications could be possible: in fundamental research to characterize the relationship between genome profile and transcriptional expression in heterogeneous cell populations from a same tissue, but also in clinical contexts such as heart regeneration or oncology.

c. Other sources of Methylation, the RNA regulation

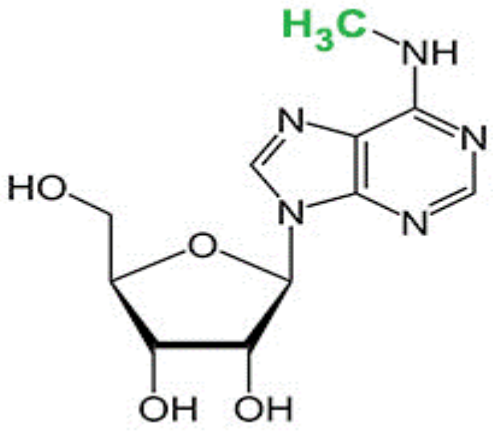
i. RNA methylation

The RNA molecule also appears to have methylations. Thus, a reversible post-transcriptional modification of the RNA corresponds to the addition of a methyl group and this one has a tremendous impact in numerous biological processes. This post-transcriptional addition is every time due to the same kind of molecule: the methyltransferase, which is present in many cell compartments (nucleus and cytosol).

Moreover, methyl addition can also be present in all forms of RNA including tRNA, rRNA, mRNA, tmRNA, snRNA, snoRNA, miRNA, and viral RNA.

The most common and abundant methylation present on the RNA corresponds to the N6-methyladenosine (m^6A) modification (Figure 11). Thereby, this particular methylation corresponds to the vast majority of RNA Methylation populations (80%), and is implicated in many physiological roles with a critical regulation during the embryonic development and the cell fate [97].

Figure 11 N6-methyladenosine (m^6A)



www.epigentek.com

But the most interesting impact is regarding the mRNA itself. Thus, this is through a phenomenon organized by the methyltransferase like 3 (MELLT3) which is responsible for an enrichment of m^6A at 3'UTRs region of the mRNA and also participates to an extreme precocious mRNA expression regulation through the YTHDF protein group [98].

5-methylcytosine (5-mC) is an epigenetic mark which also commonly occurs in various RNA molecules.

Recent studies suggest that m^6A and 5-mC RNA methylation affects the regulation of various biological processes such as RNA stability and mRNA translation [99]. Indeed, through the group of YTHDF proteins RNA could be drive to the transcription system or the p-bodies secretion; the RNA degradation process.

ii. Sequencing techniques related to RNA methylation

Research groups reported the successful conversion of RNA by bisulfite modification. They also performed this process in a sequencing context by the tagging of the methylation and coupled with an RT-PCR amplification, it is possible to detect the methylated spots on the RNA sequence directly by sequencing technology [100]. Thus, by treating RNA with bisulfite, cytosine residues are deaminated to uracil while leaving 5-methylcytosine intact easily detectable during the sequencing process.

C. Revolution of CRISPR-Cas9 and the sequencing

CRISPR-Cas9 corresponds to the most advanced genome editing technology [101, 102]. Presented as "genetic scissors", CRISPR-Cas9 needs to have a precise overview of the transcriptome in order to target a specific zone. Indeed, before making any genome engineering, it is important to know precisely the ins and outs of the genetic disease targeted. Precisely knowing the DNA sequence of the future target area is also a requirement. It is in this context that this sequencing technique grew significantly [103]. So, with the emergence of these new genome editing technologies, it is always more urgent to adapt sequencing technologies to CRISPR-Cas9 (personalized sequencing for each patient). Thus, have the possibility to sequence the whole genome of each patient would evolute CRISPR as a complete personalized gene therapy which is a very potent commitment for this new century. To do so, sequencing techniques become deeper to give suitable results and guide researcher in the selection of DNA target in order to use CRISPR-Cas9 edition [104]. In fact, without the human genome project and the overall advances in sequencing technology, CRISPR-Cas9 would be a very useless tool.

3. Sequencing technique in the heart regeneration

In the large cardiac regeneration therapeutic goals and their process of development, the sequencing technology remains a must. Like many other tissues, the sequencing methods helped first to better characterize molecular pathways in cardiac diseases [105, 106]. Moreover, in the area of cell/stem cell transplantation in the field of cardiac regeneration, many factors remain unclear and a large effort of investigation *in vitro* should be done by different kinds of analysis like genomic and transcriptomic studies. Thus, epigenetic became a powerful investigation tool in the last decade. Some doors were opened with the apparition of epigenetic analysis in the field of heart regeneration. The principal goals are to better understand the capabilities of stem/progenitor cells to increase the cardiomyocytes proliferation and better target the remodeling and dedifferentiation of cardiomyocytes during the regeneration process. Indeed, epigenetic modifications which are defined by DNA methylation, histone modifications and miRNA mediated gene regulation are also associated with cardiac degeneration and regeneration, in particular with the differential capacities of stem cells but the mechanisms are still unclear [107]. Better characterizing stem cells during the cardiac differentiation is also critical. As expected, those studies are largely made on the hPSCs-CMs process of differentiation [5, 108-110].

Nevertheless, other sources of CMs generation were also analyzed with these methods. MSC derived CMs, are also studied by a deep sequencing analysis essentially to better define the Wnt pathway mediated regulators [111, 112].

Moreover, the CDC derived CMs or the cardiac progenitor cells (CPC) do not escape from a complete analysis into RNA-seq [113]. Also, sequencing technology helped for a global transcriptional analysis of the mammalian heart regeneration capabilities [114] even though the heart is largely considered as non-regenerative.

A tremendous effort brought by the high throughput sequencing is principally due to the single cell sequencing [115]. They used single nuclei RNA sequencing on the whole cell population of the heart tissue to define the transcriptional and cellular diversity in the normal human heart. Thus, the global normal heart transcriptome characterization on a resolution of single cell is right now well-defined and opens doors to more investigation notably in pathological conditions.

4. Actual remaining gap and thesis objectives

Even if all of these methods are very powerful tools, it should be noted that they involve a physical precipitation/conversion of DNA methylated sites, either by immunoprecipitation: Antibody used to detect DNA methylation (chip-seq) or bisulfite conversion of DNA (scM&T-seq) with no possibilities of targeting specific epigenetic events. Moreover, Chip-seq is also a very robust technique to check the level of histone methylation. These techniques give information on the methylome. The methylation of CpG sites is taken as an indirect evidence of transcriptional repression at those loci. Multiple studies have recently suggested however that DNA methylation is not a primary controller of genomic TF landscapes [116, 117]. These limitations have encouraged the development of techniques focused on opened chromatin (euchromatin). Indeed, researchers always kept going further and recently developed a fast and sensitive epigenomic profiling of opened chromatin [75]. This ATAC-seq (Transposase-

accessible chromatin using sequencing) uses a particular propriety of hyperactive transposon Tn5 called "tagmentation" [118, 119] which simultaneously fragment any double stranded DNA and insert sequencing adaptors (primers) allowing the direct recovery of a high-throughput sequencing library. Also, fast analysis of epigenome expression is compatible with the clinic in terms of timescale [75]. The tagmentation process of the Tn5 is completely random and cannot be directed against specific transcription factors. Coupling the Tn5 transposase with a specific antibody would be a very powerful model to guide tagmentation on a specific epigenetic mark.

The **First thesis project** was to develop a new sequencing technology to better identify the key transcriptomic factors during the stem cell differentiation and their myocardial commitment.

Herein, we propose to address this major gap in know-how by developing a new method of single cell epigenetic analysis allowing to give a multiplexable targeted epigenomic sequencing by **Chromatin Histone Immune Coding coupled to whole genome sequencing (Chic-seq)** in genome-wide at a single cell level. The cut and paste capacity from the transposon protein present very interesting possibilities especially for an *in vitro* genomic analysis. Thus, a lot of company design kits and specific material to use the Tn5 propriety as molecular tools. Indeed, Illumina, Inc proposed their own Nextera® sequencing kit based on the Tn5 activities [120, 121]. Nevertheless, lucky for us, a research group published a protocol to help other laboratories to produce and use this 'home-made Tn5' and it is based on this work that we decided to produce our own Tn5 protein [122]. After a long

process of optimization, we finally found a design (called Tn5 loop) which seems to work for our own purpose.

nevertheless, we were forced to stop this first thesis project when we discovered a patented technology very close to our current Chic-seq design and method. But I decided to capitalize on the previous project on the Tn5 protein and came back to my thesis directors with a second thesis project idea: developing an innovative approach to sample intracellular RNA in a non-biased and non-destructive manner. Indeed, as we already know, many different transcriptomic analysis methodologies have been developed during the past decades especially in sequencing [123, 124]. Microarray mRNA profiling or RNA sequencing, for example, have allowed to push the frontiers of knowledge and give a better understanding of cell heterogeneous architectural complexity. Despite the development of many techniques with increasingly powerful resolution, a gap remains. Thus, a large majority of current techniques are cell destructive [125] and are then incompatible with *in vivo* time points studies. Conversely, it exists *in situ* hybridization in live cells with mRNA probes [58, 126], but it is difficult to use routinely, hardly multiplexable and cannot reflect the whole cell transcriptome over time.

Many efforts have been focused on reducing the number of cells in transcriptomic analysis [75], with the objective to follow specific cell types or single cell mRNA expression [127, 128]. Indeed, the goal is to reduce the experimental noise and get enough representative results with very little starting material. Even if a new approach has been developed with a method for time-resolved, longitudinal extraction and quantitative measurement of intracellular mRNA [129], this technique has been designed for *in vitro*

studies and requires specific material, unadaptable in routine or again for *in vivo* studies. Even though, each method offers specific advantages designed in function of the goal needed, none of them proposes a technique using the intracellular mechanism to secrete a representative and unbiased part of the transcriptome for a monitoring of live cell in physiological or pathological conditions. With the fast development of new therapies deriving from biological products and the advanced therapeutic products, it is urgent to have tools for the validation and the understanding of these new therapies like cells transplantation, stem cell differentiation in animals etc.

Herein, as **second thesis project**, I proposed a new methodology: TRACE-seq (**TRanscriptomic Analysis Captured in Extracellular vesicles using sequencing**) which provides a monitoring transcriptomic analysis, non-destructive and compatible *in vitro* as well as *in vivo* which gives to researchers a very powerful support tool of development for a lot of different kinds of studies like tumor monitoring, organ transplantation and cellular therapies analysis and development etc.

First Part: A novel method for highly multiplexed epigenetic analysis at single cell resolution

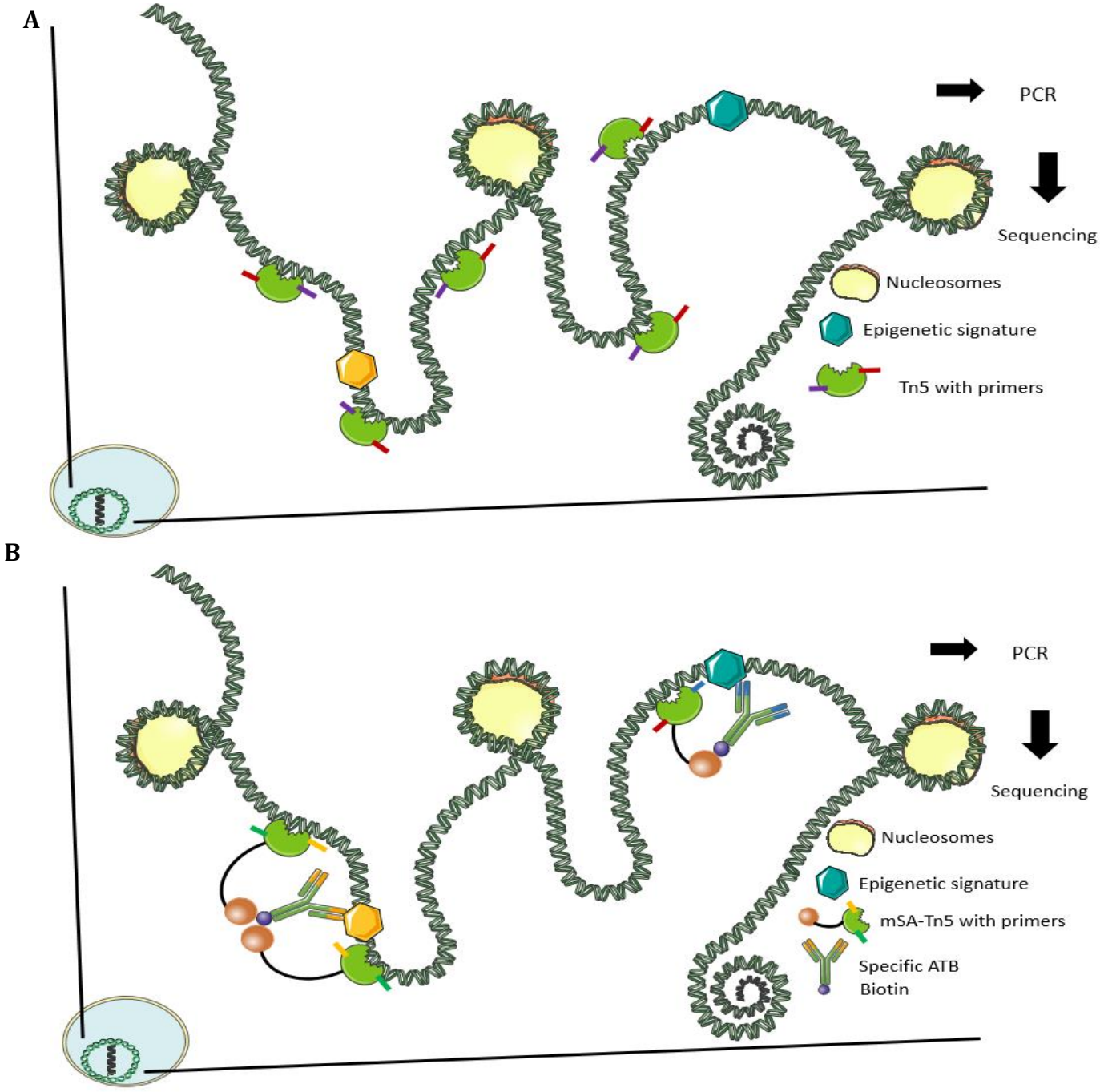
I. Introduction

Having the possibility to propose a multiplexed epigenetic analysis coupled to whole genome sequencing would be a very powerful tool in the field. **Chic-Seq (Chromatin Histone Immune Coding coupled to whole genome sequencing)** offers this possibility by adapting and merging single cell analysis and the ATAC-seq technology with an antibody targeted recognition to improve the specificity of Tn5 transposase on epigenetic or transcriptomic markers previously selected. To address this, we propose to use a modified Tn5 linked with monomeric streptavidin (mSA). In fact, the monomeric streptavidin was chosen for its small size of 12.4 kDa and it seems to be more stable and present high-affinity for biotin [130]. The same research group also showed that the mSA molecule can be used as a genetic tag to introduce biotin binding capability to a heterologous protein [131]. Therefore, we designed a new protein called Tn5-mSA and its isomer mSA-Tn5 to achieve our goal. First, we want to use specific Ab-biotin against previously selected TFs or epigenetic modifications (Fig 12). Second, we would add the Tn5-mSA and play with the complementarity of Biotin-mSA. The Tnp could do its tagmentation process by placing

specific primers (blue/red or yellow/green Fig 12) in the DNA sequence near TFs/DNA binding sites. These fragments could be easily amplified by PCR and sequenced.

Also, it could be possible to get a multiplexable analysis by playing with different antibodies with different corresponding primers. The capabilities of the Tn5 transposase were used like

Figure 12. Presentation of the new methodology



A. actual ATAC-seq B. Chic-seq design

a “foot printing genome wide sequencing”. Interesting fact, a group recently merged Crispr with the Tn5 transposase protein to create a genome editing technology capable to insert a large amount a dsDNA in a designated site [132].

II. Bibliography context

To complete our goal and create a new sequencing technology, we need to find a way to link the Tn5 transposase to a specific antibody. A long preliminary work was done in terms of design, tests and validations to carefully verify the reactivity of the new tagmentation complex. Indeed, each part of this new methodology should be designed, produced and analyzed, step by step with appropriate controls. By nature, the Tn5 transposase is an unpredictable protein [133] and requires a very large panel of experiments to determine the feasibility and robustness of a new transposon design. But before presenting all this work, we need to verify in the literature the role of each protein to be used.

1. The transposon protein family and their role in sequencing

A. The transposon superfamily

Before defining a design and planning a cloning strategy, it is very important to clearly understand the proteins functionality to achieve our goal. Also, the transposon protein mechanisms, discovered in the early 50's by Barbara McClintock [134, 135] still remain obscure today, after 70 years. Thus, the Transposon proteins are related to different groups [136] (Class I TEs or Class II TEs) present in both prokaryote and eukaryote cells. To simplify, we can note two large types of transposons: the retrotransposons, using the reverse transcription function as element of transposition and the DNA transposon, generally

mobilized by a transposase protein. Moreover, the important transposon family for the purpose of this study corresponds to the second group of DNA transposases, group which includes our Tn5 transposase protein.

B. Transposons and Tn5 protein

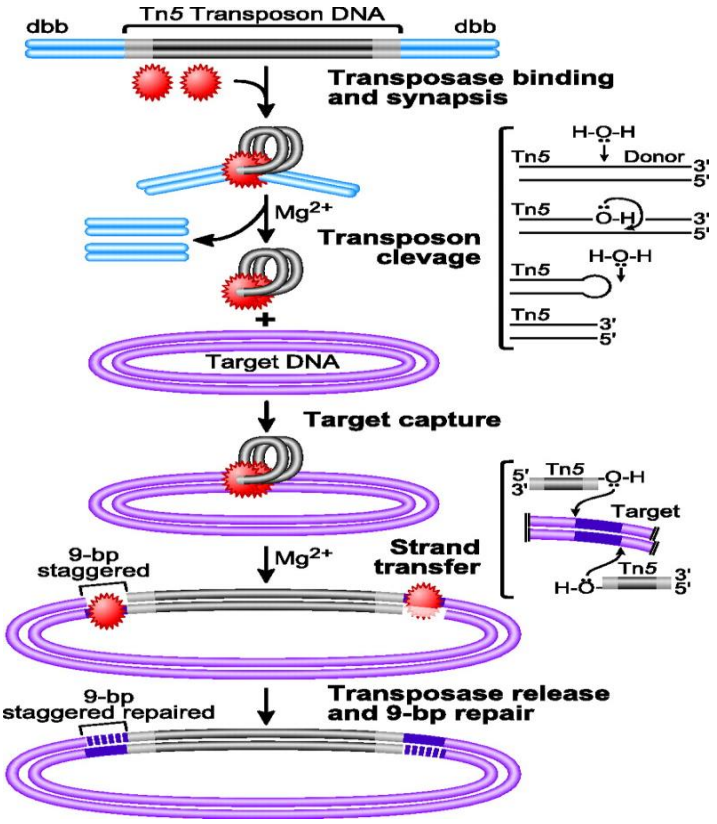
a. The Transposon class of proteins

The Tn5 transposase protein came from the subgroup of the DNA transposon (Class II TEs) like others such as Tn10 or Tn3 etc. This protein was one of the first transposases identified and discovered by Julian Davies in 1973 [137, 138] during researches in order to elucidate the neo adaptation of certain bacteria to the kanamycin antibiotic. Indeed, Berg and Davies understood the key role of the Tn5 transposase for the adaptative response of bacteria to kanamycin. So, this is by the mechanism of transposition (through the Tn5 protein) of the transposable element, the kanamycin gene, that bacteria acquired their resistance to antibiotic (kanamycin). Moreover, this is through this mechanism that cells have an adaptative response to environmental stress [139]. We can also highlight the key role of transposase recognized as potent agents for the phylogenetic adaptation [140]. Nevertheless, the mechanism of transposases was really understood by the long work of William S. Reznikoff in the early 80's [119, 133, 141]. It is after 40 years of work made by the Pr Reznikoff group that the Tn5 transposase protein became the most well-known transposon today, which justifies the interest of different researchers/companies to use its marvelous capacities to "cut-and-paste" ds DNA [142].

b. The Wild-type Tn5

Like all transposase proteins, the Tn5 has surprising transposition skills of dsDNA without any sizes requirement. So, it can relocate any dsDNA from point A to point B in the genome [143] (fig. 13). This is through a complex mechanism which can be broken down in several parts:

Figure 13. Wild type Tn5 mechanism of Transposition



T W Wiegand et al., J Bacteriol, 1992 [144]

First, the Tn5 recognizes the Mosaic target sequence ES: End sequence (light gray Fig. 13) at the ends of the Transposon DNA sequence. Two Transposases (Tnp) form a binary synaptic complex composed of two ES-bound Tnps which form the required scaffold for the subsequent catalytic steps. Each Tnp bound at each ES sequence catalyze a DNA cleavage and release those sequence from the whole initial structure. This cleave corresponds to three catalytic steps called the transferred strand which can be assimilated to a hydrolytic nick reaction catalyzed by the transposase (using water as nucleophile and coordinate by Mg^{2+}). A first activated water molecule will – through

a nucleophile attack- generate a 3' OH group in one DNA strand (Transferred strand TS). Then, this 3'OH generation on one of the DNA strands will create an unstable structure which results in an attack of the opposite strand to generate a hairpin structure. A second activated water molecule resolves the hairpin, resulting in a double-stranded DNA cleavage product (with a 3′-hydroxyl group free on one DNA strand end [on the transfer strand TS]). As a result of this cleavage process, each Tnp-ES complex present at each DNA end is bound together and form a completely activated synaptic complex ready to insert the transposon DNA through a target DNA [145, 146].

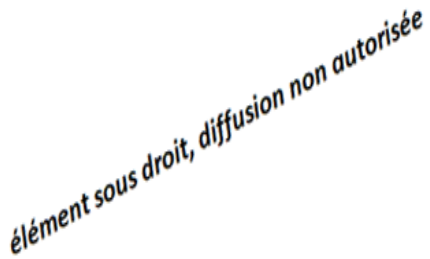
Secondly, the Loaded Tnp Dimer (with its transposon DNA sequence) can detect and bind a DNA target (any dsDNA can be targeted by a Tnp Synaptic complex) [147]. This step, called strand transfer, corresponds to the attack made by the 3′-hydroxyl group of the transposon strand transfer end on the phosphodiester backbone of target DNA. Due to the staggered strand transfer reactions, a 9bp duplication is generated in the target strand initiated by the host DNA repair mechanism [119].

By the study of this mechanism of transposition, we can now understand how it becomes possible to use an engineered Tn5 in a different approach, especially for cloning enhancer strategy and mostly for sequencing tools.

C. Applicability of the transposon in biology

a. EZ: Tn5® and cloning strategy improvement

Figure 14. EZ: Tn5® mechanism of action

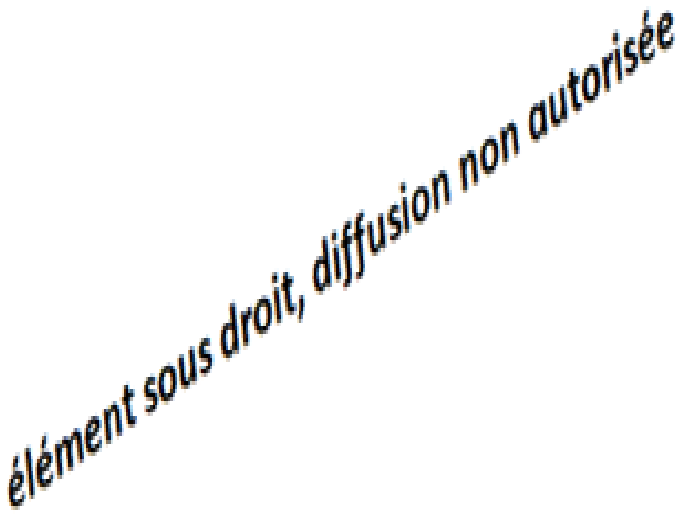


Assigned to Epicentre® (an Illumina® Company).

The first engineered function used for the Tn5 transposase was the EZ mechanism made by Epicentre® (now part of the Illumina® Company). It consists of a boost of a cloning capability through the transposition mechanism of the Tn5 transposase (Tnp) playing with the drug resistance inside a target plasmid. Indeed, the drug resistance gene is flanked by the two Mosaic sequences and the Tnp can form dimers with the synaptic complex. Once the synaptic complex is formed, it just needs to incubate it with any plasmid and the drug resistance gene could be inserted. Nevertheless, as we previously saw, a synaptic complex can target any plasmid sequence, so it is really important to control the ratio of synaptic complex/plasmid to limit the insertion to one drug resistance gene per plasmid (Fig. 14). This optimization of the Tnp function opens the way to its major role in the world of biotechnology, as a tool for sequencing library generation.

Thus, the Illumina® company thought that it was possible to use the Tnp as a sequencing library generation tool as described in the ATAC-seq paper [75]. They capitalized on all the research and development made by this group and especially on the ATAC-seq paper and the work from the Reznikoff's team. The principle is the same but instead of loading the Tn5 transposase (Tnp) dimer with the full dsDNA sequence, it loads with

Figure 15. Nextera® sequencing kit mechanism of action



Nextera® technology, Illumina® Company.

specially designed primers (with the Mosaic sequence) (Fig. 15). Each dimer of Tnp is related to a couple of oligos/primers which can tagment any dsDNA. The attacked dsDNA will be fractioned and the oligo loaded on the Tnp will be inserted. It is now easy to amplify the sequence with a second reverse complement primer corresponding to the flanking sequence of the first one and in the meantime to add specific barcode for NGS sequencing. Thus, the library is ready (fractioned, barcoded and amplified) to be sequenced with a specific machinery. The strong advantage of this technique is mainly on the time used to generate the library which is considerably reduced compared to other methodologies. Moreover, the Tnp tagmentation is an enzymatic reaction which is less variability dependent during fragmentation process than sonication.

Moreover, Illumina® constantly improves their technology. The Nextera® kit was upgraded to Nextera XT® which is compatible with very low amounts of starting material (less than 500 cells) and they also sold the Nextera flex® which is basically ready to use Tn5 transposase (Tnp) protein and sedimented on beads. So, the reactivity of the protein seems better and significantly improved the global library generation, as seen from the Illumina's brochure. All these technologies are based on an uncontrolled Tnp. Thus, the idea to couple the transposase protein with an antibody to give them a targeting ability toa specific epigenetic marks or a specific TF binding site would be a significant move forward. Thus, we decided to use one of the most common protein interaction systems to link the Tnp to an antibody: the Biotin-Streptavidin couple.

2. The streptavidin class of protein and the monomeric streptavidin mSA

A. The streptavidin family

Because we know that the Tn5 transposase is very active in its synaptic complex form after the dimerization, we chose the linking strategy of biotin/streptavidin to attach the Tn5 transposase to an antibody. The Biotin/Streptavidin couple is one of the oldest protein mechanisms used in biology [148, 149]. This "old school manner of binding" remains routinely used today and is recognized as one of the strongest noncovalent interaction in

biology [150]. Nevertheless, researchers designed different types of subgroups of streptavidin by playing with its structure to give them more reactivity and less density.

Even if the reactivity of the streptavidin to the biotin is strong, research groups tried to improve it and especially tried to reduce the number of binding sites of the streptavidin. Indeed, the streptavidin 3D conformation has 4 binding sites with a structure quite similar to the hemoglobin. Nevertheless, for some studies especially in kinetics, it was necessary to have a streptavidin with only one biotin binding site. This research group developed in 2006 the monovalent streptavidin [151, 152], which corresponds to the same density as the wild type streptavidin but with 3 blocked biotin binding sites. In our case, we needed only one biotin binding site to reduce the number of antibodies by Tn5 molecule to a 1:1 ratio. But with this streptavidin, the molecular weight was problematic for our design (52.8 kDa), and its capability to bind multiple biotins (four in total, even if ¾ were theoretically inactive) was not compatible with our purpose.

B. The monomeric streptavidin mSA

Luckily for us, another group developed a new synthetic form of the streptavidin: the Monomeric streptavidin (mSA) [130, 153]. This new streptavidin is a monomer, meaning that it corresponds to just one part of the streptavidin with one biotin binding site and a reduced molecular weight of 12.8 kDa. We know that the biotin is a very small compound, but by using this small mSA we can greatly reduce the potential problem of steric hindrance and keep the action of each site of our fusion complex safe.

III. Material and method

1. Cloning strategy

In order to save time, we directly bought the plasmid PTXB1-TN5 (addgene #60240) which is used with the NEB Impact kit to purify the Tn5 protein as used in the Picelli protocol [122]. We also prepared two other constructs PTXB1 mSA-TN5 and PTXB1 Tn5-mSA by cloning the previously amplified mSA sequence from the pSRET mSA plasmid (addgene # 39860) with the following enzymes: Xba1-Nde1 and Spe1. Moreover, between both proteins Tn5 and mSA a poly Serine-glycine chain of 15 amino acids (SGGGG x3) was added. All plasmids were produced into the DH5 alpha strain before the protein production itself.

2. Tn5 production

A. Protein Expression

After each PTXB1-TN5, PTXB1 mSA-TN5 and PTXB1 TN5-mSA are cloned into the competent cells C3013 (NEB#C3013), an aliquot of the stored solution (LB with 20%glycerol) was used and a starter culture of 5ml in a 14ml round bottom tube (falcon, 836 North St # 300, Tewksbury, MA 01876) incubated overnight (16h) at 37°C (aliquot saves by a spin 13K rpm 4°C for 10 minutes) was saved if needed by spinning and storing the bacterial pellet at –80°C. As mentioned from the Picelli protocol [122], we used the Low Temperature Expression methods: we took the whole starting culture (5ml) and brought the volume of culture up to 500ml. The cultures were shaken for 2 h at 37°C and the OD at 600nm was verified every hour. To do so, 0.3mL of culture media was transferred into a second shaker

for 20 min at 22°C. The OD value was checked every 30 min until it reached 0.6-0.9. At this step, we saved 50 µl of sample for a non-induced cell control (Sample 1). Then, we performed the IPTG induction with 0.5 mM of IPTG for 4h culture at 22°C and saved 50 µl sample for an induced cell control (Sample 2). After this IPTG incubation, the culture media was spun at 4K rpm 4°C for 10 min and a saving step is possible by freezing bacterial pellet at -20°C.

B. Protein Purification

The 3 Tn5, mSA-Tn5 and Tn5-mSA-Tagged Proteins were purified from Bacteria with the following preparation protocol.

First, we resuspended the bacterial pellet from 500ml culture in 6 ml chilled Lysis buffer (Sigma-Aldrich, Saint Louis, MO, USA) + 24µL lysozyme (Sigma-Aldrich, Saint Louis, MO, USA) + 0.5µL nuclease (NEB, Ipswich, Massachusetts, USA) + 60 µL Protease inhibitors complete (Sigma). Next, we transferred into a 50ml conical tube and left rotating for 30 minutes at 4°C. We saved 50 µl sample for lysate control (Sample 3) and transferred to a 50 ml JA20.1 centrifuge tube for a spin at 15K g at 4°C for 10 minutes. Another aliquot of 50µl was saved as a supernatant control (Sample 4). After the spin, we lysed the bacteria, transferred the supernatant in a 15 ml tube and diluted 1:3 with cold binding buffer (50 mM HEPES-NaOH ph 7.2, 1 mM EDTA, 0.5 M NaCl and 20% Glycerol) + 90 µL of protease inhibitor (Fisher Scientific, Hampton, NH, USA). While the lysed bacteria were spinning (15 K g 10 min at 4°C), we prepared the chitin beads (NEB, Ipswich, Massachusetts, USA) and used a cut tip (P1000) to take about 5 ml (50% slurry) and transferred to a 50 ml tube, brought the volume to 8 mL with cold binding buffer and spun them at 4K g at 4°C for 2 min. We removed the

supernatant and washed the beads at least 3 more times with binding buffer. After the final wash, we brought the volume to 20ml total and resuspended the beads with the diluted bacterial lysis supernatant. We let rotating at 4°C for 30 minutes to 1 hour and spun at 4K g at 4°C for 2 minutes. We removed the supernatant (except for 50µl; saved as Sample 5), washed 5 times with 2.5-5 ml of wash Buffer (50 mM HEPES-NaOH ph 7.2, 1 mM EDTA, 0.8 M NaCl, 20% Glycerol) and left it rotating at 4°C for 5min followed by a spin of 4K g 2 min (occurred between each wash). For each wash, we saved 50µl of sample (Samples 6-10). Finally, we washed 2 more times with cold elution/Cleavage buffer. After the last wash, we brought the volume to 5 ml with cold Elution Buffer + 100mM DTT (Sigma-Aldrich, Saint Louis, MO, USA) + 60µl Protease inhibitor cocktail (Fisher Scientific, Hampton, NH, USA). Elution mix was left rotating at 4°C for 60h. One hour before the end of the cleavage, we prepared the 50 kDa filter (Millipore Burlington, Massachusetts, USA) by addiction of 1 bed vol of cold dialysis buffer stored 1h on ice. After the 60h rotation, we spun down tubes, 4 K g for 2 min at 4°C and saved 50µl sample from the supernatant (sample 11 eluate sample). We also saved 80µl of chitin resin beads by adding 50µl of 3X SDS buffer (Sample 12). Next, we filtered the supernatant with the prepared Millipore filter tube (50kDa, 15 ml tubes), spun at 4K g for 15-20 min and analyzed the filtrate with Nanodrop (typical range: 1.1 mg/ml) we washed the filter 3-4 time with 2ml of cold dialysis buffer to bring down [DTT] around 1-5 mM, let the sample with Dialysis buffer with final concentration of 1 mix/ml. At that final step, we took ¼ of this sample and added 1.1 vol of 100% glycerol and incubated with the ready dsDNA oligos form the Tn5 primer ready mix and directly stored at -20°C for a short-term use. The ¾ rest was flash frozen and stored at -80°C, prepared 1 mix/ml (work aliquot).

3. Tn5 validation

A. Annealing Oligos

The ssDNA Oligo from the master solution (100mM) were mixed together according to the following solution:

-35 µL H2O

-5µL of 10X TE buffer (100mM Tris, 10mM EDTA, 1M NaCl)

-5µl of each oligo

Total volume of 50µL.

The solution was boiled for 5min at 95°C and cooled down for 60 min at RT. The same reactions were made for the Nh2 fluorescents oligos of the chic-loop. At the end of the process, oligos were ready to be captured by the Tn5 protein.

B. Tn5 oligo preparation

As referred in the Picelli paper [122] the dsDNA oligo solution were previously captured by the Tn5 (Dimerization process) according to the following solution protocol:

-0.125 vol of 100mM equimolar pre-annealed oligo

-0.4 vol of 100% glycerol

-0.36 vol of Tn5 (preparation 1.16 mg/ml)

-0.12 vol 2X Dialysis Buffer (50 mM HEPES-NaOH ph 7.2, 1 mM EDTA, 0.5 M NaCl, 20% Glycerol)

Total volume of 10µL.

The reaction was incubated for 60 min at 37°C. At the end of this step, a dimer of tn5 loaded with oligo were obtained and ready to use for tagmentation.

C. Tagmentation reaction

The tagmentation reaction was made as the following:

- DNA Target (200 ng/µl of linearized plasmid used 3µL→600ng)
- 4 times more Tn5 from Tn5 oligo solution (2.4 µg)
- 5X Tagmentation Buffer (100 mM HEPES, 50 mM MgCl2, 40% PEG 3,500)
- ddH2O to 20 µL

Total volume of 20 µL.

Tubes were incubated in a thermocycler for 7min at 55°C.

We added 0.5µL of Proteinase K (Fisher Scientific, Hampton, NH, USA) and incubated for 7min at 55°C. The reaction was PCR purified with the PCR purification kit (Qiagen, Hilden, Germany). After the tagmentation process, the reaction could be amplified by PCR or directly ran on agarose gel for tagmentation validation.

D. PCR after tagmentation

After the tagmentation process, an aliquot of the reaction was used for the PCR reaction and mixed as the following:

Buffer Mix:

-0.2µl Taq Polymerase (Sigma-Aldrich, Saint Louis, MO, USA)

-1µl of 10X polymerase Buffer (Sigma-Aldrich, Saint Louis, MO, USA)

-0.6µL MgCl2 solution (Sigma-Aldrich, Saint Louis, MO, USA)

-1µl dNtps solution (Sigma-Aldrich, Saint Louis, MO, USA)

-ddH2O to 10µL

Reaction Mix:

-used 2.6µL of PCR mix

-1µL DNA-Tn5 Treated

-ddH2O to 9µL

Linearization step for 7min at 72°C

-add 1µL ss Tn5 MEA oligo V2 for amplification

-thermocycler cycles:

-3min at 94°C

-45s at 94°C
-30s at 55°C
-90s at 72°C
} 35 cycles

-10min at 72°C

At the end of this process of amplification, the tagmentation could be analyzed on an agarose gel. It is also possible to replace the regular oligos with Nh2 fluorescents oligos (Fisher Scientific, Hampton, NH, USA) or chic-loop oligo design (Fisher Scientific, Hampton, NH, USA).

E. Tn5 Streptavidin complex formation

Just before the tagmentation reaction, when the Tn5-Chic-loop complex was formed, the Tn5 dimer-chic-loop complex was mixed with the Streptavidin protein (Fisher Scientific, Hampton, NH, USA) as the following:

-15 µL Tn5 chic-loop solution (at 1mg/ml)

-5 µL of the Streptavidin protein (at 1mg/ml)

Total volume of 20µL.

The reaction was incubated for 30 min at RT. After this complexification process was done, the tagmentation reaction was processed as usual and a PCR was also made.

IV. Results and Discussion

1. Experimental results

A. Design and theorical cloning strategy

To be sure of the potential success of our design, we decided to take time to carefully validate each part of the design by a complete *in silico* study.

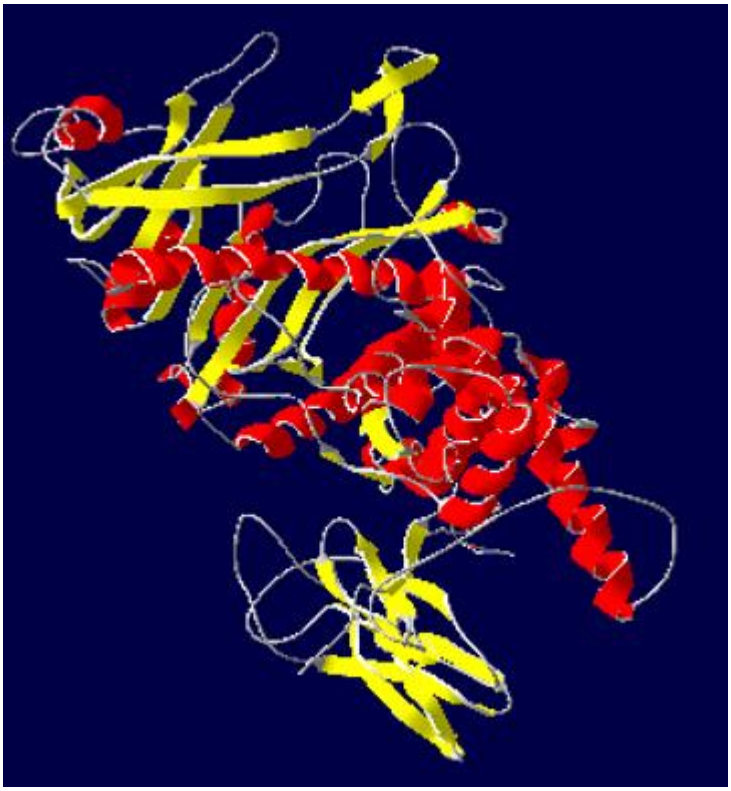
a. In silico studies for the two-fusion proteins

First of all, it made sense for us to use a biotinylated antibody because the technology is well established [154], and the fusion protein would be a Tn5/mSA couple. To give us more chances of success, we decided to make two different designs: Tn5-mSA and mSA-Tn5. To be sure that both proteins complete their own function properly without steric hindrance issues, we decided to link them with a long tail of repeated serine and glycine amino acids known for their flexibility and inertia (SGGGG x3). Moreover, a Tn5 control will be produced and used for further validation functional tests. All *in silico* analysis were made with the iTasser free software analysis technology and gave us a very good idea of the potential 3D structure of our new fusion proteins. Moreover, if the software can recognize each part of the fusion protein, it can give us a theorical score of functionalities, ligand prediction binding and GO terms analysis detection [155].

b. Tn5-mSA and mSA-Tn5 prediction

i. In silico Tn5-mSA

Figure 16. Predicted structure of Tn5-mSA by iTasser

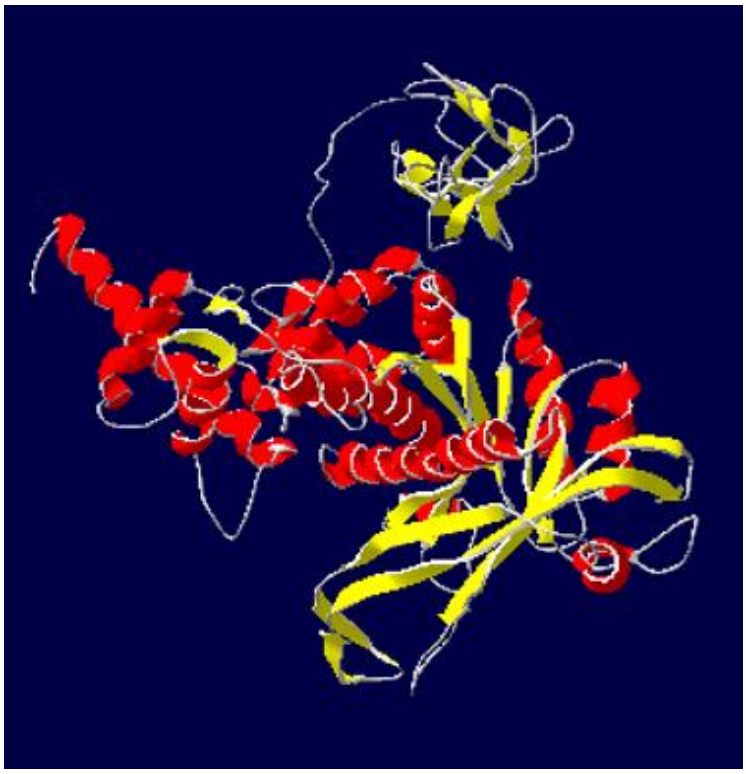


First, Fig 16 shows the result from the software iTasser for the TN5-mSA fusion protein which seems to have the best 3D architecture.

As we can see, the two-protein conformation was detected and the C score which corresponds to the confidence rate of the prediction was for this model C= -1.86 (basically, a model is very strong in terms of prediction if the C score is close to 2 or -2). Moreover, other analysis like the ligand prediction (Fig 18) shows that the Tn5 could bind to nucleic acids and the Mg^{2+} molecule, while the mSA could bind to Biotin, which means that this *in silico* form seems functional for both proteins.

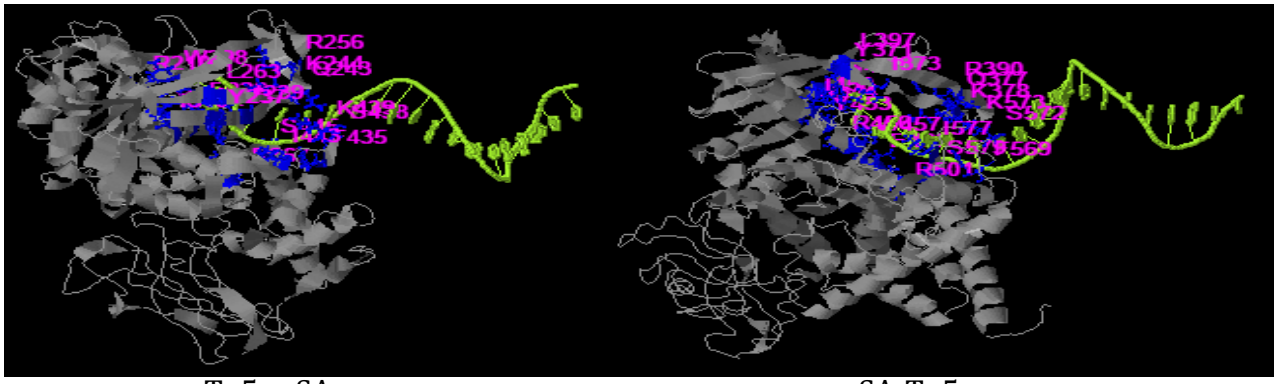
ii. In silico mSA-Tn5

Figure 17. Predicted structure of mSA-Tn5 by iTasser



Secondly, the results from iTasser for the mSA-Tn5 fusion proteins were at the same level of quality as the first one. The two fusion proteins conformation was detected and validated by the software with a C score of -1.84. Moreover, other analysis like the ligand prediction (Fig 18) shows that the Tn5 could bind to nucleic acids and Mg^{2+} molecules, and the mSA could bind to Biotin. It means that these two *in silico* conformations of both fusion proteins seem to respect their functional goals.

Figure 18. Predicted Binding site of Tn5-mSA and mSA-Tn5 by iTasser



Tn5-mSA

Rank	C-score	Cluster size	PDB Hit	Lig Name
1	0.53	15	4dm0A	Nuc.Acid
2	0.29	8	1muhA	MN
3	0.22	6	1mm8A	Nuc.Acid
4	0.11	4	1mm8A	MN
5	0.10	3	1muhA	MG

mSA-Tn5

Rank	C-score	Cluster size	PDB Hit	Lig Name
1	0.50	15	4dm0A	Nuc.Acid
2	0.49	9	1mm8A	MN
3	0.21	6	1mm8A	Nuc.Acid
4	0.10	4	1mm8A	MN
5	0.10	3	1muhA	MG

c. Theorical cloning design for the two-fusion protein plasmids: Tn5-mSA and mSA-Tn5 and the control

After making both *in silico* predictions, we designed our cloning strategy. As my mentor Ibrahim was used to say "The best cloning strategy is the less complicated design". To do so, we used the backbone of protein production plasmid NEB PTXB1, which presents many advantages in terms of protein purification. The cloning step can be summarized first, by inserting the Tn5 inside the plasmid and secondly the mSA sequence was added in both end sites (N-term or C-term) to generate the Tn5-mSA or mSA-Tn5 fusion proteins. Finally, a Tn5 alone sequence was also generated without any mSA (for more details see material and methods part).

B. Production and purification of the two isoforms Tn5/mSA fusions proteins complex

After preparing the three plasmids Tn5 control, Tn5-mSA and mSA-Tn5, each of these three proteins are produced regarding a specific production and purification protocol. Even if the protein production and purification is a well-established protocol, producing proteins is always a real challenge.

a. Theorical design protocol and justification of the production/purification strategy

As mentioned, the protein purification system is well established and a lot of different protocols and technologies were developed over the years. Nevertheless, even if protein purification has been developed a long time ago, getting a good protein quality and quantity requires a good experience.

i. Quick resume of the different classes of protein purification

Protein purification became useful due to the diversity of drug medications and some active principles which must be purified in that manner. It is also very diversified and a lot of different techniques exist. Without entering into too much details, we presented here a brief overview of the most popular technologies used.

A very well-established method is the chromatography. It exists a lot of different kinds: size exclusion, ion-exchange or by affinity. Most of them are designed in column especially for routine use [156]. They present a very good adaptabilities and are compatible for very different scales of production. From the 60's, it was always something researchers had to deal with, and today many continue to use it [157]. But even if this technique could be adapted for many kinds of protein parameters such as pH and buffers, it is quite restrictive and the scale is more adapted for large production [158]. This technique could be highly efficient and versatile, but sometimes it is difficult to set up especially when combined with very recent technologies like gas chromatography with mass spectrometry [159] or newest gradient HPLC [160].

Another form of purification, also very well established, is by tag protein purification, which is associated with chromatography. The principle advantage of this sort of purification is that the manner of purification is closely related to the production design [161]. Indeed, tagged proteins are needed, which means that during the cloning step, a histidine tag (of six histidine for the most common [162]) is added to the N-term or C-term of the protein target. And based on this tag, the purification is adapted, e.g. Ni-NTA column for the His tag. This manner of purification is very ligand dependent but more efficient [163]. Thus, due to this kind of isolation, only the purification itself is focused on the protein neo synthetized and does not require multiple additional steps of purification and avoid or at least limit the contamination problem. This methodology is also faster and very easy to set up on a compatible scale for research projects. Nevertheless, it exists a large panel of different systems of protein tagging and purification. Some proteins activity may not be compatible with the addition of a tag even if this one is very small and inert (especially due to problems of 3D folding and steric hindrance) [164]. This is based on these elements that we chose one technique of purification by protein tagging without a remaining tag on the proteins after the purification, to give the Tn5 a chance to make the dimerization process [165].

ii. Justification of the Neb IMPACT® strategy of purification

To give us the best chance to have at the end of the process a functional fusion protein, we decided to perform a protein purification process by tagging but with an auto-cleavable tag technology. Indeed, the NEB company provides a purification technique which can purify a protein almost in the native folding.

Thus, with the NEB IMPACT® strategy, it was possible to produce all of our fusion proteins without any tag at the end of them (Fig. 19). This is for this reason that we chose to clone all our interest sequences in the PTXB1 plasmid. Moreover, following the design and the protocol made in the Picelli paper, [122], we decided to put the Intein Tag at the end of all of our proteins. As mentioned in Figure 19, the technology itself can be summarized by an auto-cleavable Intein peptide obtained by an inducible cleavage in a specific buffer (4^0C, in DTT solution). It is just with the thiol mediated cleavage (N-S shift) that the Intein peptide, already bound into the Chitin bead column (Chitin binding domain attached to the Chitin beads), is cleaved. The target protein is finally released and just a S-CH_2-CH_2-SO_3^- molecule is present at the C-terminal part of the protein. This design of purification corresponds to the paper in the field [122] of the Tn5 purification and we were confident to have a functional protein at the end of this process.

Figure 19. Neb IMPACT® technology summary

Neb®, www.Neb.com.

b. Optimization and validation of the protocol of purification and conservation

After choosing our purification strategy, we decided to follow the Picelli paper and adapted this strategy for our own case.

i. Introduction and conceptual protocol

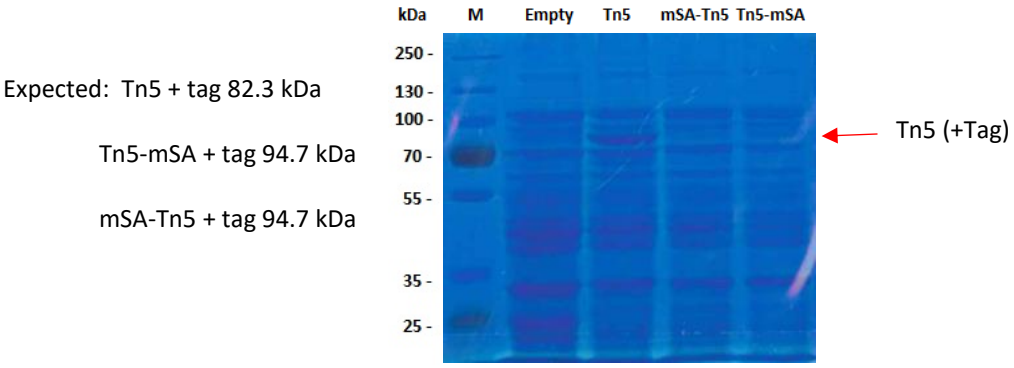
We produced our pTXB1-Tn5 as recommended in the paper [122]. We also added the monomeric streptavidin before or after the Tn5 with, as previously mentioned, a polylinker between both proteins: serine and glycine (SGGGG x3). As mentioned, all purification steps were used according to the paper, but we actually noticed for the cleavage step that the DTT action seemed to have a yield of maximum 30-40%. To increase the yield, we decided to use the MESNA molecule as recommended in the protocol of the IMPACT® manufacturer and we got a yield way much better (70%).

ii. Results and optimization

After producing our Tn5/mSA plasmids, we transduced them by heat chock to the C3013 competent cells from NEB (T7 express lys Y/I^q). These cells had the particularity to allow the production toxic proteins and difficult molecules. The first part of our production step was to verify the protein production in the whole protein lysate from the cells. Bacteria

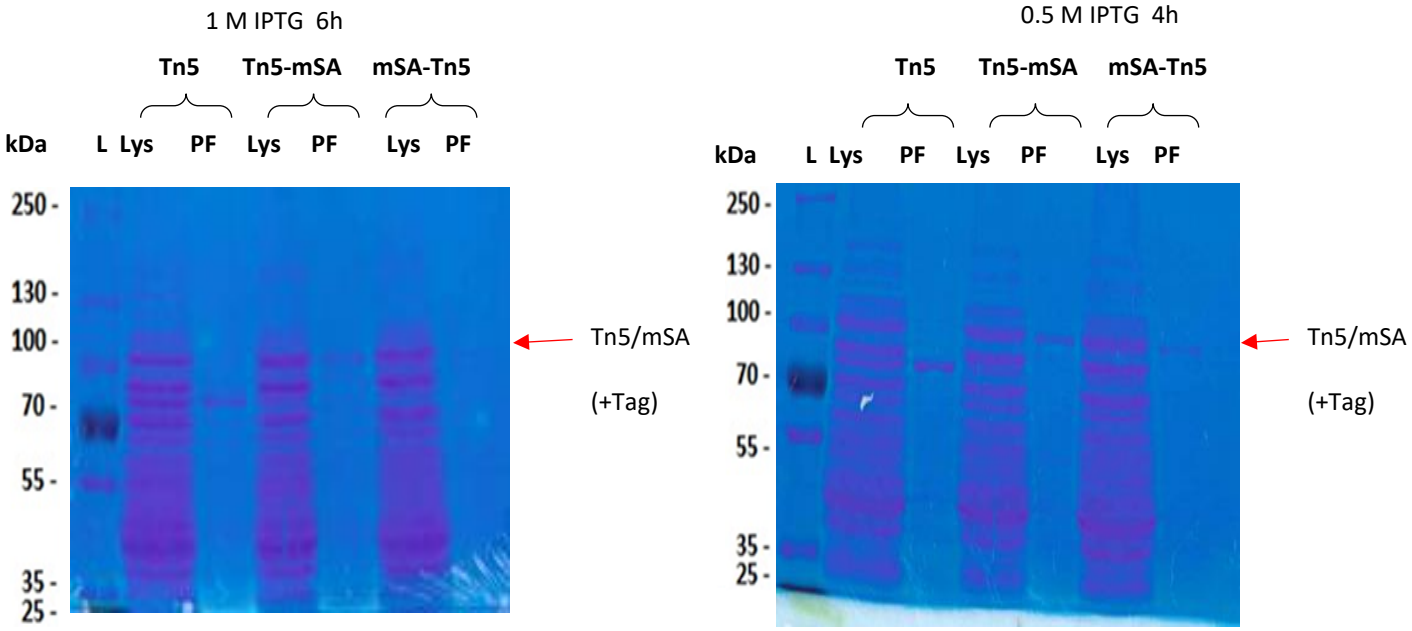
are grown during 4 hours before being induced by IPTG (regular induction of 1M). As shown on the SDS gel in figure 20, the expected bands were found especially for the Tn5 alone. For the Tn5-mSA and mSA-Tn5 bands, they are present but in very low quantity.

Figure 20. Tn5 production in C3013 Cells total lysate analysis



Next, we determined the best IPTG concentration for the induction of the bacteria. We tried to optimize the induction time as well (shown in the SDS gel Fig 21). As shown in the figure 21, we tested different IPTG concentrations and induction times and it was clear

Figure 21. Tn5 production optimization on [IPTG] and induction time



that 0.5M IPTG for 4hours was the best incubation time for all proteins production. Indeed, in the purification fraction (PF) the bands appear clearly.

The next step was to optimize the salt concentration in the washing buffer, after the binding of our proteins on the chitin beads columns. As shown in figure 22, for both proteins, 0.75M NaCl seems to be the best concentration even if for mSA-Tn5 some degradation appears.

Moreover, a test on the DTT efficiency was realized to determine the best cleavage time (Fig. 23).

Figure 22. Tn5 production optimization on [NaCl]

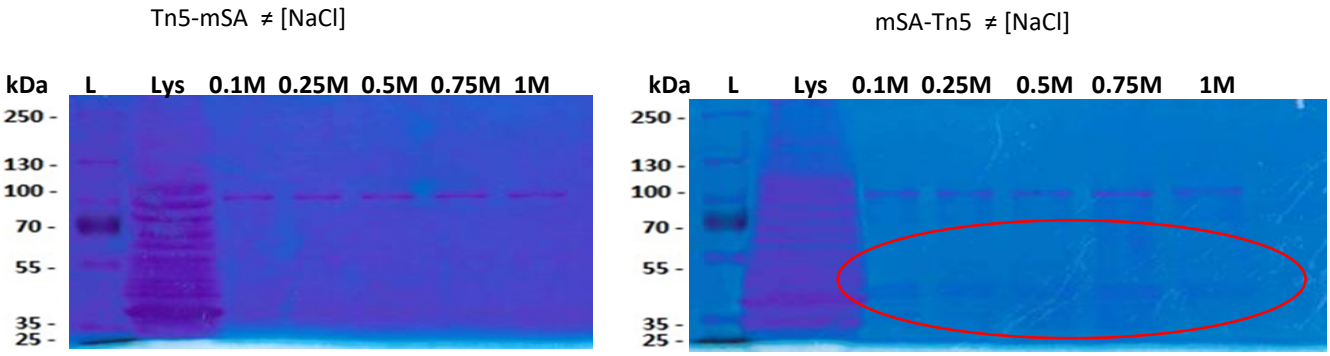
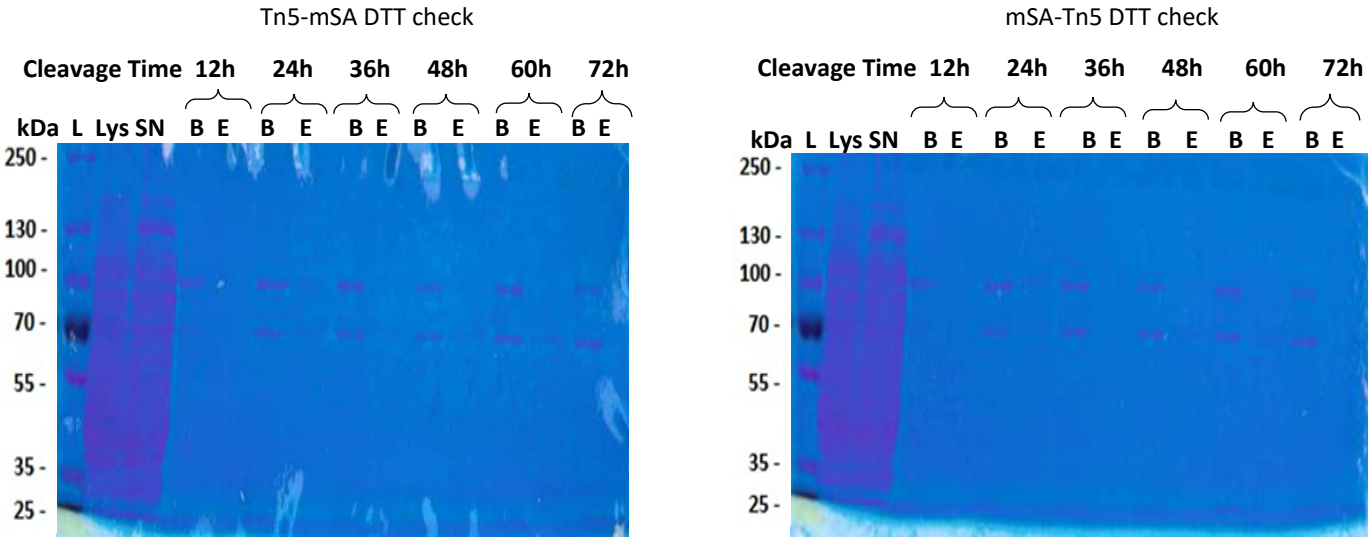


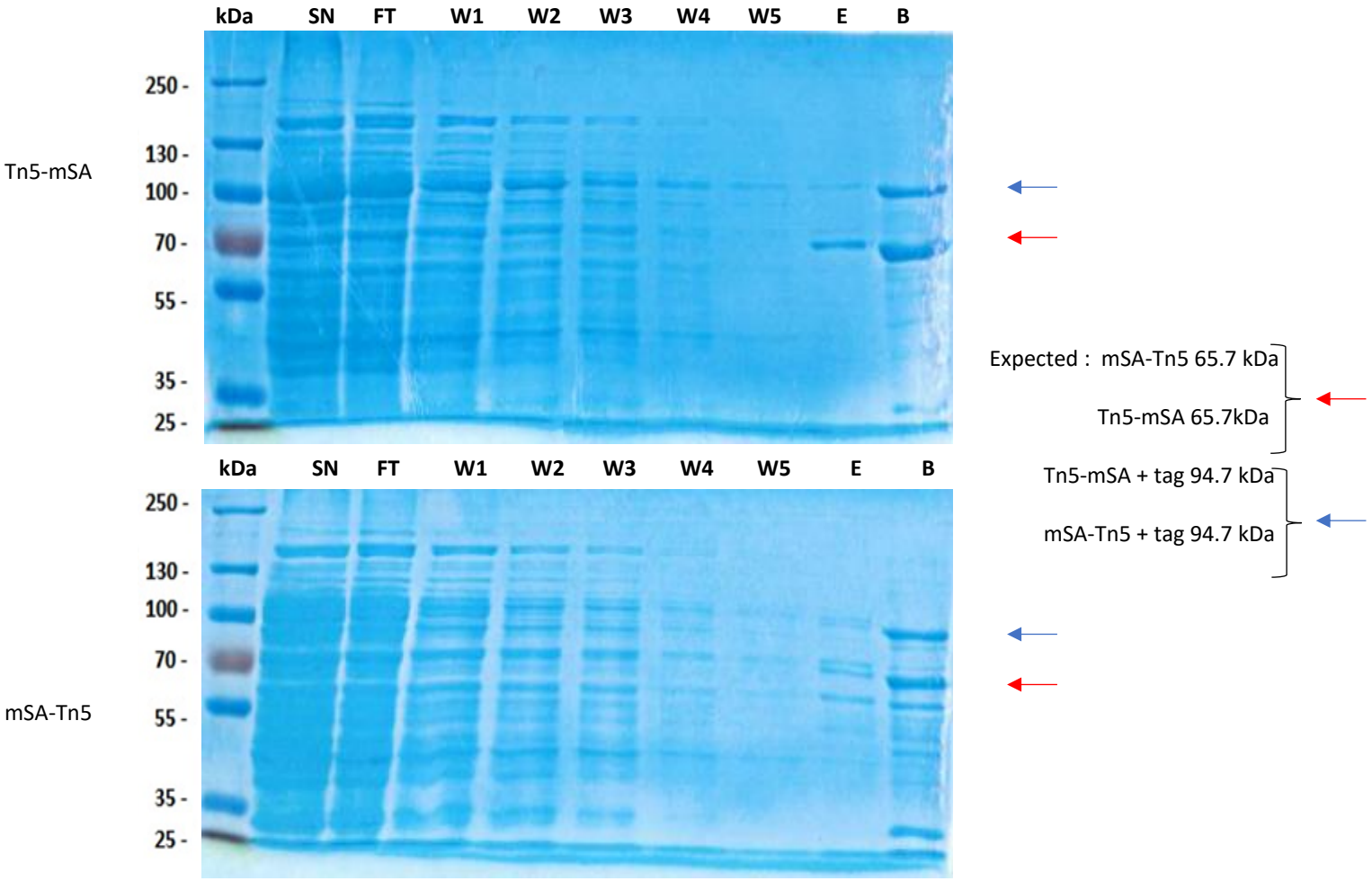
Figure 23. Tn5 production optimization on DTT cleavage time



As shown, for each cleavage time, an elution fraction (E) and a binding (B) fraction are presented. The binding fraction corresponds to the protein still on the column. The yield of purification was not as good as expected but we still can purify both isoforms.

With all these results, we decided to move on the complete purification by DTT process (Fig 24).

Figure 24. Gel of production of the Tn5/mSA DTT cleavage method



In figure 24, we have the whole production gel with the Supernatant (SN), the Flow though (FT), 5 successive washes (W), the Elution fraction (E) and the bound fraction (B). As seen, the binding fraction has still the vast majority of the protein production. Just few of them are eluted. Moreover, for the mSA-Tn5 isoform, it seems that the protein was degraded. For the Tn5 control alone, the DTT production was very fine even if the yield was not as good as expected (Fig 25).

In regards of these results, we decided to optimize the production process and move to the MESNA reagent for the cleavage, but we used it for the Tn5 control alone due to the reactivity test of our Tn5/mSA isoforms. For this one, the yield was way much better (Fig. 26).

Figure 25. Gel of production of the Tn5 DTT cleavage method

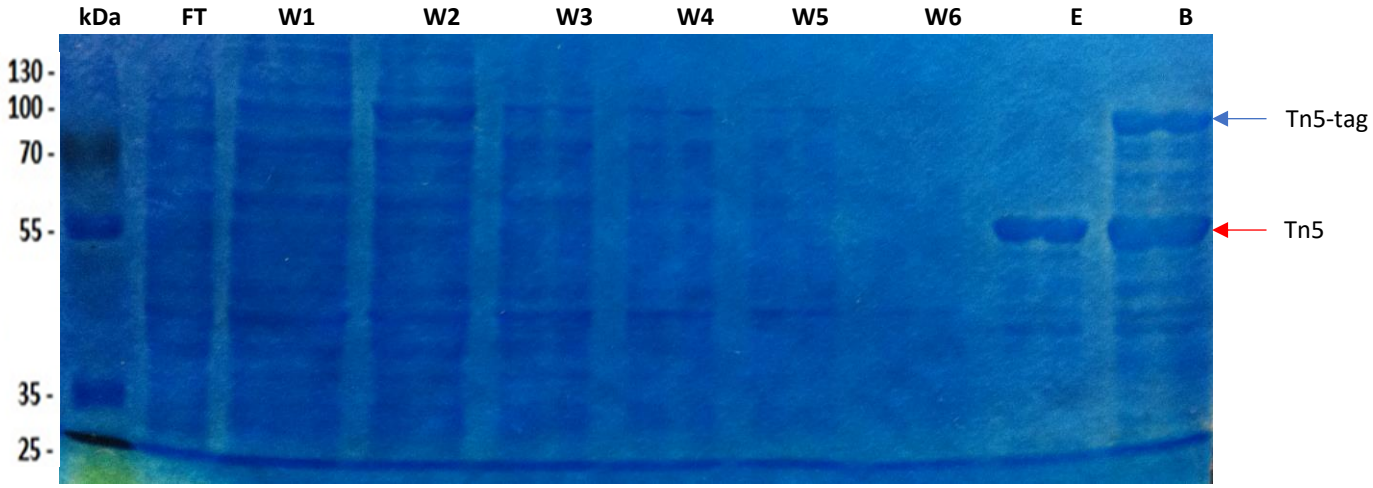
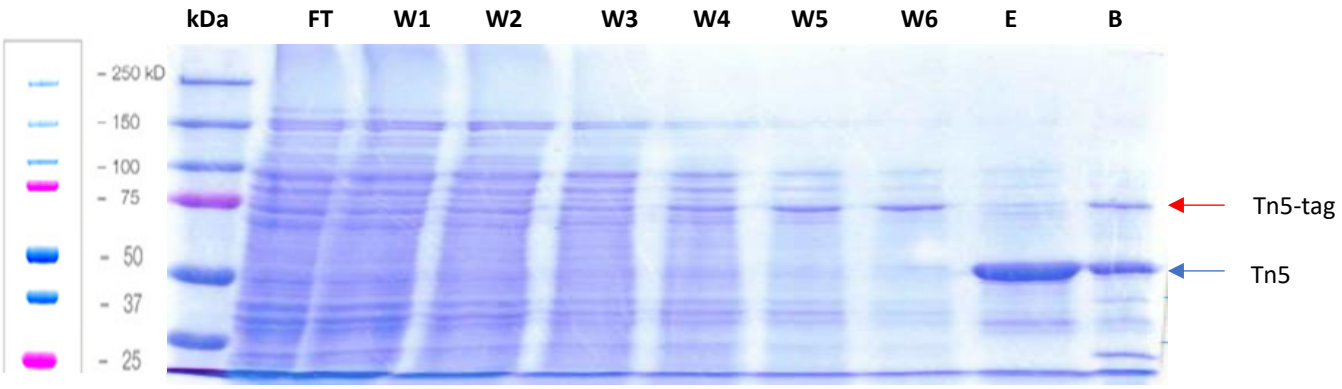


Figure 26. Gel of production of the Tn5 MESNA (200mM) cleavage method



C. Validation of the Tn5/mSA fusion protein

After producing with success all isoforms (our two fusion proteins plus the Tn5 control), we decided to test the activity of each protein. The Tn5 activity test was easy and just consisted in verifying its capability to tagment (degraded) a dsDNA plasmid. For the mSA, the check was not so complicated as well. We just needed to verify its capability to bind fluorescent biotin oligos.

a. Verification of Tn5 tagmentation by DNA and plasmids fragmentation

To verify the Tn5 tagmentation capability, we needed to know first if our two isoforms could react with fluorescent oligos (Alexa 488, same design as the Nextera® oligo) and form the synaptic complex.

i. Tn5 fusion and focus on the tagmentation reaction

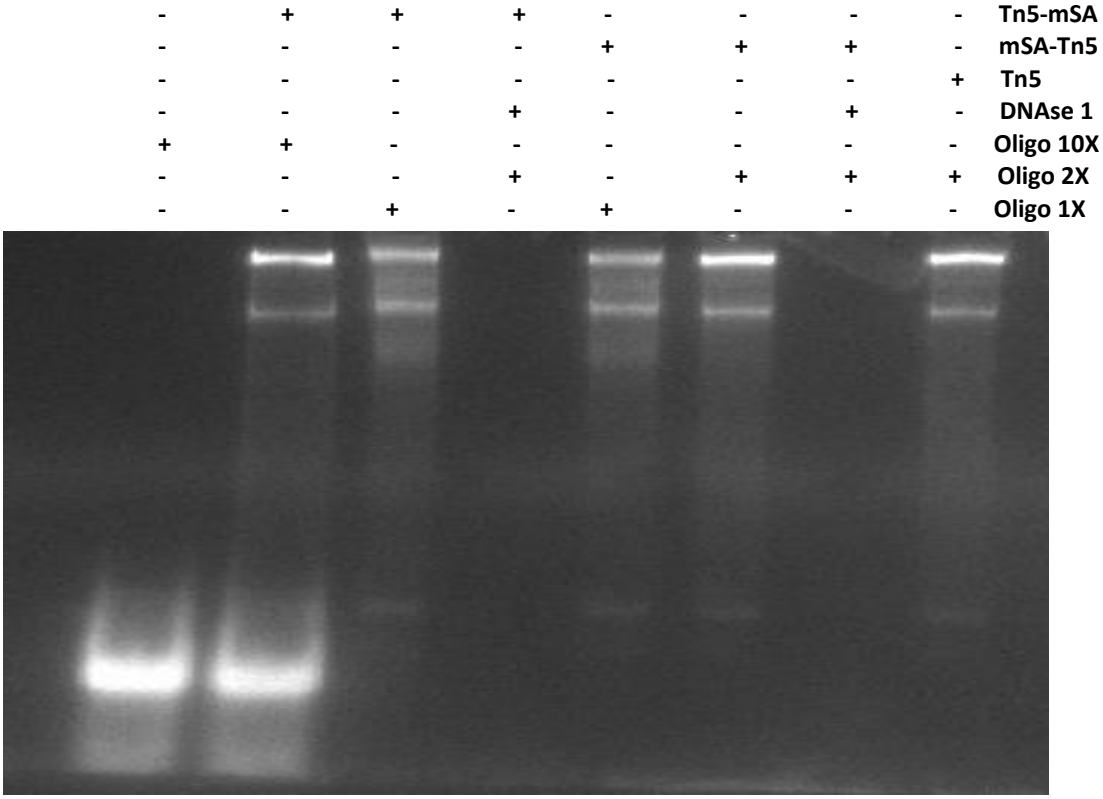
Indeed, as we already know, before performing its tagmentation function, the Tn5 transposase has to make its dimerization process in the shape of the synaptic complex. To verify that, we planned to use fluorescent oligos.

Synaptic complex formation

After making the dimerization reaction (50 mM HEPES-NaOH ph 7.2, 1 mM EDTA, 0.1 M NaCl, 10% Glycerol at 37^0C 1 hour) with the different types of isoforms and the Alexa488-oligos, we ran an Electrophoresis mobility shift assay (EMSA) on 8% polyacrylamide gel. In Figure 27,

we tested three proteins (two fusion and the Tn5 control) and for all of them, a complex was made with different concentrations of fluo-oligo. As we can see, the oligos bound to the Tn5/mSA the same way as the control which is proved by the retardation on the top of the gel. As expected, when we used DNAse 1 during the binding complex formation, no signal was detected, which proves that all different isoforms produced have the capability to assemble the synaptic complex.

Figure 27. Tn5/mSA synaptic complex formation verify by EMSA

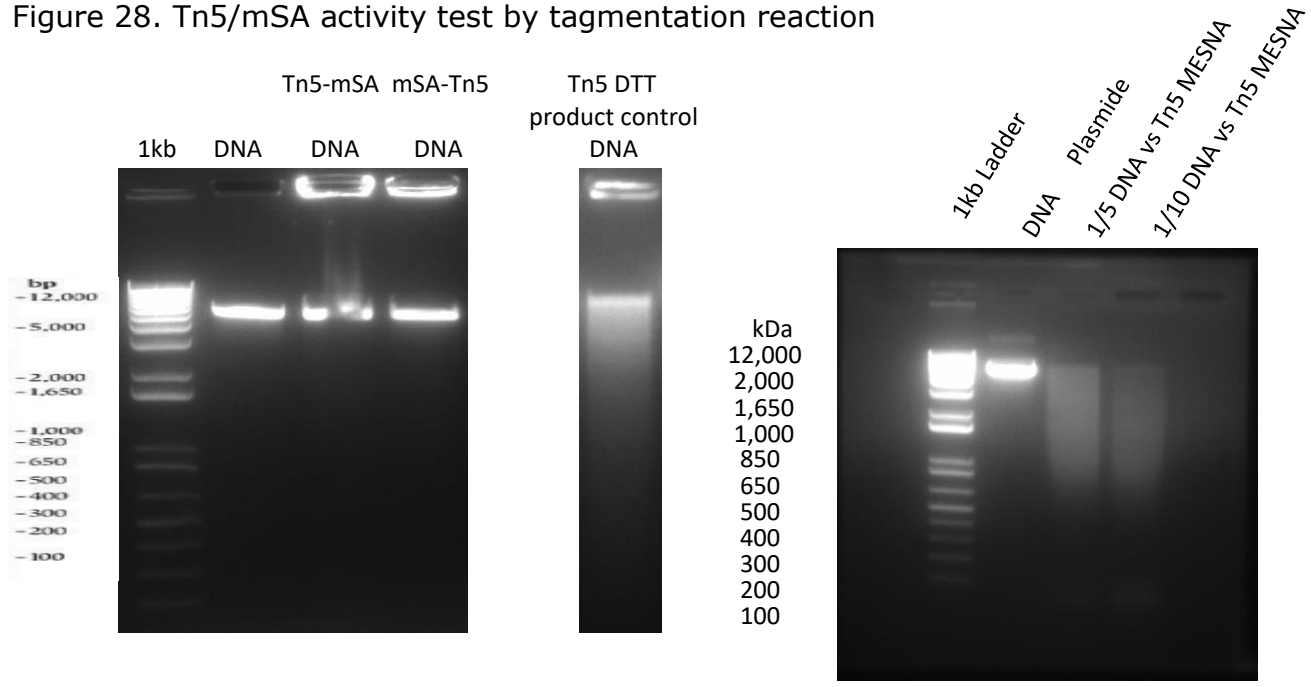


Tagmentation reaction

In a second step, we decided to test all of these synaptic complexes Tn5/mSA in a reactivity assay by testing their tagmentation capability. We let all different Tn5/mSA fusion proteins and our Tn5 control in a tagmentation buffer (5X 100 mM HEPES, 50 mM MgCl2, 40% PEG 3,500 during 7 min at 55^0C) with a dsDNA target: a plasmid. If the Tn5 tagment, the plasmid DNA band should appear as a smear on an agarose gel. To block the reaction, 10 µg of proteinase k was added per tube after the 7 min tagmentation reaction for another 7 min at 55^0C.

As found in the Figure 28, unfortunately, for both of our fusion proteins no degradation of the plasmid was detected on the gel. In fact, no smear was seen except for the Tn5 control (the best condition of DNA degradation was found with the Tn5 control the MESNA production).

Figure 28. Tn5/mSA activity test by tagmentation reaction



It seems clear that the addition of the monomeric streptavidin on the N-term or C-term of the Tn5 protein does not make an issue for the dimerization process but more for the tagmentation activity. A certain steric hindrance probably appears during the refolding of the protein just before the Tn5 meet the DNA target (before the strand transfer itself). It also seems that an interaction appears for the Tn5-mSA isoform which is suggested by the band on the top of the gel (above the DNA plasmid band). It means that a part of the DNA interacts with the protein, but no degradation was seen even if a hypothetical interaction appears, which proves that both fusion isoforms cannot integrate (cut and paste) oligos into

the targeted plasmid. Nevertheless, we continued the validation process and tried to focus on the monomeric streptavidin part.

b. Verification of mSA/biotin binding with EMSA Tn5-mSA/Biotin-Dna probe

Thus, we focused on the monomeric streptavidin side and designed an experiment to validate its activity and capability to bind the biotin molecule.

i. mSA binding reaction

As we previously mentioned, the monomeric streptavidin was developed to reduce the whole steric hinderance of the streptavidin protein by the isolation of just one biotin binding site and enhance its binding capability [130]. The reaction biotin/streptavidin is known as one of the strongest in biology, well used in different manners notably for purification [166].

ii. Validation of the two-fusion protein focus on the mSA function

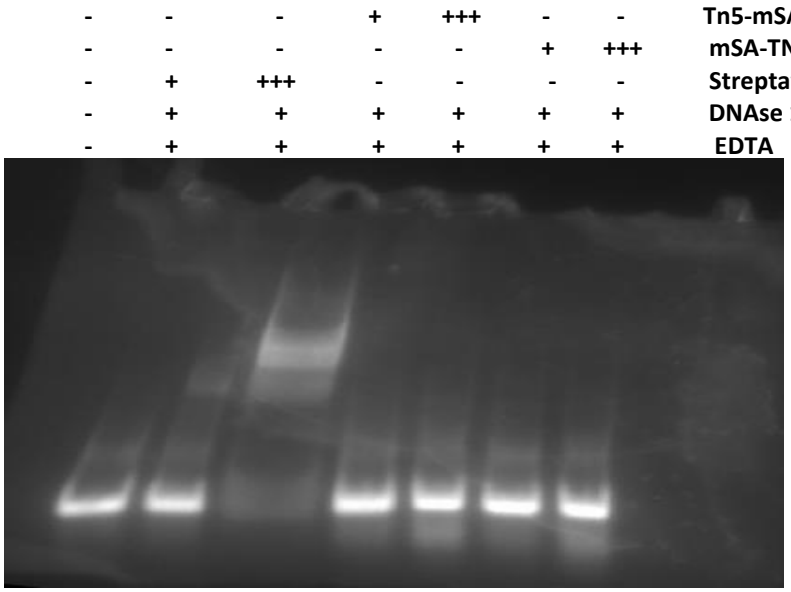
Introduction and conceptual propose

To know if the mSA works as expected, we just need to test it with the biotin molecule. To do so, we decided to proceed as previously, meaning that we designed the biotinylated fluorescent oligo (Alexa 488) and tested it in an EMSA. If the two isoforms Tn5/mSA can bind the biotin, a retard on the gel should appear. As a positive control, we decided to use the streptavidin molecule directly bought from commercial supplier.

Results and optimization

As we can see in the Figure 29, multiple experiments were made. Each protein (Tn5-mSA, mSA-Tn5 and streptavidin control) were used in two different concentrations with Biotin single stranded (ss) oligos-Alexa488. To be sure to not have any dsDNA contamination, the DNAse1 was added. To block the Tn5 activity, we used the cation chelator EDTA, which can trap the Mg2+. All reactions were made at room temperature for 1 hour and a 5% polyacrylamide gel was run. As seen, unfortunately for us, just the streptavidin positive control showed a retard on the oligo migration, proving its binding capacity. For both produced isoforms nothing happened.

Figure 29. mSA activity for both Tn5/mSA isoforms test by EMSA gel



It became clear that the fusion Tn5/mSA creates a steric hinderance and inhibits both protein activities. To find another way to complete our goal, we came back to the Tn5 control

because we know it works perfectly. From there, we moved to chemical reactions to bind it to the streptavidin or the antibody.

c. Direct binding by chemical reaction on the Tn5 protein itself

A lot of different binding chemicals are available to modify a protein. The most popular one is the NHS binding which corresponds to a binding on an amine group. Also, a lot of amine NH_2 molecules are usually available (on polar amino acids) on proteins and the Tn5 do not escape this rule. This kind of technique is well established and still used today [167], but these amines are present in almost all the Tn5 sequences and binding them may not be compatible with its activity process. Moreover, and luckily for us, two cysteine (amino acid which has the capability to make S-S bonds) are present on our Tn5 protein as reported in the Annexes figure 1 and 2. One is present inside the core of the protein which is inaccessible [168] (Cys 187) and a second one is very close to the C-terminal part of the Tn5 (Cys 402), very accessible [169]. Based on that, we decided to target the second one with a specific chemical reaction like the maleimide complexification.

i. Maleimide complexification test

We decided to go straight forward and tryied the reaction which seems to be the best regarding the feasibility and potential preservation of the Tn5 capabilities.

The maleimide reaction

The maleimide reaction was developed in the 60's [170] and is still very used in Science and Chemistry today [171]. The maleimide itself is a chemical compound $H_2C_2(CO)_2NH$ and the reaction corresponds to a Diels-Alder reaction on available thiol groups.

In our case, to an available cysteine on the Tn5 C-Term sequence (Fig.30) we decided to mimic this reaction done by many groups and tried on the Tn5. The most important thing for this kind of chemical reaction is the buffer control and the pH stability between 6.5-7.5 to get the reaction done.

Figure 30. Maleimide Diels-Alder reaction on a Thiols group

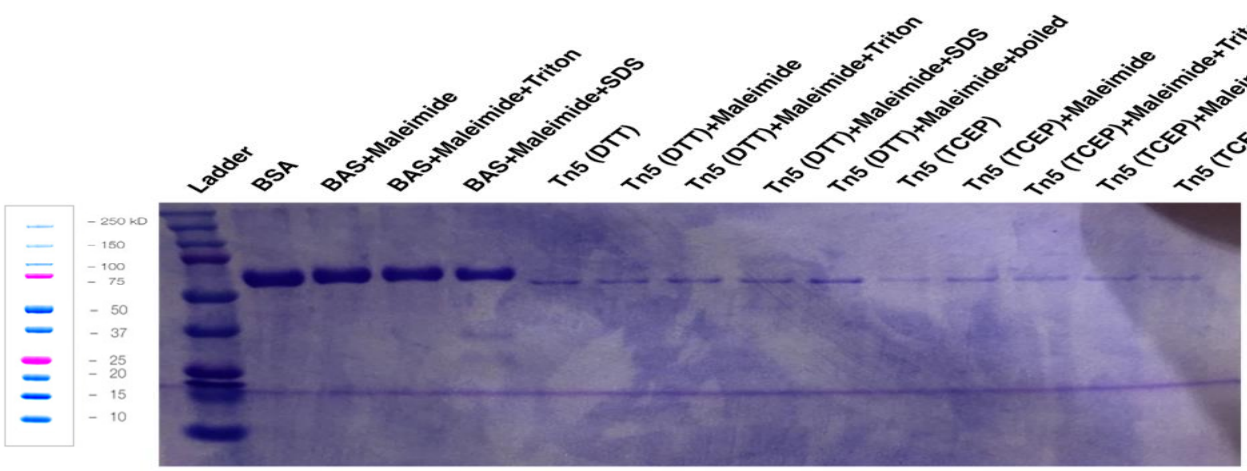
ThermoFisher Scientific®, www.ThermoFisher.com.

Results and optimization

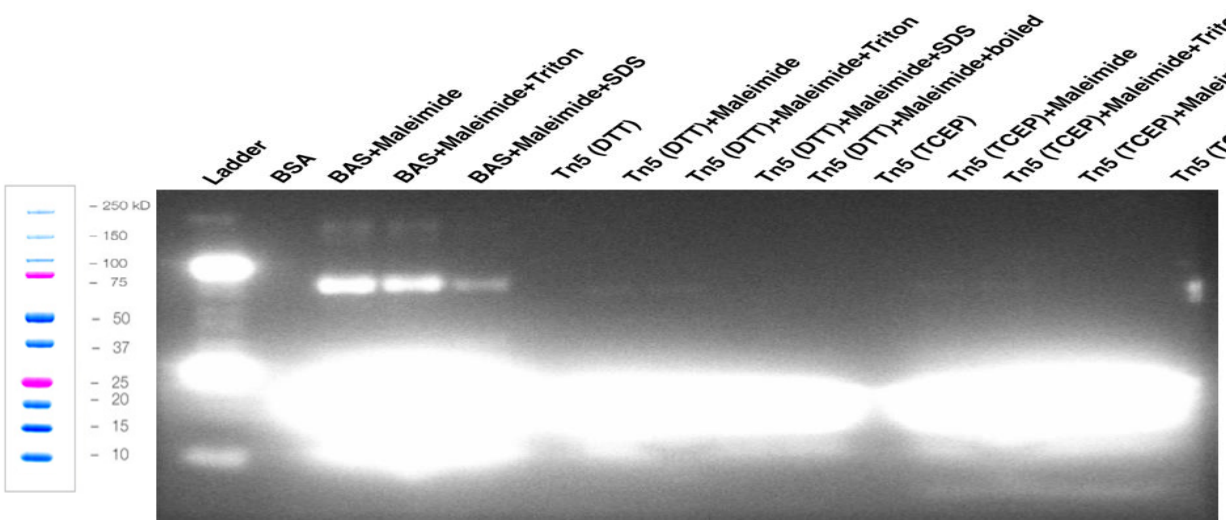
First, we made the reaction on our home-made produced Tn5 just with the maleimide compound itself complexed and an Alexa-488 to identify the reaction. We made the reaction as recommended by the field [160] and the manufacturer where we bought our maleimide-alexa488 (ThermoFisher®). Moreover, we kept the reaction in an acid buffer with a pH of around 6.8 (100mM sodium phosphate, 5-10mM EDTA, pH 7.2) which becomes very acidic with the addition of the maleimide. The reaction is made with different ratios of Maleimide-

Alexa488 vs Tn5 and was let for 90 min at 37^{0}C. BSA (Bovine serum albumin) was used as a positive control and different productions of Tn5 were used, traditional one DTT and some

Figure 31. Maleimide-Alexa488 Diels-Alder reaction on the Tn5



SDS PAGE



Signal Alexa488

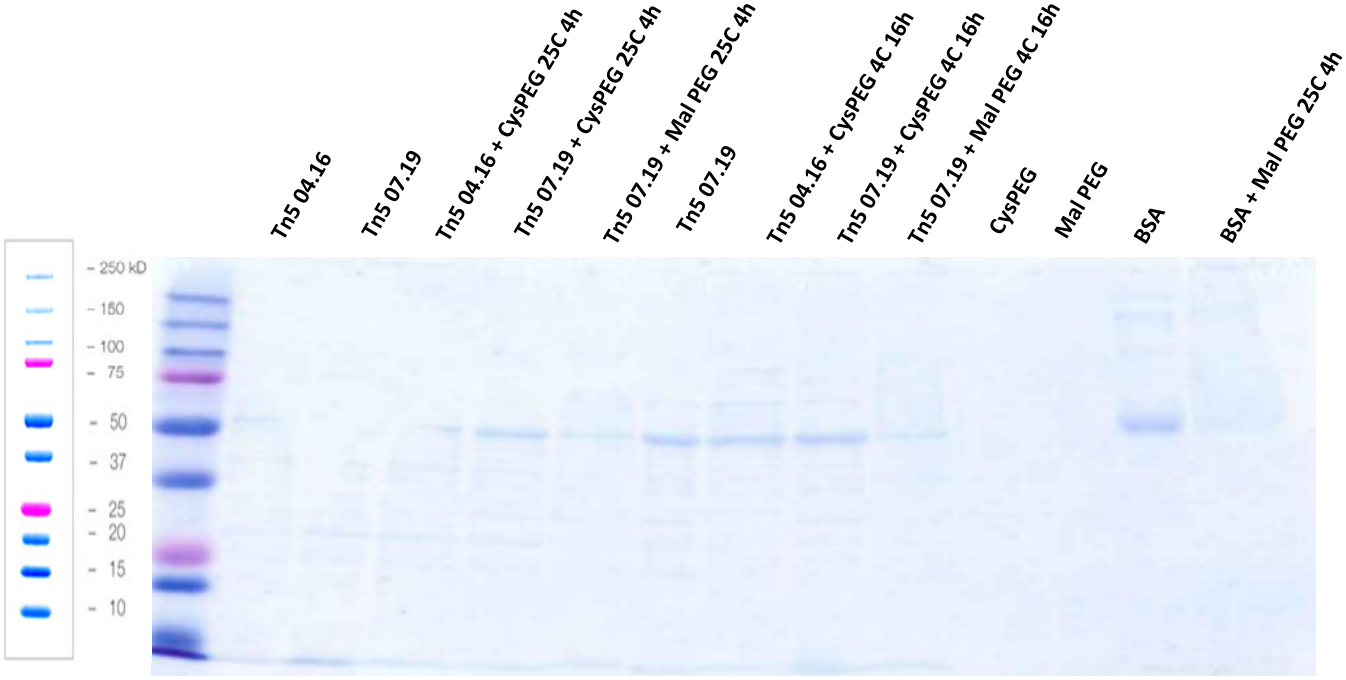
other productions conserved with TCEP buffer (corresponding to the DTT Tn5 production with additional steps of buffer exchange DTT to TCEP). Some other reagents were used to help the reaction (to linearize the Tn5 protein: Triton, SDS or Boiling).

As we can note in the Figure 31, both proteins BSA and Tn5 can be found on the SDS migration gel and when we analyze the same gel under the UV, we clearly see at the bottom the unbound maleimide-Alexa488 compound. Moreover, on top of it, the BSA Maleimide

complexification reaction appears as a band, which proves that the maleimide can react very well with BSA, with or without the help of triton (For the SDS binding is reduce). For the Tn5 samples, a very weak signal can be found but it corresponds by proportionality to the amount of proteins in the SDS band. Like the BSA, the maleimide can react with the Tn5 with or without reagent like triton and TCEP preparation (which as no significant difference with the DTT one). Based on these evidences, we moved forward and planned to attach more complexed molecules to our Tn5.

ii. The Mal-PEG, Cys-PEG and NHS-PEG complexification

Figure 32. Multiple binding reactions with Mal-PEG, Cys-PEG and NHS-PEG on the Tn5



Signal Alexa488

We reproduced the same kind of experiment, but we attached on the maleimide compound a polyethylene glycol (PEG) and the Alexa488. Because we know that we need a

flexible linker between the Tn5 and the antibody, we chose the 2K PEG which is commercially available.

In order to get better results, we decided to make multiple reactions with: Cys-PEG-Alexa488 by creating S-S bond with the cysteine available on the Tn5 and the NHS-PEG-Alexa488 (reaction previously develop). Different incubation times and temperatures were tested as well and the BSA molecule was used as a control just for the most promising binding: the Mal-PEG-Alexa488.

As presented in the figure 32, the same kind of results as before was obtained. Thus, an EMSA has been used on a SDS gel migration and on the Coomassie blue signal, all the proteins could be found (except for the negative control Mal-PEG, Cys-PEG alone). If we switch on the UV signal, all PEGs were found at the bottom of the gel and if we look at the top, unfortunately for us, just a retard signal was found for the positive control BSA molecule with the Mal-PEG-Alexa488 and nothing for the other reactions with the Tn5. Based on these results, we tried many other kinds of reactions, part of them shown in the Annexes figure 3 and figure 4. In these results, we managed a lot of different parameters to improve the maleimide-PEG reaction on the Tn5. As shown, we successfully increased the binding of Maleimide-alexa488 alone with the Tn5, but nothing showed up for the Mal-PEG-Alexa488 with the Tn5. Thus, in our despair, none of them worked and we admitted that this kind of chemistry binding was too hard to manage and uneconomically valuable for our propose. Nevertheless, we refused to stop there and tried to find another way to complete our goal, the one which finally gave us success: the direct oligo binding.

D. New design and validation strategy based on oligo customization

After this ascertainment, we decided to capitalize on what we discovered so far. It became clear for us that due to the particular multiple folding steps that the Tn5 molecule undergoes during its reactivity process (dimerization to generate the synaptic complex, strike on dsDNA and strand transfer), touching the Tn5 protein itself created too much steric hindrance and killed its reactivity. But, starting on the initial postulate that the Tn5 cannot be modified in any manner, another kind of binding exists. Indeed, the Tn5 is used as a sequencing tool in the Nextera® Kit proposed by the company Ilumina®. In this technology, the Tn5 is dimerized with well-designed oligos and adapted for the sequencing library generation. We used these same oligos for our project and generated our sequencing tools but why not modifying these oligos to give them a double role: inserted by the Tn5 during the tagmentation process and create a link between the Tn5 and the antibody itself.

a. Oligo customization design and strategy

Before moving forward with a successful design, we needed to better understand the Tn5 tagmentation with the actual oligos. Moreover, after this step, we played with different kinds of designs and proceeded step by step to isolate the most promising one and adapt it for our purpose.

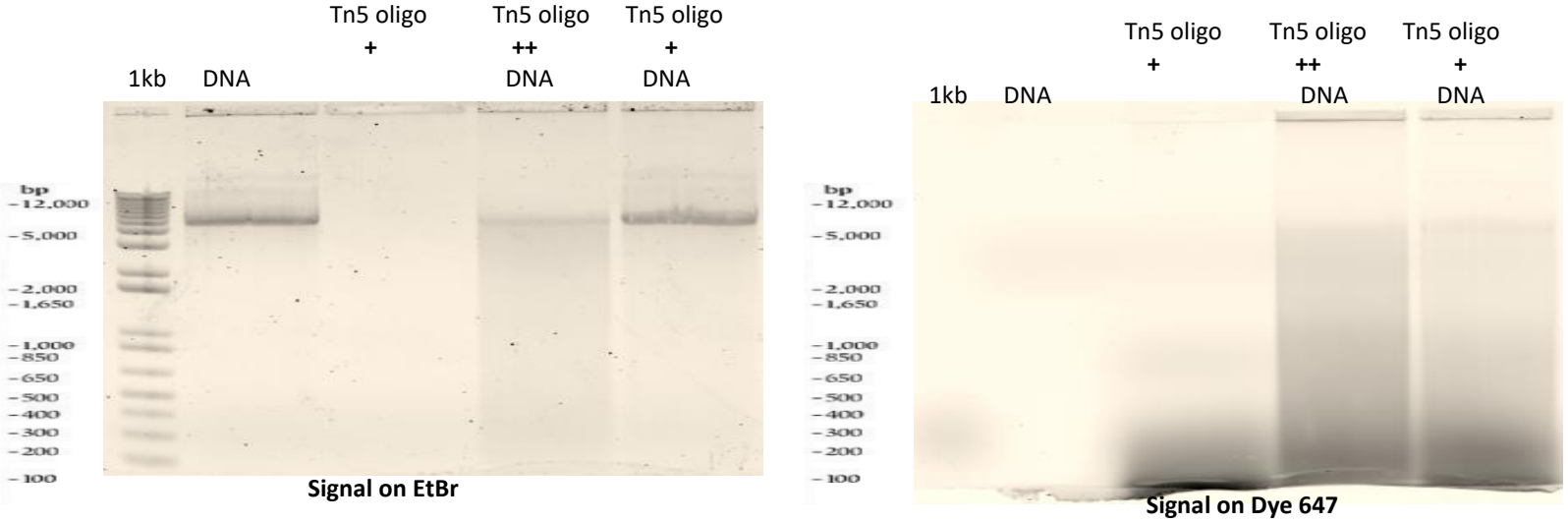
i. Return to the Tn5 production and more careful validation

As previously mentioned, we first performed different tests of our homemade Tn5 production to better understand basic parameters linked to the oligos and which could impact the tagmentation reaction.

Tagmentation test on regular oligo design

As we already know, we used oligos like the one from the Nextera® kit and described in the Picelli paper [122]. We tried to redo the tagmentation validation reaction but analyzing on different kind of experiments and in different conditions of reaction.

Figure 33. Tn5 Tagmentation assay with fluoresent Alexa647-oligo

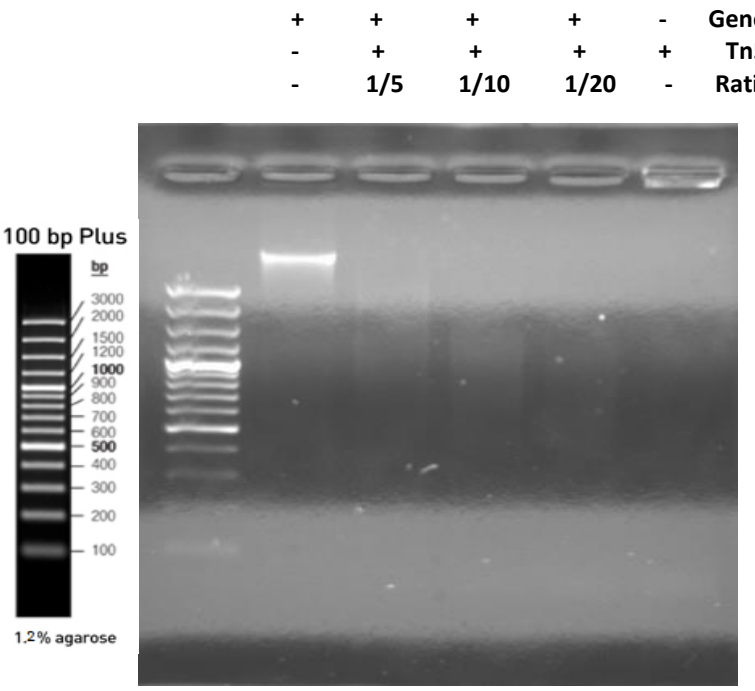


First, we decided to better follow the tagmentation reaction itself by adding an Alexa647 dye to the regular Nextera® oligos. We ran our reaction as usual, 55^0C, 7min targeted on a DNA plasmid and ran an agarose gel to check the smear of DNA fragmentation made by the Tn5. As expected, in Figure 33, the signal on the same gel is definitely better for the oligo Alexa647 compared to the EtBr. The smear is clearly detectable and with the

increase of Tn5 quantity vs the DNA target, the band of DNA is almost completely digested and fragmented. The oligo signal can be found at the bottom of the Alexa647 gel as well.

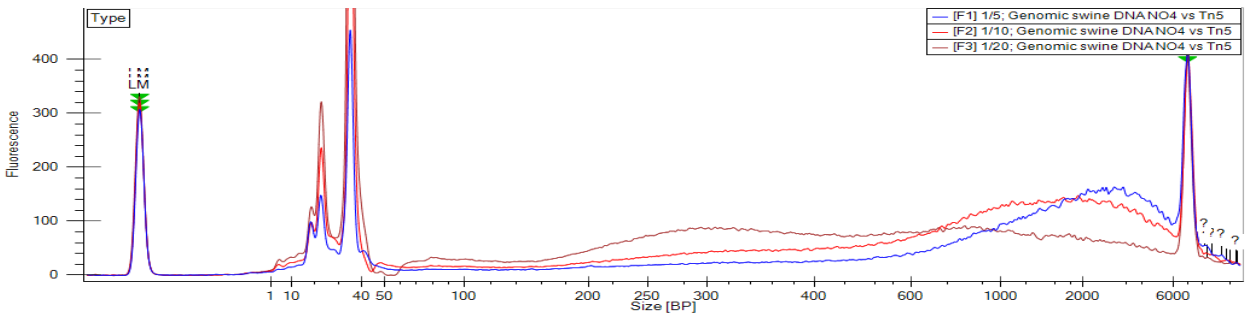
Secondly, we decided to know more about this relation, especially the ratio Tn5 vs DNA target, and made different kinds of experiments to check it. As shown in the experiment figure 34, we fixed the genomic DNA target quantity and increased the Tn5 quantity. As expected, with more Tn5 we obtained more tagmentation, indicated by the reduction of the size of the smear (same results on the Annexes Fig.5).

Figure 34. Tn5 vs DNA with fixed Tn5 quantity and increased genomic DNA quantity



To really be sure of the linear relationship between the Tn5 quantity and its DNA target, we checked a fraction of the same samples by bioanalyzer and saw the fragmentation efficiency. Indeed, in figure 35, we can see the three DNA fragment populations and appreciate that more the Tn5 is concentrated, shorter the fragments are. We continued our

Figure 35. Tn5 vs DNA with fixed Tn5 quantity and increased genomic DNA quantity Bioanalyzer results

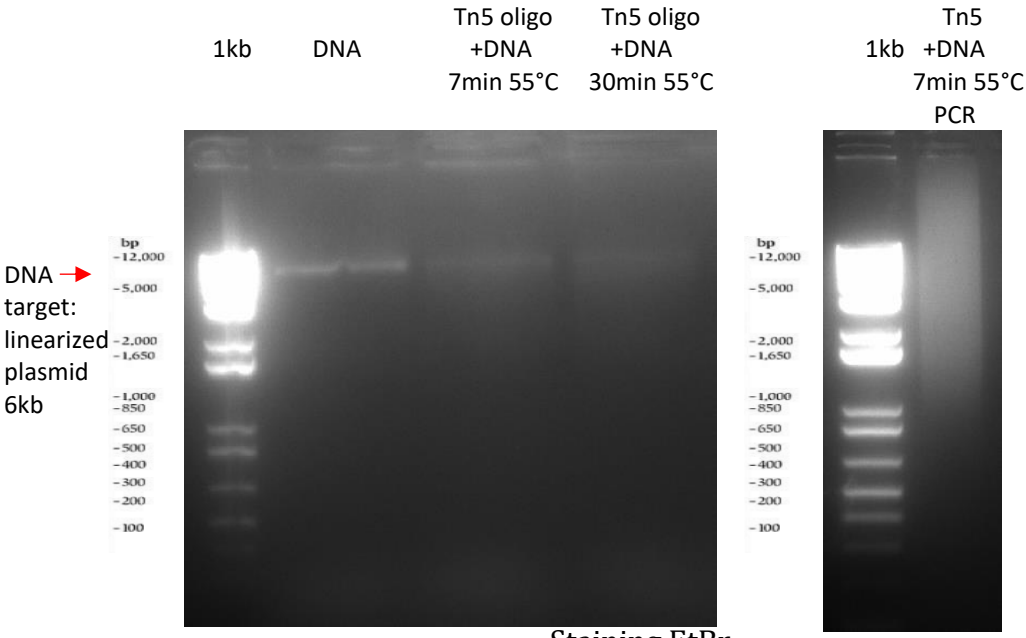


series of experiments and tried to manage the next step of the library protocol: the PCR amplification of the tagmented DNA.

`After the Tagmentation, the PCR amplification step`

We tried to follow the protocol developed by Illumina® and continued to use the same regular oligos (MeRev, MeA and MeB Illumina® catalogue: FC-121-1030 and FC-121-1031). After the tagmentation occured, we took the same samples and amplified the fragments with a second couple of primers flanked with the barcode for the sequencing (illumina® FC-121-1012 + FC-121-1011 primers). As shown in figure 36 and figure 6 Annexes (same experiment plus release via proteinase k or SDS and PCR primers complexed with Alexa 647 for a better view), the PCR amplified fragments generated a smear easily detectable on the gel. There is right now no doubt on the Tn5 action and tagmentation illustrated with both mechanisms: the fragmentation and the oligo insertion (proved by the specific PCR amplification). Based on these facts, we decided to work on the oligo binding efficiency and play with different designs to adapt them to our purpose.

Figure 36. Tn5 vs DNA followed by PCR amplification



ii. Verification of tagmentation on different designs

Figure 37. Transposase sedimented to beads or well wall.

Agilent Technologies Inc. US Patent EP2527438 A1 2012

.

Beginning with the design made for the purpose of the ATAC-seq [75], we decided to develop new kinds of oligos. We kept the Mosaic sequence (Me), essential for the well function of the Tn5 (synaptic complex formation) but decided to add more bases based on the contiguous A or B sequences, the reverse oligo Me Rev will not change depending on the design.

Introduction and conceptual propose

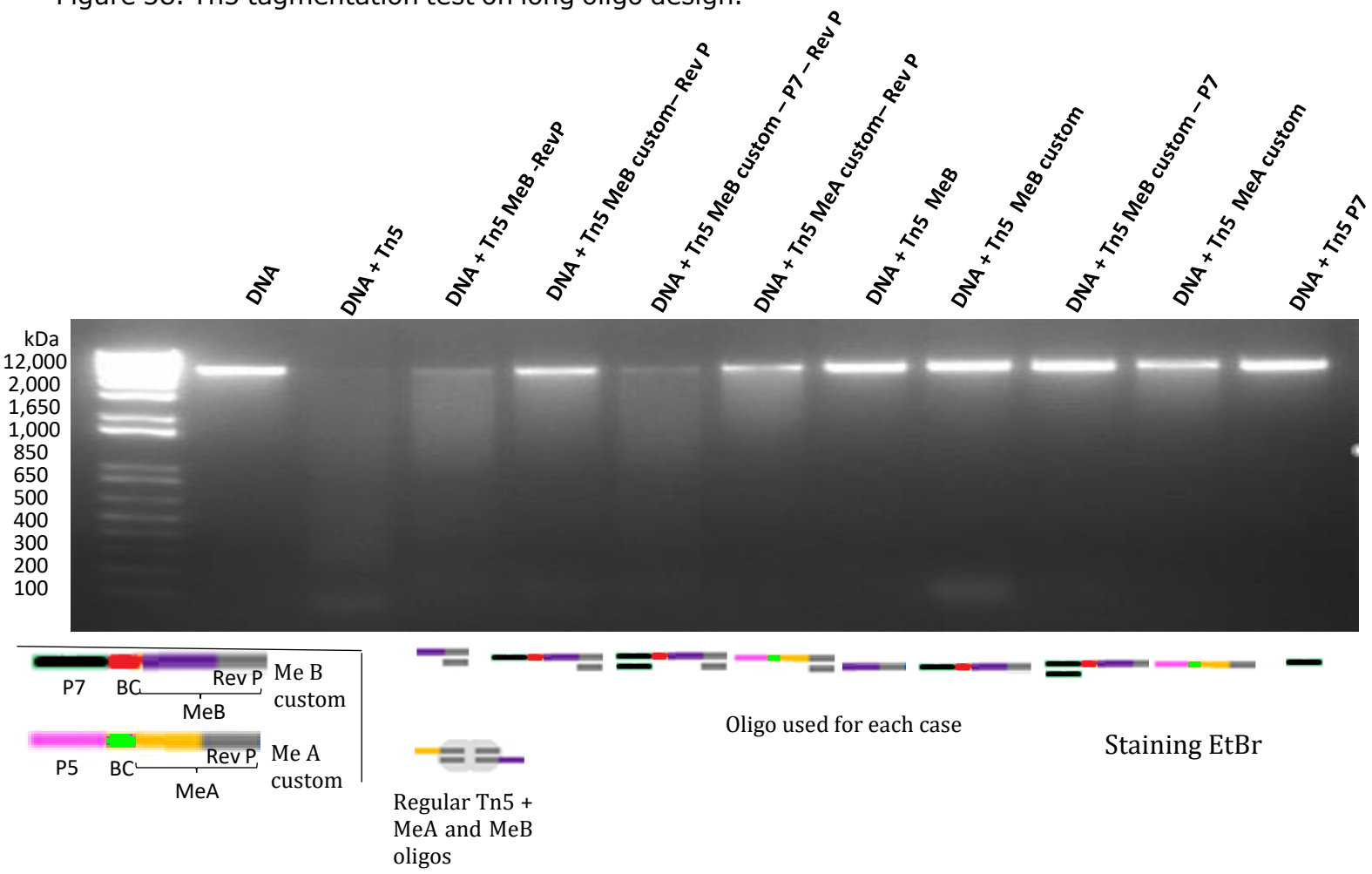
Any sequence and design are possible, different research groups focused on the design of bisulfite oligo for sequencing library generation [86], [172, 173], but we preferred to move from the regular design and tried to add different kinds of sequences. As presented in figure 37, (Agilent® patent), the Tn5 itself is sedimented in the well

through the oligo designed to adhere to the wall of the well. This design should give a more sensitive technique, increasing the tagmentation and the PCR amplification fragments. We decided to play around these ideas and made a succession of TTTTTTTT bases which is a flexible and inert design. Then we proposed long oligos “all in one” with the detection barcode sequencing primers. Moreover, we played on the annealing type of each design and saw what was possible for the Tn5 to use during tagmentation.

Results and optimization

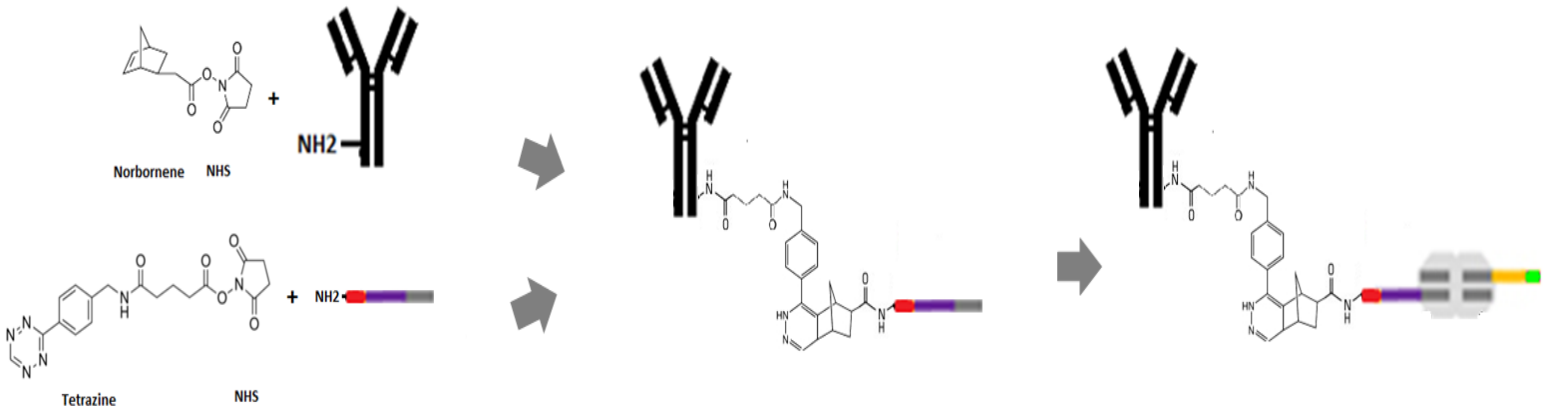
First, we modified the regular oligos MeA and MeB by adding part of amplified pieces from the PCR. In figure 38, the Barcode (BC and the annealing flanking part (P7 or P5) was added.

Figure 38. Tn5 tagmentation test on long oligo design.



Different kinds of annealing possibilities tested are all summarized in figure 38. Thus, after annealing each oligo (95^{0}C 5min and room temp for 60 min), they were added to the Tn5 to create the synaptic complex (37^{0}C for 60 min). This is after all this process that the tagmentation was tested (7 min at 55^{0}C + reaction block with 10µg proteinase k 7min at 55^{0}C) and an agarose gel was run (fig 38). As seen, the different designs work differently compared to the negative control (all oligos which are not a RevP part (Mosaic sequence double stranded) do not tagment as expected). Nevertheless, one condition seems to be very promising: MeB custom P7 which works very well. We also saw that the MeB oligo seems more available for custom design due to probably a more inert hairpin complex formation. Based on these results, we moved to our other choice of design, the poly TTTTT tail with an NH_2 at the end and which could give us the NHS reaction (previously seen) linked with the antibody (through the Norbornene/tetrazine molecule chemistry) as presented in figure 39.

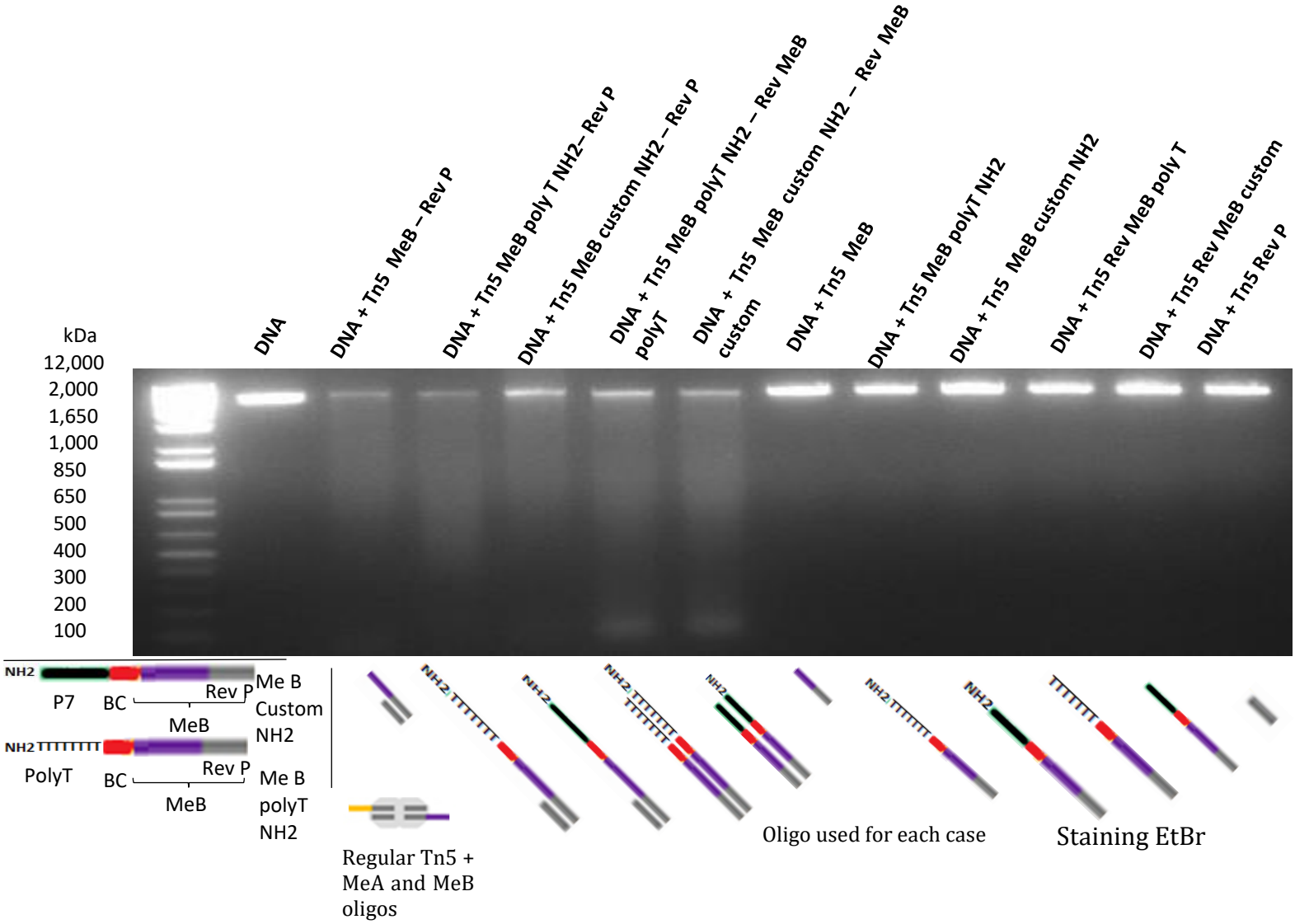
Figure 39. NH_2 oligo/antibody binding design



As presented in the figure 40, the same kind of experiment was made, but this time focused on the MeB oligo custom design and tested with a poly T tails plus NH_2 (same protocol as above was used).

As we can see, all the negative controls (same as above) still remain untagmented and the design tested seems very functional and gave us the same kind of degradation smear result compared to the positive control (Tn5 +MeB-RevP). So, the design MeB polyT NH_2-RevP

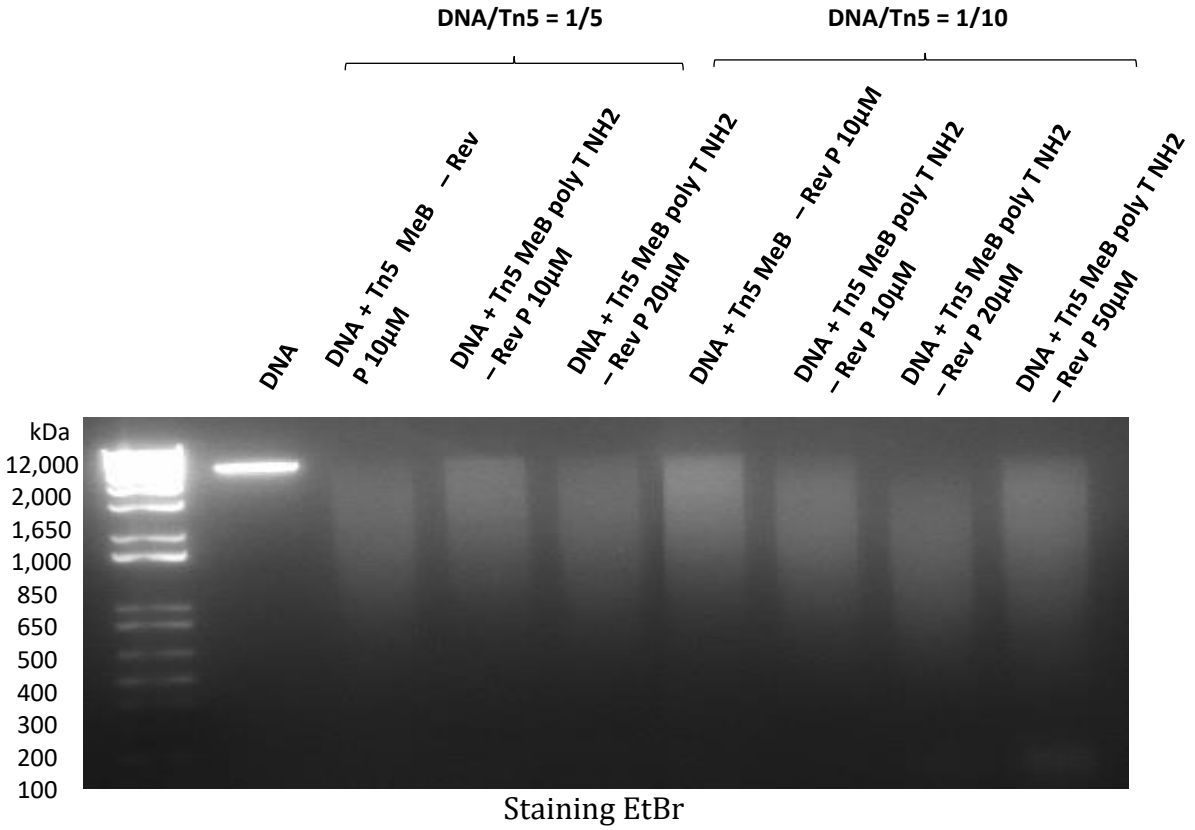
Figure 40. Tn5 tagmentation test on MeB long oligo and NH_2 polyT custom design



seems very promising for us and as expected, the PolyT tail increased the reactivity by inhibiting the potential hairpin formation.

Moreover, we focused on the MeB-polyT NH_2 custom oligo and tried different oligos concentrations with two fixed Tn5 concentrations (all of them tested in a tagmentation assay). As presented on the agarose gel in figure 41, we used different ratios of DNA plasmid vs Tn5 and different oligo concentrations (control MeB-RevP: regular oligo and the custom one MeB-polyT-NH_2). As expected, the custom design works as well as the positive control oligo and surprisingly with a very concentrated Tn5 (ratio 1/10), better than the positive control.

Figure 41. Tn5 tagmentation test on MeB NH_2 polyT custom design with more or less DNA/Tn5



Staining EtBr

Very satisfied with these positive results, we decided to move forward and tried to develop our final and definitive oligo design and tested the whole complex.

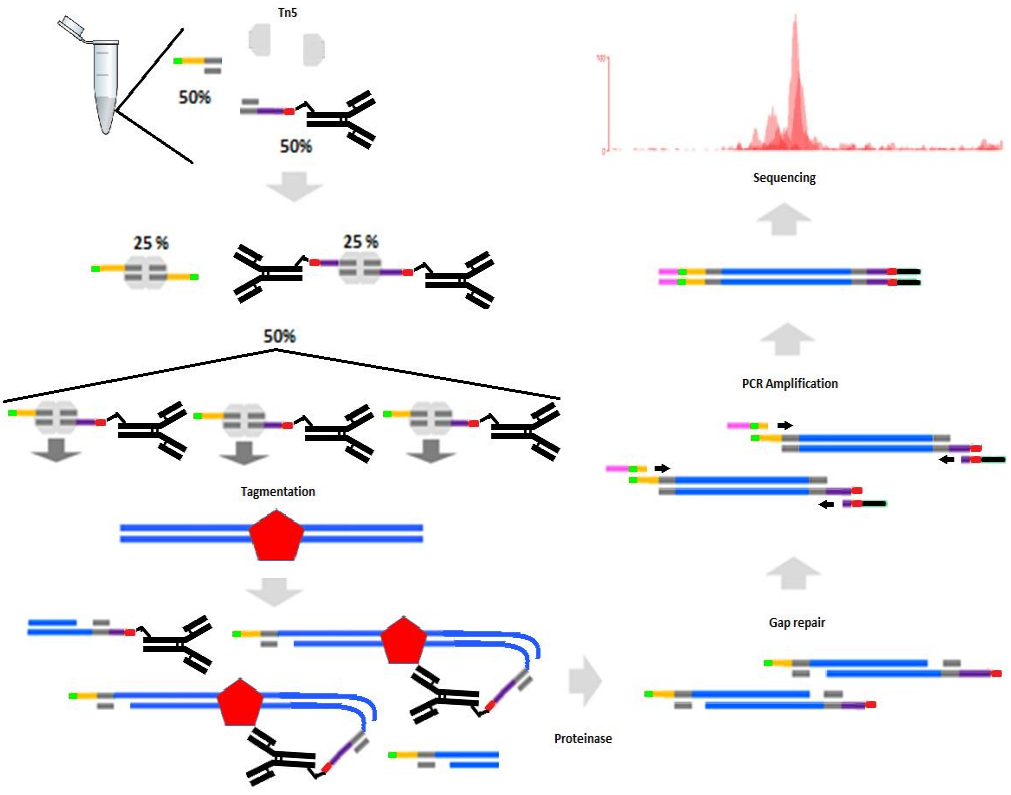
b. Definitive design and validation process of the new Oligo-Tn5 custom

Before testing both functionalities of our potential Tn5/Antibody complex, we needed to find the definitive design of our custom oligo. As we know now, we needed to capitalize on the polyT tails to create an effective design.

i. The final design chosen

After using the linear oligo MeB-polyT custom design, a problem appeared. Indeed, as resumed in figure 42, if one oligo (MeB custom) has the NH_2 antibody complexification, just 75% of the total mix of the synaptic complex can bind the antibody. Even if 75% is a high percentage, in fact, the problem came from the rest, 25%. Thus, it means that 25% of the activated Tn5 dimer has the capability to tagment anywhere in the genome without any control

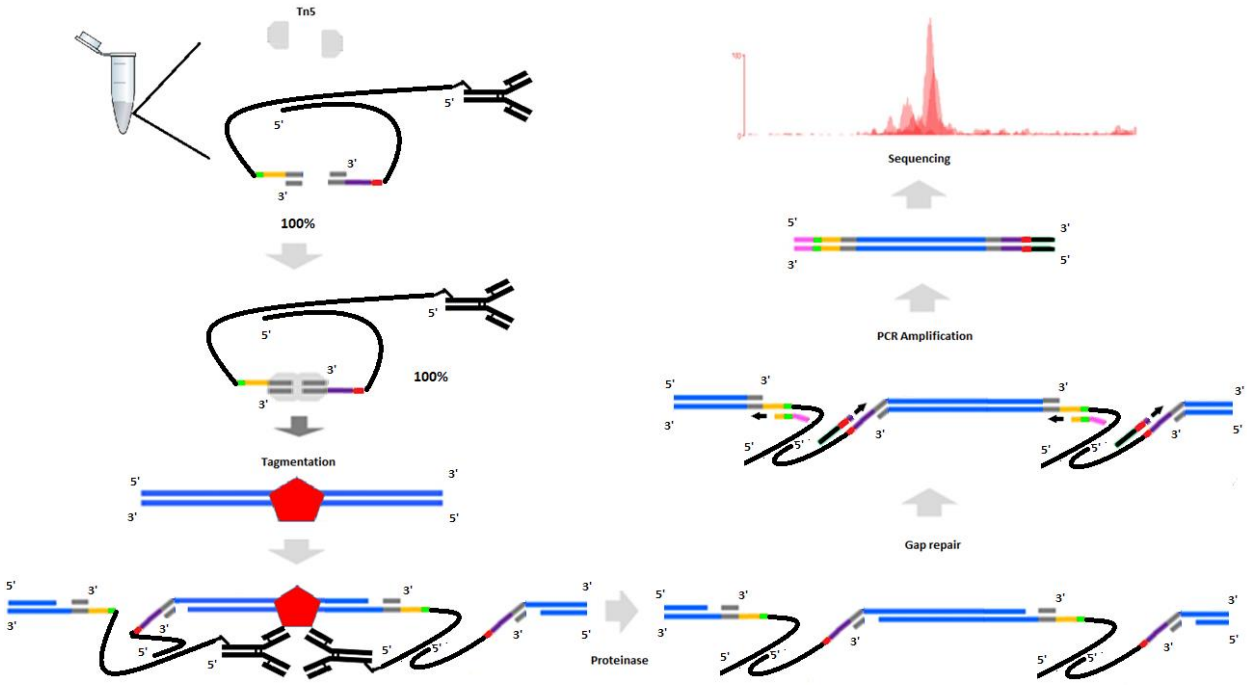
Figure 42. Final process strategy with the MeB NH_2 polyT custom



which is completely incompatible with the new methodology we are trying to develop!

So, we came back to our pens and papers and tried to find THE design which could give us optimum results. After a long day of brainstorming, a very basic idea came out: why not using the wild type synaptic complex form of the Tn5 to our advantage? Indeed, we can reproduce a sort of loop which would link both oligos together with the antibody. Next, we just need to add the Tn5 and form the synaptic complex. Thus, as presented in figure 43, this kind of design would be very useful especially that it gives just one possibility of synaptic complex formation and allows to get one dimer of Tn5 per antibody as well.

Figure 43. Final process strategy with the Chic-loop oligo design



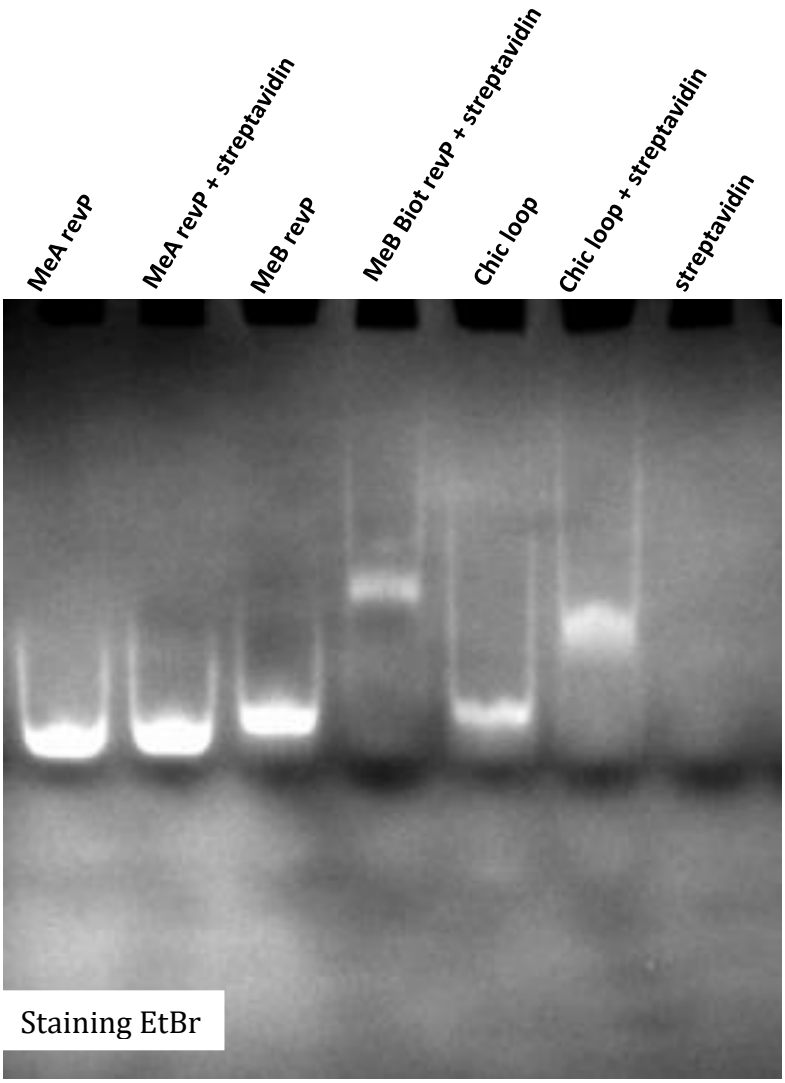
We decided to focus on this idea and designed a quite complex form of oligo with between each annealing part, a polyT tail, to reduce the potential hairpin formation. Moreover, we used different RNA/DNA 3D conformation predictive softwares to help us on

our design; RNA designer and RNA secondary structure prediction online. All oligo sequences and the final oligo designs are reported in the annexed table 2.

ii. Verification of Streptavidin binding on Chic-loop biotin oligo

To test our new loop design, we came back to the biotin/streptavidin complex test and started from the postulate that if the tagmentation works with the Tn5-oligo loop-biotin/streptavidin, this design could be reproduced with a Tn5-oligo loop-NH_2-antibody.

Figure 44. Streptavidin complex formation with custom oligo



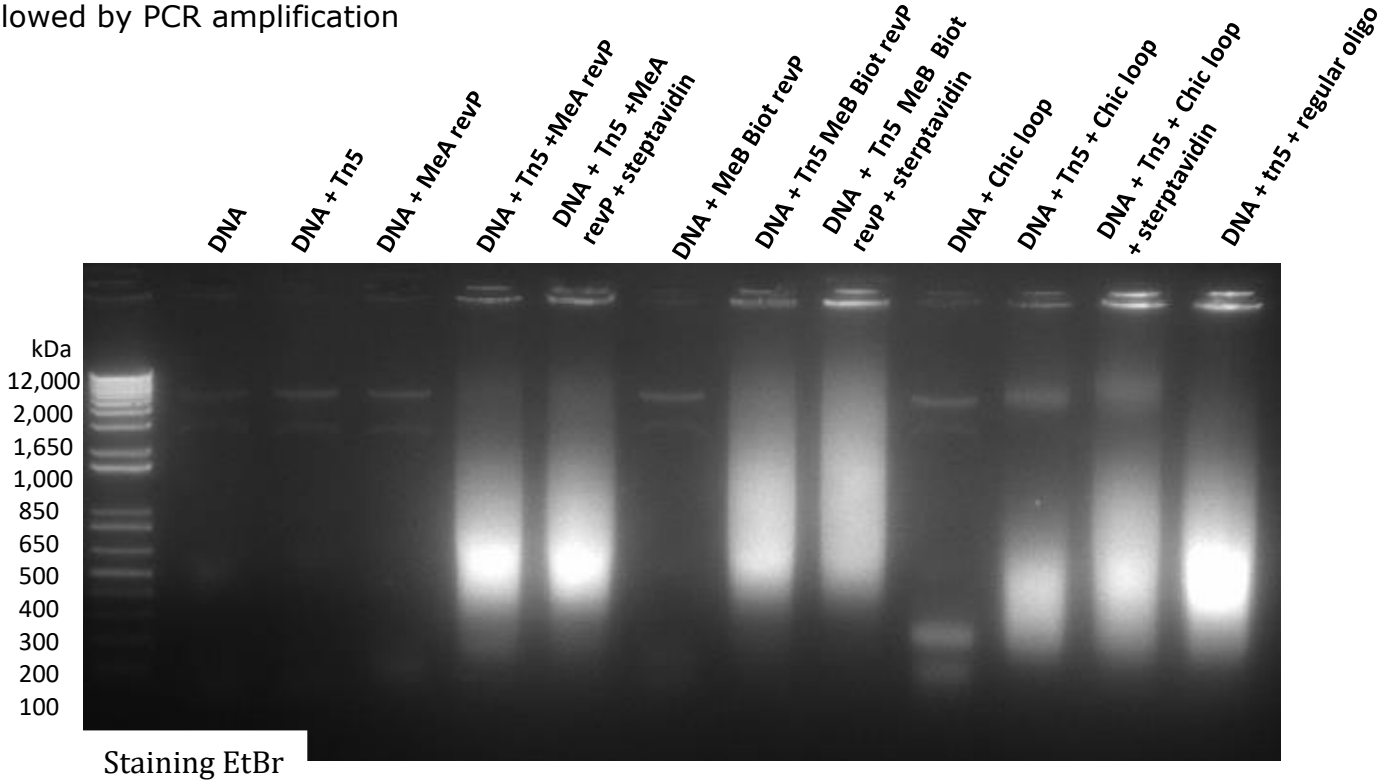
We decided, after finding the best design possible *in silico*, to order the chic-loop in pieces of ssDNA long oligo biotinylated (switch NH_2 site by the biotin) and anneal them together (95^0C 5min plus 60min at room temp) to give them the final Chic-loop form. We finally had four oligos annealed: custom MeA-polyT, custom MeB-polyT- Biotin and RevP (x2) composing the Chic-loop. We bound our Chic-loop biotin form and controls (RevP+custom MeA-polyT and RevP+custom MeB-polyT-

biotin) with the streptavidin just by adding the streptavidin (5µg) with our custom oligo (1µg) and let at RT for 30min, then a polyacrylamide gel (5%) was run.

As shown in figure 44, with the custom made MeA which has no biotin, no retarded signal is detected but for the Chic-loop biotin the same kind of retard as the positif MeB custom biotin control is observed. This result proved the binding of our chic-loop biotin with the streptavidin and we proceeded with the same oligo preparation for the final and most important experiment: take this chic-loop-SA and try to form the synaptic complex with the Tn5 to test tagmentation.

iii. Verification of tagmentation whole complex

Figure 45. Streptavidin oligo custom complexed with Tn5 test of tagmentation Followed by PCR amplification



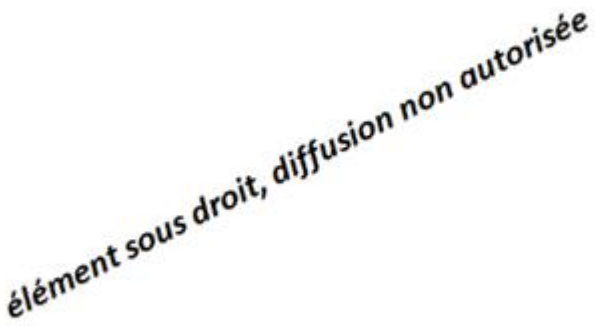
Thus, after the SA binding checked, we used the same oligo preparation and created a synaptic complex (adding Tn5 at 5µg) at 37^0C for 60 min. We made the tagmentation reaction (Tn5 vs DNA 5µg vs 0.5µg) at 55^0C for 7min (reaction blocked by 10µg proteinase k at 55^0C, 7min). We made this protocol for all previous samples of oligo alone or oligo+streptavidin and all of them were amplified by PCR reaction with the regular secondary primers used in the Picelli paper [122].

As we can find in figure 45, all expected negative controls stay with no amplified signal. For all oligos complexed with Tn5 (synaptic complexification), a specific tagmentation appears and an amplified signal is detected. More importantly, our chic-loop-Biotin-streptavidin Tn5 complex got a good amplified signal as the same range on the other Chic-loop-biotin Tn5 control and the regular oligo Tn5 positive control. Other complexes, oligo MeB custom and oligo MeA custom with or without streptavidin binding, work very well. These results are absolutely stunning and we are looking on moving forward and test the Chic-loop-NH_2-antibody.

2. Discussion

Close to realize the final *in vitro* assay with a panel of specific targeting antibodies and our designed "Tn5 loops", a brutal discovery recalls to us that our field is also very competitive. Indeed, a company was developing, at the same time, a kit based on this exact same idea. This new sequencing technology called Tam-chip™ from the Active motif® company was very similar in terms of design to our final Tn5 loop- antibody complex. I decided to do more research on it and found a patent (US 20150111788 A1) made the same year by the Active Motif™ company which developed the exact same product as the Chic-seq (Figure 46).

Figure 46. Tam-ChiP® sequencing technology develloped by Active Motif™



https://www.activemotif.com/catalog/1218/tam-chip

Indeed, after asking directly to the team who managed and developed this technology, they gave me more details in particular in terms of strategic design which appears at the end to be very similar to the one we developed (figure 47).

Figure 47. Tam-ChiP® flow chart

https://www.activemotif.com/catalog/1218/tam-chip

As we can see on the Tam-ChiP® flow chart in figure 47, the methodology is exactly the same except for minor details (loop in the oligo…). Moreover, and more importantly, they understood, like us, that it was possible to play with the Tn5 protein just by using different oligo designs and its through them that the link with the antibody is possible. Based on this fact, we decided to stop this project which took a lot of efforts for something certainly less efficient than the Tam-ChiP® technology. Despite of this unfortunate news, I decided to capitalize on this work and continue to develop new sequencing tools.

V. Conclusion

Before presenting another project, which became my second thesis project and corresponds to a manuscript which is currently in revision, it is the moment to analyze why we failed.

In my opinion, the reason why we stopped this project is essentially due to a time management problem. Indeed, the basic idea of this project came from a very simplistic fact based on the ATAC-seq paper [70]: Transforming the ATAC-seq technology to create an epigenomic technique coupled with a regular transcriptomic analysis technology. The idea itself is risky, because it does not correspond to a completely novel technique but just propose an upgrade of an existing methodology. In fact, after reading the ATAC-seq paper, it came naturally to propose a possible strategy to "guide the Tn5" through an antibody binding. But even if we knew it existed a real risk to not be the first, this one is acceptable especially when you develop a sequencing technology. Our failure in my opinion is more due to the "schedule of development".

Thus, the ATAC-seq paper was published in the beginning of 2013 and we just started to work on the project by the beginning of 2016. In fact, starting to work on this technology so late was our principle fail. Even if I knew the risk and I have been trying to compensate it by hardworking, it was an un-sufficient strategy. Indeed, as I know in the industry, when an idea is considered as "hot", they put a maximum of resources especially in the R&D team with generally multiple engineers and researchers, while I was the only one on this project and started with years of delay.

In terms of comparative design of their propose and our Chic-seq, both are based on the exact same goal. Nevertheless, the two designs are a bit different, for the Chic-seq this is a complete loop which includes both primers and binds then to the antibody. This design gave us the capability to be sure to have 100 percent of the desired product. In the opposite, the design claimed by our competitor, Active motif, seems to present a one end antibody binding (one oligo is used to the link). Even though we do not know their complete protocol of purification, their manner of proceeding, to my opinion, suffers from issues in term of resources and time, specially to obtain the right product (one antibody per dimer of transposase). But despite this point, theirs concept can possibly use not one but two antibodies which would make the technique much more flexible and this parameter was not considered by our team. A valorization of this work through an article might be possible but regarding the time and effort invested in this project, my thesis mentors and I decided to stop its development. As our competitor got a serious design very close to ours, we anticipated a significant amount of time to generated competitive data for publication.

I can, however, be proud of the work I accomplished, especially that I obtained a design very close to a designed kit from a private company. Thereby, since every cloud has a silver lining, this unexpected event had at least the advantage of convincing me of the quality of my scientific work as a scientific creator of new tools to analyze the living.

So, I decided to continue in the field of the development of sequencing technologies but with an idea I had during the reading of the Chic-seq bibliography. This time, I proposed to both of my thesis directors a completely new technology, very innovative, based on engineered extracellular vesicles.

Second Part: Engineered Extracellular vesicles for over time unbiased mRNA sampling of living cells

I. Introduction

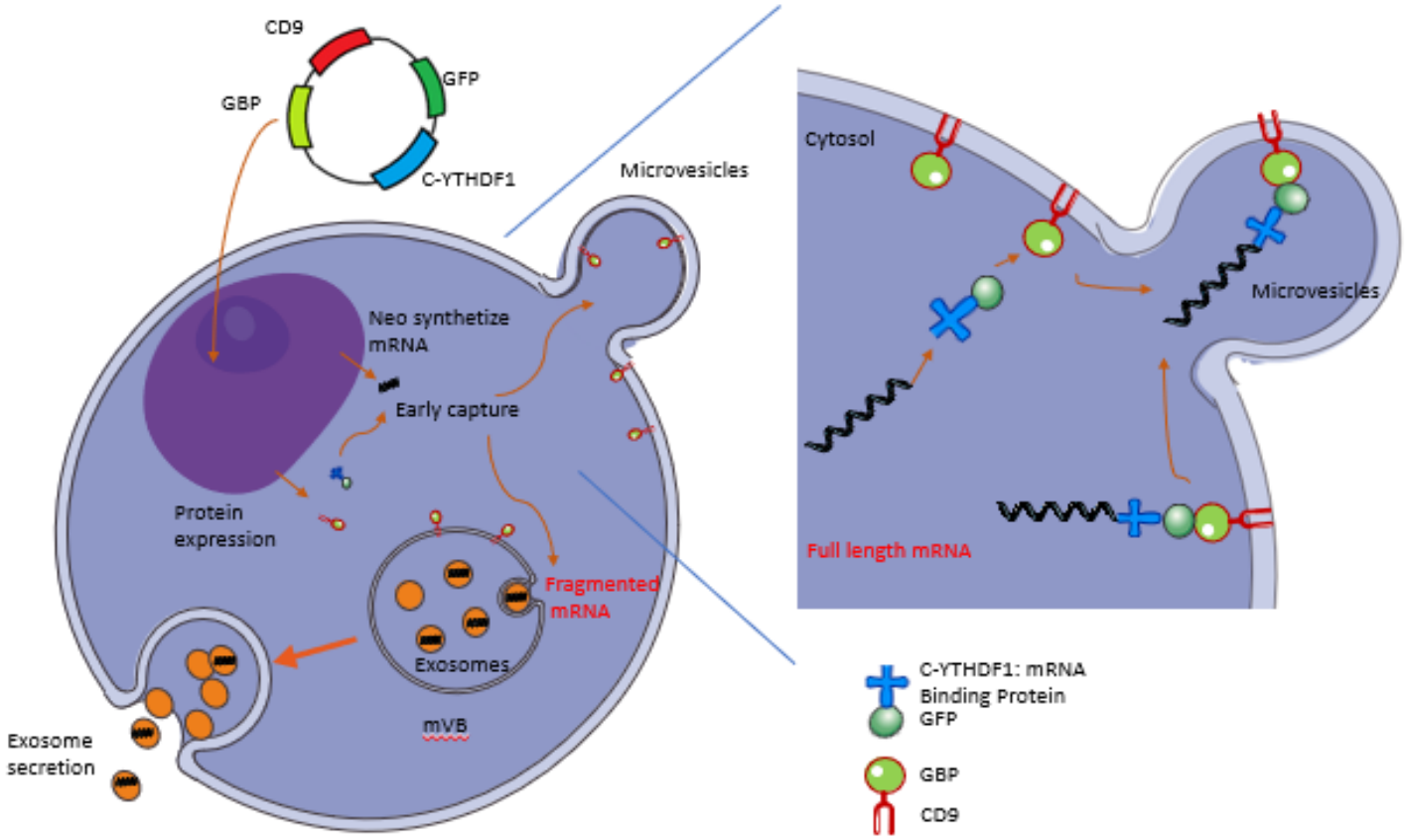
Cell heterogeneity and fluctuant genetic expression in specific microenvironments remains poorly understood. Current methodologies that seek to interrogate gene expression at a molecular level require sampling of cellular transcriptome and therefore lysis of the cell preventing serial analysis of cellular transcriptome. To address this area of unmet need, we have recently developed a new technology allowing transcriptomic analysis over time without cellular destruction. Our novel method, **TRACE-seq (TRanscriptomic Analysis Captured in Extracellular vesicles using sequencing)**, is characterized by a cell-type specific transgene expression. It provides data on a representative part of the cell transcriptome inside extracellular vesicles. Thus, the transcriptome of cells expressing TRACE can be followed over time in a non-destructive manner, which is a powerful tool for many fields of fundamental and translational biology research. We also know that the Extracellular Vesicle (EVs) RNA population is not representative of the intracellular transcriptome with a majority of fragmented mRNA [174-178]. Also, a gap

remains, and it could be possible to specifically design engineered extracellular vesicles used as carrier for transport of representative part of cell transcriptome.

Thus, to complete this goal, TRACE-seq uses a cell-type specific transgene providing an mRNA import inside the Extracellular vesicles and more specifically Microvesicles (MVs), which are used as carriers (Figure 38). A similar kind of technology already exists for protein transportation and uses engineered Extracellular vesicles as carriers [179]. Based on this technology, our method can be broken down into two parts. First of all, a mRNA "Catcher" which corresponds to a fusion protein composed of a C terminal part of YTHDF1 protein and the fluorescent transmitter enhancer GFP. Recently reported, the YTH protein domain recognizes m^6A, one of the most abundant internal modifications in eukaryotic mRNA [180, 181]. Moreover, the isoform 1 of YTH protein domain enhances the translation efficiency and binds to the mRNA close to the nucleus membrane after its translocation from the nucleus to the cytosol [182, 183]. Secondly, we designed another fusion protein constituted of a GFP binding protein 1 (GBP1), which binds to EGFP while enhancing the fluorescent signal [184, 185] and CD9 protein known as one of the most common Extracellular vesicle markers [186]. Thus, the mRNA "Catcher" EGFP-C-YTHDF1 can trap the neo synthetized mRNA, then bind to the second fusion protein through the EGFP/GBP1 affinity and bring the mRNA into

Extracellular Vesicles via the CD9 proteins, as shown in Fig.48. The justification and design will be developed after to explain our choice.

Figure 48. TRACE-seq methodology overview



II. Bibliography context

Developing a new technology requires a very careful and solid bibliography. Because of the concurrence in the field, a large portion of the development time is related to bibliographic research in the know-how for scientific publications, but patents are also a good source of investigation as well.

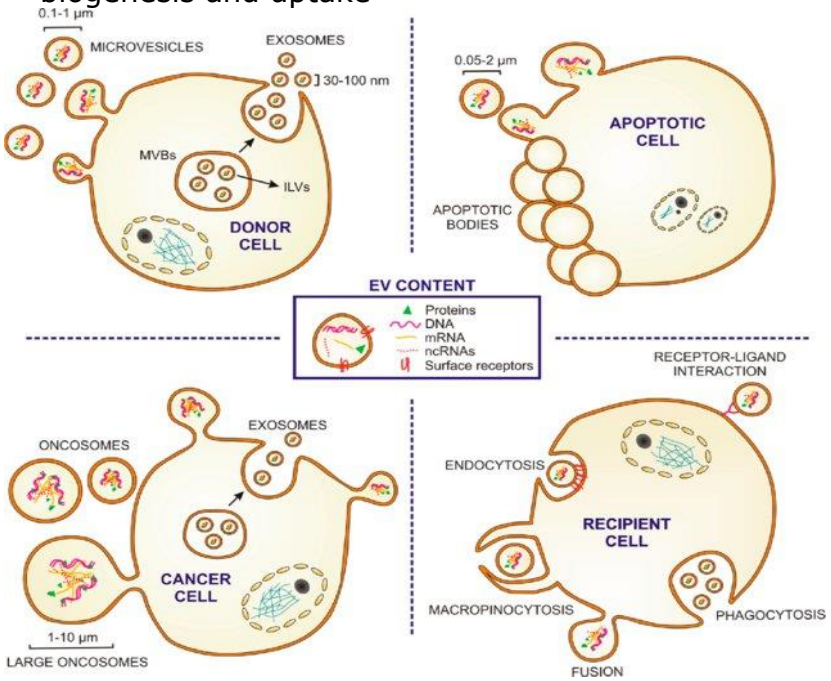
This part is not just an exhaustive listing of what we could find in the literature, but a selection of the most interesting techniques we could use to help us in our future design. It is more a way to reflect a mental path which gave us a clear analysis for the future of our proposition to not reproduce the previous failures.

1. EVs landscape, a big family.

The EVs field is very diversified, indeed, it exists different kinds of vesicles. Far from giving an exhaustive list, we will synthesize and present the major groups of EVs with their characteristics.

First, we should give some details about the classification of EVs (Figure 49).

Figure49.Extracellular vesicles: types, sizes, content, biogenesis and uptake



Esperanza R Matarredona Angel M Pastor, Cells, 2019 [187]

In fact, the EVs classification is related to vesicle size, secretion pathway and transported molecules. With 40-160nm, there are Exosomes, which come from the Multivesicular bodies produced by the mechanism ESCRT [188]. They may differ upon cell origin and can contain RNA, DNA, proteins, metabolites and surface proteins. The main role of exosomes is to remove excess or unnecessary molecules from the cells [189]. But they seem to play an important role in cell to cell signaling suggesting that they may have a function in the intercellular cell regulation [190]. In terms of purification and characterization protocols, many research groups gathered their findings several years ago in order to standardize protocols in the EVs field [191, 192].

Exosomes are not the only vesicle type present and secreted by cells. Microvesicles which are much bigger: 0.2-1μm [193], also play a role in the cell-cell communication and seem to present a large variety of RNA, DNA and proteins non directly related to the cell renewing metabolism [194, 195]. In fact, Microvesicles seem to be a more adequate choice for our purpose because they are bigger and are non-related to any RNA decay mechanism like the exosomes [196]. Moreover, the protocol for the isolation and the characterization of the microvesicles is also well established and easy to use [197].

The third type of vesicles that we should definitely consider for this project is the apoptotic bodies. They are not related to healthy cells and are associated to the apoptotic process. Because we know that in our experiment there will be some of these contaminant apoptotic bodies, this is a population of vesicles we should definitely care about. The Apoptotic bodies are bigger than the other types of EVs: 0.5-2μm [198] and even though at first we can think that they are just inert debris, they play a messenger role and transport a lot of material especially some full-length mRNA [199, 200]. Thus, we need to avoid as much

as possible these potential contaminants for our future experiments. Different methods could be used to complete this goal, our choice will be specified in the detailed protocol (mat&med part).

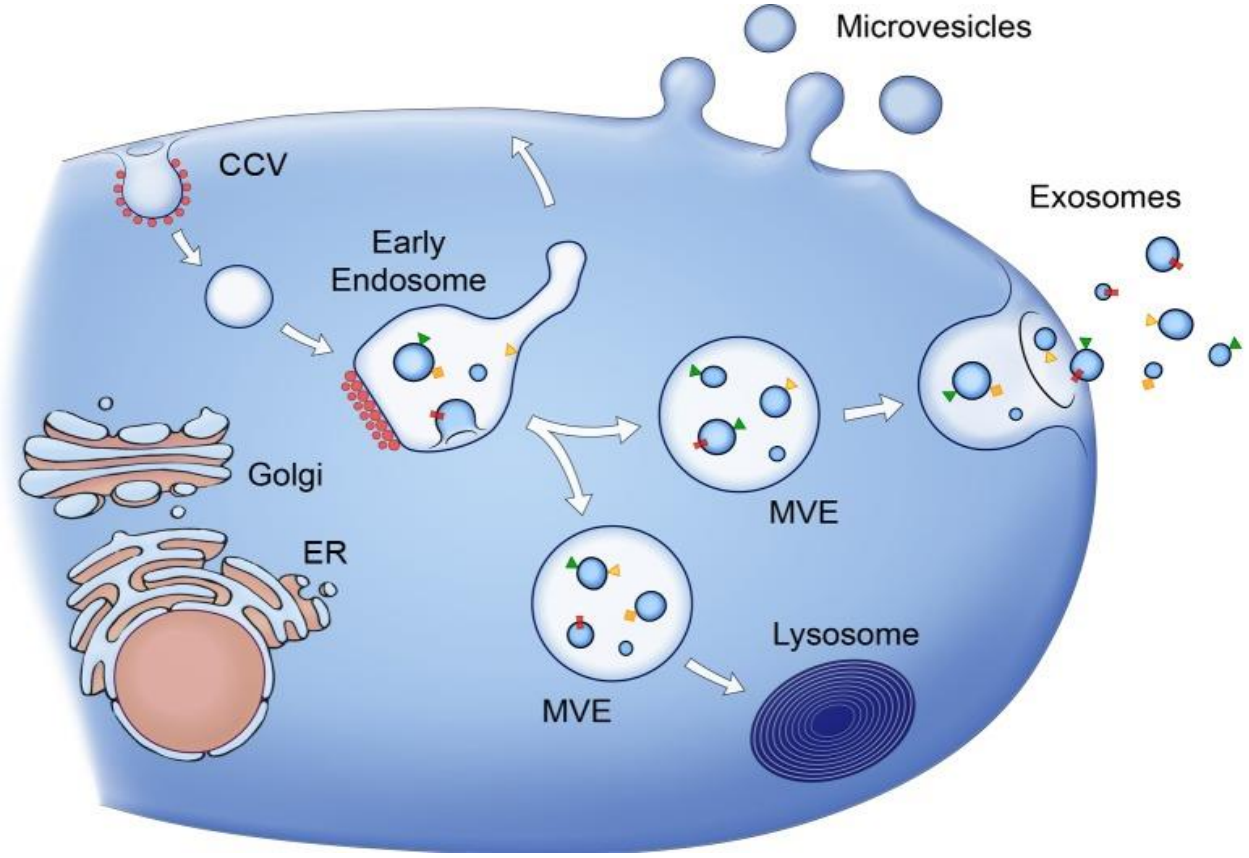
2. RNA content in vesicles, a world where everyone does not agree

Scientists disagree regarding the RNA content of EVs. Indeed, original studies treated for this question pushed more in the fact that EVs do not transport full length RNA but smaller RNAs or fragmented RNAs, not always related to the cellular one [174, 176, 201, 202]. Even though some researchers claim that EVs may contain a small fraction of full-length mRNA [203], the vast majority of the field agreed by far to say that the mRNA inside the EVs is fragmented [204, 205]. Moreover, we should keep in mind that the RNA content inside EVs depends on the parental cells and is not the same between cell types [206, 207] or cell state [208]. Thus, a gap remains and a new technique could be developed to bring a representative part of a real full-length mRNA inside EVs. Nevertheless, we first need to know more about which kind of technologies were developed to bioengineer EVs and decide which class of RNA binding proteins will be a promising candidate.

3. Binding protein systems: a link between mRNA catcher and the EV Import protein

A lot of different technologies were developed for drug delivery [209, 210], regeneration medicine [211] or diagnostic [212]. First, we need to focus on the proteins which coat the vesicle membrane. Different groups of proteins are found in EVs. Soluble ones inside the vesicles like ALIX and TSG101 or membrane bound like the tetraspanin family (CD63, CD9, CD81) [211, 213] are the most representative proteins in EVs. Moreover, the mechanism of secretion and production needs to be understood, it exists 2 major ones:

Figure 50. EVs secretion pathway

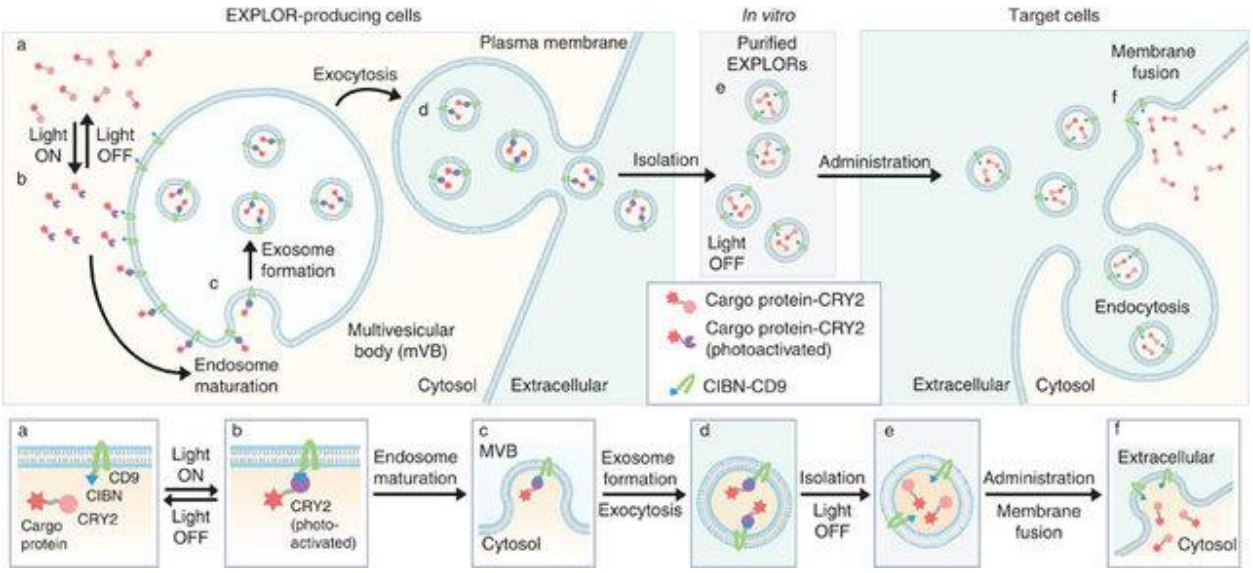


Graça Raposo and Willem Stoorvogel JCB 2013 [214]

the membrane direct invagination for the Microvesicles secretion or the Multivesicular Bodies for the secretion of exosomes [214] (Fig.50).

Instead of the natural mechanism of vesicles which represents a lot of different promising aspects for modern therapeutics, researchers mostly use vesicles as carriers like protein production/therapeutic [215] or close to TRACE aim like RBP trapping [216] or selectable RNA loading [217]. But none of them propose a new sequencing technology for an analysis of the transcriptome in live cells over time. From all different techniques of loading, one of them retained our attention, the EXPLOR technology [179, 218] Figure 51.

Figure 51. Schematic diagram of EXPLOR technology



N. Yim et al. Nature Communications 2017 [179]

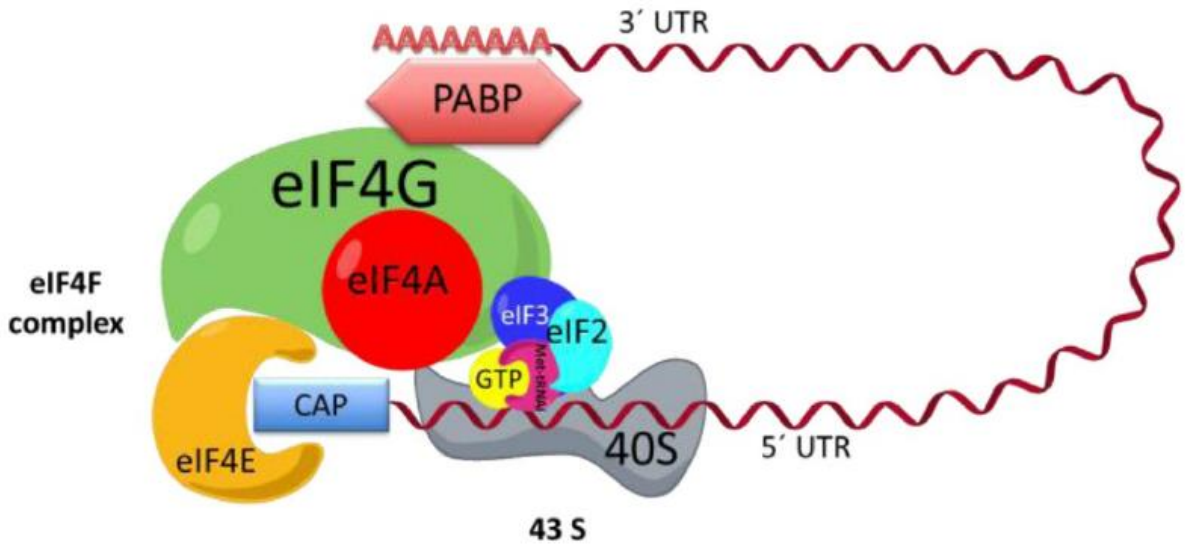
This methodology offers a very plastic possibility allowed by the binding mechanism of the CD9 protein which is a common membrane marker in EVs such as microvesicles as well as exosomes [186]. The fact that this technology has a certain kind of fluorescence is something that we should care about. Indeed, we need to think about a system to easily purify the vesicles from cell culture media or biofluids. Recent new technologies were developed to purify EVs from any kind of fluids based on fluorescence [219]. Related to the development of the nanoFACS, this very high FACS technology can sort EVs by fluorescence. We know that we need to find a mechanism to bind two fusion proteins at least: one attached to the CD9 for an EVs loading package and another one attached to a potential mRNA binding protein. Moreover, as said, this binding process should be fluorescent. After a lot of research, we finally found THE perfect mechanism of protein-protein interaction: the GFP-GBP1 protein binding, mostly known as the nanobodies complexes [184, 185]. Already used *in vivo* [220], this complex is based on the super variable binding part of the GFP antibody isolated and called GBP for GFP binding protein. In this class of new protein binders, the isoform one (GBP1) binds the GFP and increase its fluorescence (close to 100% increase) on the opposite of the isoform 4 (GBP4) [185]. Based on these findings, we decided to create an interaction complex based on two fusion proteins: CD9-GBP1 and an RBP-GFP.

4. RNA "catcher" in a jungle of potential candidates

A. The whole RBP landscape

The world of the RBPs is vast with a lot of different possible candidates [221, 222]. Because the choice will be cornelian, we decided to first focus on the mRNA looping complex just before translation. Indeed, the mRNA makes a loop with a complex of proteins to stabilize (it is also part of the decay mechanism) [223]. As shown in figure 52, the translation initiation complex involves a lot of different proteins which all play a role in the loop shape of the mRNA and promote the translation with the polysome formation.

Figure 52. mRNA translation and stability



H. Montero et al. 2015 Viruses [224]

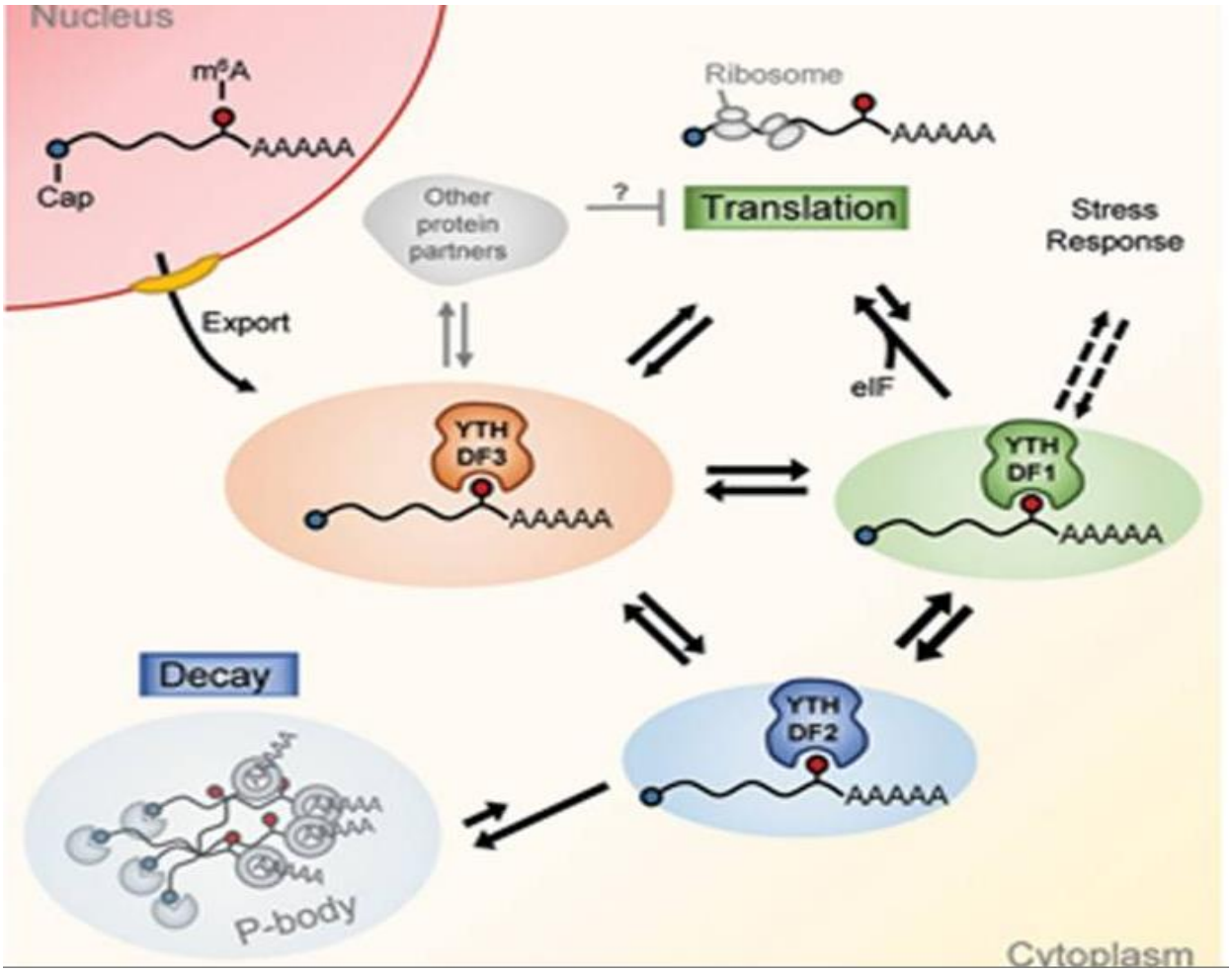
We focused on the first protein which binds the methylated cap m^7G of the 5' UTR end of the mRNA. This protein, eIF4E, is the first protein forming the complex with also multiple PABP proteins which stabilize the polyA tail. The eIF4E protein is very critical for the translation mechanistic [224] and because it arrives first and recruits the other proteins like eIF4G, eIF4A, it may be a very good candidate for us [225, 226]. Different groups decided to use it as a mRNA immunoprecipitation target and try to produce it [227] or modify it [228], but despite a lot of efforts, the principal problem of this protein is a Kd=100-200 [228], which is normal regarding the role of this protein. Indeed, the translation complex is a dynamic mechanism non fixed in the time and the eIF4E cap protein binds and unbinds multiple times before the complex gets officially formed [229]. Nevertheless, we decided to select this potential candidate and compare it with others. After multiple brainstormings, we finally got to the conclusion that we need to find a protein which binds the mRNA very quickly after it moves to the cytosol.

B. YTHDF candidates and the m^6A mRNA methylation modification

Thereby, we decided to focus on a protein which has already been used in Immunoprecipitation technology (IP) for mRNA transcriptomic analysis. We found the best candidates from these aims, thus, the YTH class protein family replies to all goals previously fixed. Indeed, the YTH protein class appears to play a critical role during the early step of mRNA regulation with 3 different isoforms all with a specific role during the mRNA regulation [181, 183]. Indeed, the YTHDF2 isoform plays a role for the decay and early mRNA

degradation response [180, 230], the isoform YTHDF1 promotes translation [182] and the isoform YTHDF3 seems to regulate the two others [181] (Figure 53).

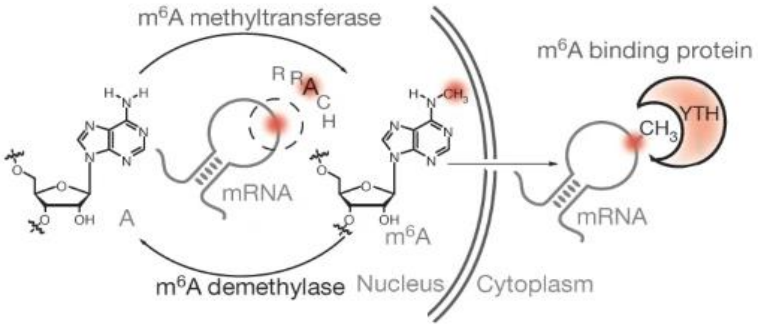
Figure 53. YTHDF group of proteins form an interconnected network in the cytosol.



S. Shi et al. Cell Research 2014 [181]

All these three protein activities are closely related to the m^{6}A Methyltransferase 3 (METTL3) which modifies the mRNA by adding a methyl group on an adenine usually in the 3'UTR region [98]. The binding of the methylated mRNA appears just after its import through the nucleus membrane (Figure 54).

Figure 54. YTHDF mRNA translation and stability



X. Wana et al. Nature 2013 [180]

The mechanism of post transcriptional modification helps the cell to reply very quickly to a mRNA demand by tagging mRNA with m^{6}A which can promote translation or the decay [231, 232]. Thus, a non-negligible quantity of the whole transcriptome is methylated and groups of specialists in the field know it has a very strong impact on the cell stress response and adaptation [182, 233, 234]. Moreover, our attention was retained by the YTHDF1 protein which promotes translation and binds the m^{6}A tagged mRNA very early, right after its translocation through the nucleus [182]. Research groups used it as a tool for immunoprecipitation of the methylated transcriptome and made RNA-seq analysis, e.g PAR-Clip methodology [182, 235]. After analyzing the 3D conformation of this protein, we concluded that the C-terminal part of the protein which corresponds to the YTH domain interacts with the mRNA m^{6}A [236, 237]. Finally, we decided to include a third different kind of RNA binding protein, something that we know we could use as a positive control for our next experiment even if we know it probably will not work for our propose. The RPL10A protein, which is part of the big subunit of the ribosome (60S), is used as a IP method like in the TRAP-seq methodology [238, 239], but we know that the Ribosome of polysome are probably too big to enter the EVs with the mRNA, so we used it as a positive control for our experiment for the first step of mRNA catcher selection.

5. Current sequencing library generation protocol compatible with an EVs context

A. Purification of nucleic acid in EVs

Analyze nucleic acids from EVs is currently a routine technique. A lot of kits and specific reagents are available on the market (e.g. qiagen exoRNeasy) and find a protocol to extract RNA from EVs is not a big challenge. To summarize it, the protocol can be broken down into three major steps. First, the isolation of EVs, different techniques can be used: with a size exclusion chromatography column, immunoprecipitation, or by the differential centrifugation technique with or without a sucrose gradient [186]. We decided to use the most commonly used technique of the ultracentrifugation with multiple centrifugation steps because this method offers the possibility to focus on different EVs population in a timely fashion.

Secondly, the RNA extraction is usually made by breaking the EVs population with organic solvent like Phenol-Chloroform. Right after the phase discrimination, the top phase (acid nucleic fraction, RNA) is isolated through RNA isolation columns.

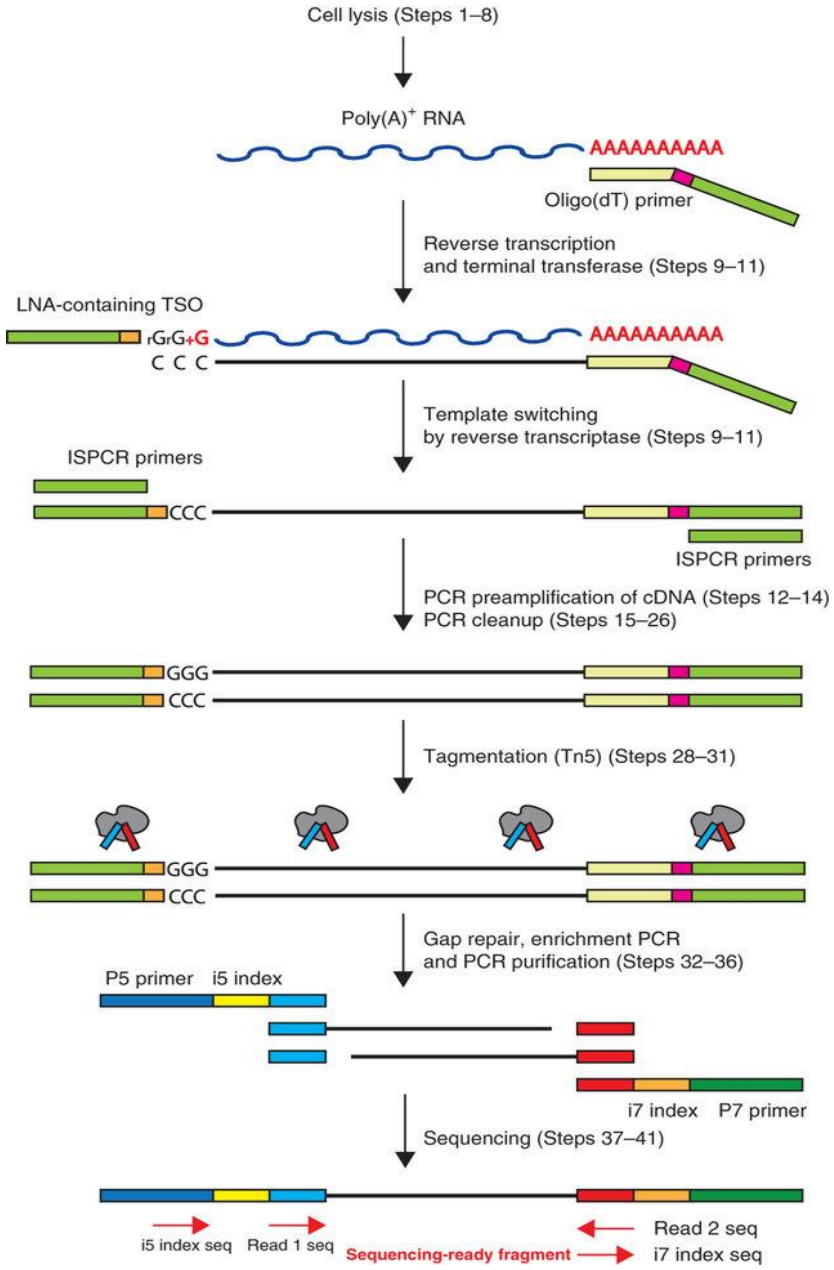
Finally, this RNA is retro-transcipted into cDNA with different kind of kit. In our case, this step is very critical and requires a much more specific technique than the traditional cDNA synthesis kits.

B. Sequencing library compatible with EVs low content

Full-length mRNA presence in EVs is subject to controversial debates. But to really know if there is full-length mRNA in EVs, the recurrent problem remains in the technique

used for the analysis. Indeed, using the regular RNA-seq library preparation is not representative because they convert every piece of the transcriptome in cDNA [123].

Figure 55. Flowchart for Smart-seq2 library preparation.



S.Picelli et al. Nature Protocols 2014 [240]

Thus, in the most common techniques, the RNA is fragmented, and random primer annealing and extension transforms it into cDNA. Also, even if everything is mapped at the end, the addition of these pieces and fragments brings a bias and some groups claim to find full-length mRNA but in reality, it corresponds more of an addition for multiple pieces of the same type of mRNA and not a real full-length mRNA. Moreover, as we saw, other techniques propose an immunoprecipitation but, in our case, the amount of starting material will be so low that it will not be viable to use them [241, 242]. There is here a necessity to find a different approach and propose a different protocol. As defined, it exists some other RNA-seq techniques which transform full-length mRNA in cDNA [243] and they are compatible with single cell analysis [240, 244-247]. We decided to base our future protocol on the SMART-seq2 protocol from S. Picelli *et al*. 2014 Figure 55. Indeed, the actual technology is compatible with the technique we are trying to develop, and we were confident for the next steps of development of the TRACE-seq methodology.

III. Material and method

1. Cells and cloning protocol

A. Cells

HEK 293T cells Human Embryonic Kidney (CRL-3216 ATCC, Manassas,VA, USA), were prepared and cultured as described by ATCC in regular complete culture media: Dulbeccco's modified Eagle's media (D-MEM; Gibco Gaithersnurg, MD, USA) supplemented with 10% fetal bovine serum (FBS; Gibco) 1% penicillin/streptomycin (Gibco). Cells were cultured in a 10cm dish (Corning, NY, USA) previously coated with Matrigel (Corning, NY, USA) diluted in F-12 media (F-12 Nutrient Mixture, Gibco Gaithersnurg, MD, USA) as recommended by the manufacturer. During the vesicles collection, a specific media was used and we called it Exo-media: Dulbeccco's modified Eagle's media (D-MEM; Gibco Gaithersnurg, MD, USA) supplemented with 10% depleted exosome fetal bovine serum (Exosome-Depleted FBS; Gibco) and 1% penicillin/streptomycin (Gibco).

B. Plasmid and cloning strategy

All PCR primers and plasmids used for this work are listed in the supplementary Table 3 and 4. The GBP1 Fragment was PCR amplified (a flexible linker GGGGGGGGGG on the C-terminal part of the sequence was added) from the *pCAG-GBP1-10gly-Gal4DBD* plasmid which was a gift from Connie Cepko (Addgene, #49438). This fragment was cloned in the *pCMV-mCherry-CD9-10* plasmid which was a gift from Michael Davidson (Addgene, #55013) to generate, the *pCMV-GBP1-CD9* plasmid. The C terminal part of the YTHDF1 gene was

amplified (a flexible linker SGGGGGGGGGG on the N-terminal part of the sequence was added) from *pGEx-4T-1-YTHDF1* which was a gift from Chuan He (Addgene, #70087) and cloned into the *pT7-eGFP-HseIF4E* gifted from Elisa Izaurralde (Addgene, #79437) and gave the *pCMV-eGFP-C-YTHDF1*. The two neo-generated plasmids were merged by cloning the pCMV-EGFP-C-YTHDF1 fragment into the *pCMV-GBP1-CD9* plasmid. The PCMV-EGFP-C-YTHDF1 fragment was purified by using the two digestions site PCI1 and MLU1, and brought into the *PCMV-GBP1-CD9* plasmid by digestion of PCI1 enzyme followed by blunting ends and cloning procedure for both fragments. The final construct obtained was *pCMV-eGFP-C-YTHDF1/pCMV-GBP1-CD9*. A negative control plasmid was generated by the double digestion with Xho1 and Acl1 endonucleases and blunt ligated the plasmid which allowed to remove the C-YTHDF1 sequence and obtain the following plasmid: *pCMV-eGFP/pCMV-GBP1-CD9*. In the meantime, another kind of plasmid was created. First, the self-cleavable peptide T2A sequence was directly included in the reverse primer for the PCR amplification of the C-term part of the CD9 sequence from the *pCMV-GBP1-CD9* with the flanking restriction enzyme sequences BBS1 and EcoR1. This C-term CD9-T2A fragment was cloned into the *pCMV-GBP1-CD9* plasmid with the same couple of enzymes and had allowed to bring back a full length CD9 sequence without Stop codon. The next step was to digest the new plasmid with the enzyme Mef1 and Afe1 and get the GBP1-CD9-T2A inserted and cloned that fragment into the GFP-C-YTHDF1 plasmid with the same couple of digestion enzymes. Thus, we finally got the plasmid GBP1-CD9-T2A-GFP-C-YTHDF1. Secondly, we added the EIF4E fragment amplified by PCR from the original the *pT7-eGFP-HseIF4E* plasmid and cloned into the *pCMV-eGFP-C-YTHDF1/pCMV-GBP1-CD9* with the digestion enzyme Bglll and Xho1 to finally get the *pCMV-eGFP-EIF4E-C-YTHDF1/pCMV-GBP1-CD9* plasmid.

For the Lentivirus construct, the plasmid and the two endonucleases Afe1 and Mlu1, were used to put the fragment *eGFP-C-YTHDF1/pCMV-GBP1-CD9* or *eGFP/pCMV-GBP1-CD9* in the lentivirus backbone. The pCW57.1 (which was a gift from David Root, Addgene plasmid # 41393) was previously opened with the two enzymes BMT1 and Age1 and a fragment cleaved with the same enzymes, which also contained the two endonucleases Afe1 and Mlu1 were added. The new pCW57.1 plasmid containing Afe1 and Mlu1 was ready to receive each construct which were cloned with these two endonucleases. We obtained at the end the inducible tet-on lentivirus plasmid: *pCW57.1 TRE-eGFP-C-YTHDF1/pCMV-GBP1-CD9* and *pCW57.1 TRE-eGFP/pCMV-GBP1-CD9* for the plasmid control. TRE promoter with tet operators were also obtained from PCR to the pCW57.1 Afe1 Mlu1 plasmid. The generated TRE promoter fragment was cloned in each plasmid construct with NsiI and SgrAI enzymes to remove pCMV promoter and get the GBP1-CD9 fusion protein expression under control of the tet-on induction system. We finally obtained the two following plasmids: *pCW57.1 TRE-eGFP-C-YTHDF1/TRE-GBP1-CD9* and *pCW57.1 TRE-eGFP/TRE-GBP1-CD9* for the plasmid control. Exact same method (cloning steps and enzymes) was used to generate the plasmid *pCW57.1 TRE-eGFP-EIF4E-C-YTHDF1/TRE-GBP1-CD9* from the original *pCMV-eGFP-EIF4E-C-YTHDF1/pCMV-GBP1-CD9* plasmid.

C. Transfection, Transduction and virus Production

To use our new methodology, we decided to developpe three ways of delivery which are related to the capability of *in vitro /in vivo* monitoring, cell line fate analyzing, cell transcript reaction to biomaterials, drugs… As a proof of principle, we directly transfected

the *pCMV-eGFP-C-YTHDF1/pCMV-GBP1-CD9* and *pCMV-eGFP/pCMV-GBP1-CD9* plasmids to 1 or 4 Million HEK 293T cell in a 10 cm dish by lipofectamine 3000 (Invitrogen, Carlsbad, CA, USA) with the protocol recommended by the manufacturer. A negative control with regular HEK293T (untransfected cells) was also included and treated as samples like the following. After 12 hours incubation at 37^0C, all the plates were washed two times with PBS (Gibco Gaithersnurg, MD, USA) and 10mL of fresh exo-media was added complemented with RNAse A (Qiagen, Hilden, Germany) at a final concentration of 30µg/mL. Cells were left in the collection media for 24H at 37^0C for EVs production.

Regarding the lentivirus, three plasmids were transfected at the same time, one of the two constructs: *pCW57.1 TRE-eGFP-C-YTHDF1/TRE-GBP1-CD9* and *pCW57.1 TRE-eGFP/TRE-GBP1-CD9* for the plasmid control mixed with the psPAX2 and the pMD2.G plasmids. All three were transfected in HEK293T in 10 cm dishes with lipofectamine 3000 with the protocol recommended by the manufacturer. Twelve hours after transfection, the media was replaced and collected two times 36 and 52 hours after transfection. The collection media was spun at 1,000g for 5 min, supernatant was filtered with 0.8 µm filters (Millipore Co, Burlington, MA, USA) and let in incubation overnight with PEG-it virus precipitation solution (System Biosciences LLC, Palo Alto, CA, USA). After incubation, the Lentivirus solution was spun at 1,500g for 30 min and the virus pellet was resuspend in 1 ml PBS aliquoted in 100µL and stored at -80^0C. A fraction of each lentivirus production was used for a functional virus titration on HEK293T cells and gave us the MOI of 4.2.

D. Stable cell lines generation

All HEK293T stable cell lines were obtained by infecting with the lentivirus generated from the two following constructs *pCW57.1 TRE-eGFP-C-YTHDF1/TRE-GBP1-CD9* and *pCW57.1 TRE-eGFP/TRE-GBP1-CD9* for the plasmid control. Regarding the MOI of the virus, 10K cells where plated in 6 wells dish in regular complete media and incubated for 24 hours. After incubation, cell media was removed and replaced with 100µL Lentivirus PBS solution from -80C stock + 1.9ml of DMEM complete + 10 µg/mL polybrene infection reagent (Sigma-Aldrich, Saint Louis, MO, USA) and incubate for 48 hours. After the infection time, the media was switched to selection media (regular complete media + 2µg/ml Doxycycline (Fisher Scientific, Hampton, NH, USA) with 2µg/ml puromycin (Fisher Scientific, Hampton, NH, USA)) and incubated for 2 weeks with 2 cell passages with Accutase (STEMCELL Technologies Inc, Vancouver, Canada), selection media being refreshed every two days. At the last cell passage, the remaining green cells were counted and a subcloning in a 96 well plate was performed in order to dilute the cells down to 1 per well in 100µL of the selection media with puromycin 0.5 µg/ml. After the apparition of cell clusters, the best clone of each stable cell line was selected regarding GFP membranous signal and passaged with Accutase in 6 well dishes with regular complete culture media without doxycycline. At 70% confluency, each stable cell lines is ready to use for experiments and a fraction of them was stocked in liquid nitrogen according to our standard procedure (complete media + 10% DMSO) (Sigma-Aldrich, Saint Louis, MO, USA). To reduce cell stress as much as possible and protect the membranous proteins (especially CD9), all passages were executed with Accutase.

2. Extracellular vesicles and RNA purification

A. Extracellular vesicles purification

MVs were directly purified from the supernatant of 1-4 million transfected cell or 6 million cells from cell line HEK 293T**.** It is very important to keep the cells healthy and under 75% confluency as much as possible in order to minimize the release of apoptotic bodies. For the transfected cells, they were washed two times with PBS after a 24 hours incubation in the transfection Exo-media complemented with RNAse A at 30µg/mL. For the lentivirus cell line, cells were previously treated with Doxycycline (Fisher Scientific, Hampton, NH, USA) at a final concentration of 1µg/ml two times at 24 and 48 hours before the incubation in the exo-media. Next, cells were washed two times with PBS and incubated in the exo-media complemented with RNAse A at 30µg/mL and 2µg/ml Doxycycline was added and let incubate for 24 hours. After incubation, the exo-media was collected, and multiple centrifugations were performed. First, a 300g spin at 4^0C for 10 min to remove cell debris was performed. The supernatant was collected and spun a second time at 1500g at 4^0C for 15 min to eliminate the large majority of the apoptotic bodies. To get more chance to reduce and completely limit the apoptotic bodies contamination, we filtered all supernatants with 0.8µm filters (Millipore Co, Burlington, MA, USA) and made a final spin at 16.500g, 4^0C for 30 min. After this spin, the remaining microvesicles pellet was carefully resuspended in PBS supplemented with vanadyl ribonucleoside complex (NEB, Ipswich, MA, USA) at a final concentration of 10 mM to protect any mRNA released from clamped vesicles. To purify exosomes, a last spin of the supernatant was performed at 100000g for 70 min at 4^0C and the exosomes pellet was resuspended in 200 µL PBS supplemented with vanadyl

ribonucleoside complex. EVs population are also analyzed and sorted based on their green signal by nanoFACS: Beckman Coulter MoFlo AstriosEQ 4 laser system.

B. Cell purification

For each EVs samples, an aliquot of corresponding cells is isolated by a treatment of trypsin solution at 0.25% (Gibco, Trypsin 0.025%) for 1min at 37^0C and the reaction was stopped with regular complete media. A fraction of each cell population (10000 Cells) was spun down at 1,500g for 5min at 4^0C, the residual media was removed and cells were resuspended in 200 µL PBS and ready for the next step of the TRACE protocol.

C. RNA purification

Right after the EVs and cells isolation, the RNA purification protocol is started. Each EVs and cells solution was treated with 700µL TRIzol LS (Fisher Scientific, Hampton, NH, USA) and vortexed 5 sec and incubate 25^0C for 3 min. A pure solution of 140 µL of Chloroform (Fisher Scientific, Hampton, NH, USA) was added and the tubes were vortexed for 15 sec and incubated at 25^0C for 5min. Right after the incubation, the tubes were spun down at 12,000g for 15 min at 4^0C. The upper phase of each tube was taken carefully and mixed with 2 vol of 100% Ethanol solution (Fisher Scientific, Hampton, NH, USA). Each solution was added to the RNA clean up Qiagen column (RNeasy MinElute Cleanup Kit Qiagen, Hilden, Germany) and spun down at 8,000g for 30 sec. There, only for the cell lysates, the column filters were treated with DNAse 1 solution from Qiagen as the protocol recommended by the manufacturer. Columns were then washed two times with a 70% ethanol solution, a volume

of 700µL and 500µL for the second one was used, then spun for the two washes at 8,000g for 30 sec and at 8,000g for 1min for the last one. The columns were also dried at 8,000g for 1min and eluted with 10µL of RNAse free water (Qiagen, Hilden, Germany) with incubation of 1 min at 25^0C and spun down at 8,000g for 1min. To certify the presence and quality of the purified RNA, each sample was analyzed on a RNA 6000 Pico chip from Agilent technologies (Agilent Technologies, Santa Clara, CA, USA) by loading 1µL of each EVs eluant and 1µL of a diluted RNA from cell lysate (usually correspond to 5ng RNA for the lysate).

3. RNA modification and analysis

A. Reverse transcription and PCR preamplification

The reverse transcription protocol, PCR preamplification and was performed according to the Smart-Seq 2 paper (S. Picelli *et al*. 2014) except for the TSO oligo which were designed and ordered through the Qiagen custom LNA oligonucleotides tool (www.qiagen.com, Qiagen, Hilden, Germany). Moreover, volumes were adapted as the following: In 0.2mL PCR tube: 9µL form each solution and EVs RNA were mixed with 1µL of 10µM $Vnd_{30}T$ oligo and 1µL of dNTPs mix at 2mM each. 0.5µL of each cell lysate which corresponds to 30ng) was used with 8 µL RNAse free Water and with 1µL of 10µM Vnd30T oligo and 1µL of dNTPs mix at 2mM each. Samples were annealed in a thermocycler as recommended in the SMART-Seq2 protocol and cDNA reaction was processed:

Table 1: Mix preparation cDNA generation

Component	Volume (µL)	Final Concentration
SuperScript II (200 U μL^{-1})	1	100 U
RNAse Inhibitor (40 U μL^{-1})	0.5	10 U
SuperSript II Buffer (5X)	4	1X
DTT (100 mM)	1	5 mM
Betaine (5 M)	2	0.5 M
MgCl2 (1 M)	0.1	5 mM
TSO (100 µM)	0.2	1 µM
NRAse free water	0.2	-
Total volume	9	-

The 9µL of Master reaction mix was added to each sample and to obtain a final reaction volume of 20µL. After mixing the solution gently and spinning down 700g for 10s at 25^0C, samples were thermal cycled as recommended in the Smart-Seq2 protocol.

Right after the RT the PCR mix for the preamplification is prepare as the following:

Table 2: Mix preparation First amplification reaction

Component	Volume (µL)	Final Concentration
First Strand reaction	20	-
KAPA HiFi HotStart ReadyMix (2X)	22.5	1X
IS PCR primers (10 µM)	0.45	0.1 µM
NRAse free water	2.05	-
Total volume	45	-

The PCR in thermocycler run was performed as the Smart-Seq2 protocol. Finally, each sample weas cleaned up using Zymo select a size column (Zymo research, Irvine, CA, USA) to cut of small fragment and keep up to 300nt. Each sample was loaded in 40 µL of elution buffer. To judge of the quality of all pre-amplified cDNA, all samples were analyzed on High sensitivity DNA chip assay from Agilent technologies (Agilent Technologies, Santa Clara, CA, USA) by loading 1µL of each EVs eluant and 1µL of a diluted DNA from cell lysate (usually correspond to 5ng DNA for the lysate) and an expected broad peak with an average size of 600-2000 bp would be observed.

B. Tagmentation reaction and amplification of adapter-ligated fragments

These following steps were made as recommended in the Smart-Seq2 protocol (S. Picelli *et al*. 2014) and using the Illumina Nextera XT DNA kit. As specified in the manufacturer protocol, each sample is normalized on the same volume for all EVs fraction and as well for the cell lysate preparation, but each of them should never exceed 1 ng. In our case, 500pg were used, and the same amount of reagent from the Nextera XT DNA sample preparation was use, as the following, in 0.2mL PCR tube:

Table 3: Mix preparation tagmentation preparation

Component	**Volume (µL)**	**Final Concentration**
Tagment DNA buffer	10	1X
Amplicon tagment mix	5	-
Amplified DNA sample	Variable	-
Nuclease free water	Variable	-
Total volume	20	-

Samples were incubated in a thermocycler as recommended by the manufacturer. The tagmentation reaction was stopped by adding 5µL of Nextera NT Buffer, incubated for 5min at RT to each tagmentation mix.

To all tagmented samples, the following PCR mix was added:

Table 4: Mix preparation for Tagmented samples followed by PCR amplification reaction

Component	**Volume (µL)**	**Final Concentration**
Tagmented DNA Sample	25	-
Nextera PCR master mix	15	-
Index 1 primers (N7xx)	5	-
Index 2 primers (N5xx)	5	-
Total volume	50	-

All samples were incubated in thermocycler and the PCR amplification steps program was performed as recommended by the illumina Nextera protocol. The amount of PCR cycles during this amplification depends on the starting DNA material originally used as recommended in the Smart-Seq2 protocol (S. Picelli *et al*. 2014). We used 500pg so 10 cycles were performed for this amplification. After this final amplification step, each sample was cleaned up using AmpureXP beads (Beckman Coulter Co. Brea, CA, USA) as recommended in the Smart-seq2 paper.

To judge of all cDNA library quality, each sample was analyzed on High sensitivity DNA chip assay from Agilent technologies (an expected broad peak with an average size of 300-800pb will be observed) and quantified with the *Qubit* 2 Fluorometer (Invitrogen, Carlsbad, CA, USA).

Referring to this two quality control values, the concentration (ng. μL^{-1}) and an average size obtained on the Bioanalyzer were used to calculate the relative molarity of each final library. Based on these molarity values, each sample was properly diluted to get final library solution at 2nM each. At this step, all samples from the Nextera XT kit were pooled together by following the recommendation from the manufacturer. The pooled library was again checked by Qubit and adjusted if necessary, to keep solution at 2nM.

C. cDNA sequencing

Primary data processing. Data were collected using 50 × 8 × 50 reads on a HiSeq. Reads were aligned to hg19 using STAR and counting reads associated genes were detected with the FeatureCounts module. Then we proceed to a filtering to remove low expressed genes with cpm function with the edgeR package. In this dataset, we chosen to retain genes if they are expressed at a counts-per-million (CPM) > 0.5 in at least two or more counts.

Differential expression. A normalization was performed on the Data set to eliminate composition biases between libraries using the calcNormFactor function. Normalized data were tested for differential expression by using the limma-voom package. Data were voom transformed and differential expression test was made using the lmFit function form the limma package. Moreover, a differential expression test between TRACE-seq experiments and results from the POSATR database [29,30] (corresponding to the methylated mRNA from YTHDF1 IP lysate from Hela cells) was made through the same way of analysis as above.

RNA coverage analysis. Coverage is first computed using the bedTools package. Then a sliding-window approach is used to identify and quantify consistently covered regions of detected transcripts. Briefly, using a window of 50nt we require an average of at least 10 reads in order to class this window as 'covered'. Then the fraction of windows across each transcript that satisfy this detection criteria is reported as the fraction covered.

Cell line in oxidative stress testing. GFP-EIF4E-C-YTHDF1/GBP1-CD9 Cell line and regular HEK 293T cells were tested in H_2O_2 oxidative stress. Twelve plates of 150 mm were made (6 for the TRACE cell line and 6 for the regular HEK 293T cells). When reached 15M cells per well, cells were washed two times with PBS and placed into Exo-media with RNAseA at 30µg/mL final concentration. Moreover, half of each cell population (3 GFP-EIF4E-C-YTHDF1/GBP1-CD9 cell line and 3 None regular cell) was treated with H2O2 peroxide (hydrogen peroxide solution 30% (v/v) from Sigma-Aldrich, Saint Louis, MO, USA) at final concentration of 50µM. After 24 hours of culture at 37^0C, 5% CO2, conditioned media from each plate was collected and processed according to the EVs purification protocol previously described and, in the meantime, each cell lysate was collected as previously described. For each sample, RNA was extracted by phenol/chloroform extraction and RNA clean up Qiagen columns were used to isolate RNA population from each sample. Directly after their isolation, the RNA was reverse transcript into cDNA and pre-amplified as previously described. Nevertheless, the pre amplification step was a little different as a total number of 25 cycles was performed this time. Samples were purified as usual with zymo kit and samples were eluted into 40µL for qPCR analysis on a batch of 8 specific oxidative genes, 1µL of cDNA per reaction and the 2X SYBR^{tm} Green PCR Master mix (Applied Biosystems^{tm} Foster City, CA,

USA). The qPCR was run in triplicates with 10µL per reaction on a QuantStudio 6 384-well formats (Applied Biosystems^{tm} Foster City, CA, USA).

IV. Results and Discussion

1. Experimental results

A. Design and theorical cloning strategy

a. Process of production and validation of the basic roles to validate a final design

After a thorough bibliographic research, we now have a sketched plan and need to search a plasmid commercially available and as easy to use as possible to complete our goal. Moreover, all primer sequences used to make the TRACE construct are referred in the annexed 3 table. We first tried to validate the basic concept by a very fast and easy experiment to then move to a much more complex design.

i. A complex cloning strategy

First, because our process of development is a complicated strategy of cloning, this is better to quickly summarize here the overall steps. For a much more detailed protocol, see the material and method part.

We decided to test our two major goals by breaking down our strategy in two different plasmids coding for two different fusion proteins. One for the CD9-GBP1 and the other one for the RBP-GFP. First, we got the GBP1 sequence (Addgene #49438), and

amplified it. We next cloned this GBP1 fragment into the mCherry-CD9 plasmid (Addgene #55013) instead of the mCherry sequence.

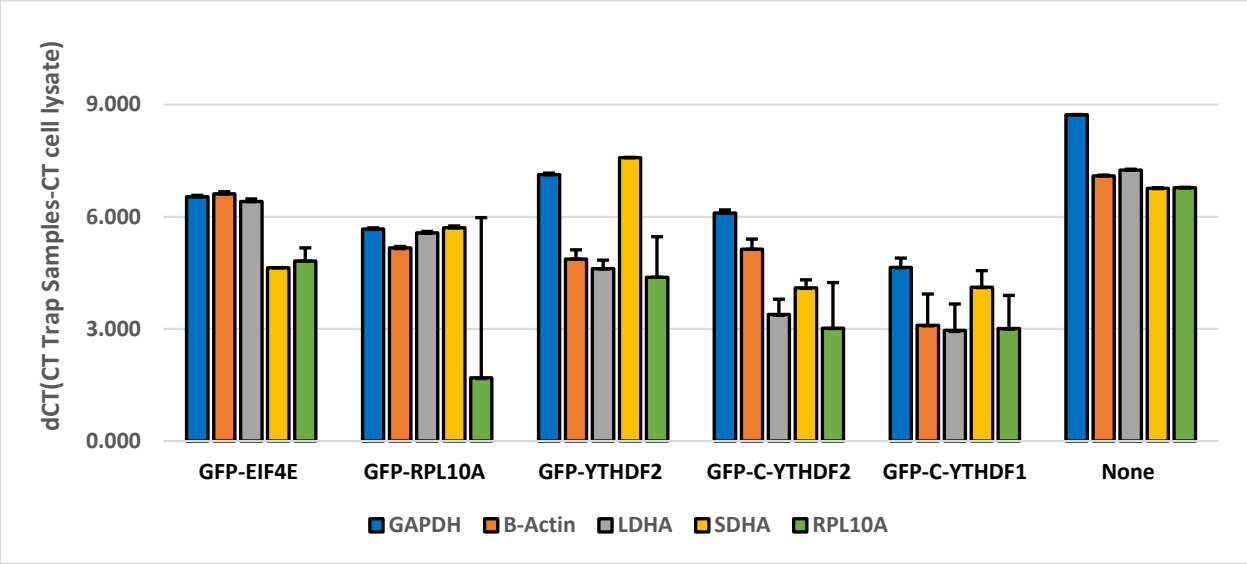
For the other RNA binding protein, two of them were amplified from the plasmid (Addgene #70087). It corresponds to the protein C-Term-YTHDF1. Other proteins: the full YTHDF2, C-Term-YTHDF2 and the RPL10A were amplified directly from an cDNA bank obtained by mRNA extraction from HEK293T cells. All fragments were cloned into the Addgene #79437 instead of the EIF4E fragment, note that the full plasmid eGFP-EIF4E was also used for the next experiments. For a better visualization of the complete cloning strategy, please refer to figure 58.

ii. Validation of the basic role of both parts of the constructs

First, we needed to know if our different eGFP-RBP could catch the mRNA. To do so, we transfected HEK 293T cells and extracted the mRNA from all the different cells. All cell populations were: HEK transfected with eGFP-RPL10A, eGFP-C-YTHDF1, eGFP-YTHDF2, eGFP-C-YTHDF2 or eGFP-EIF4E plasmids and the regular HEK293T was added. All mRNAs were converted into cDNA and we ran a qPCR experiment based on primer from Annexed Table 4.

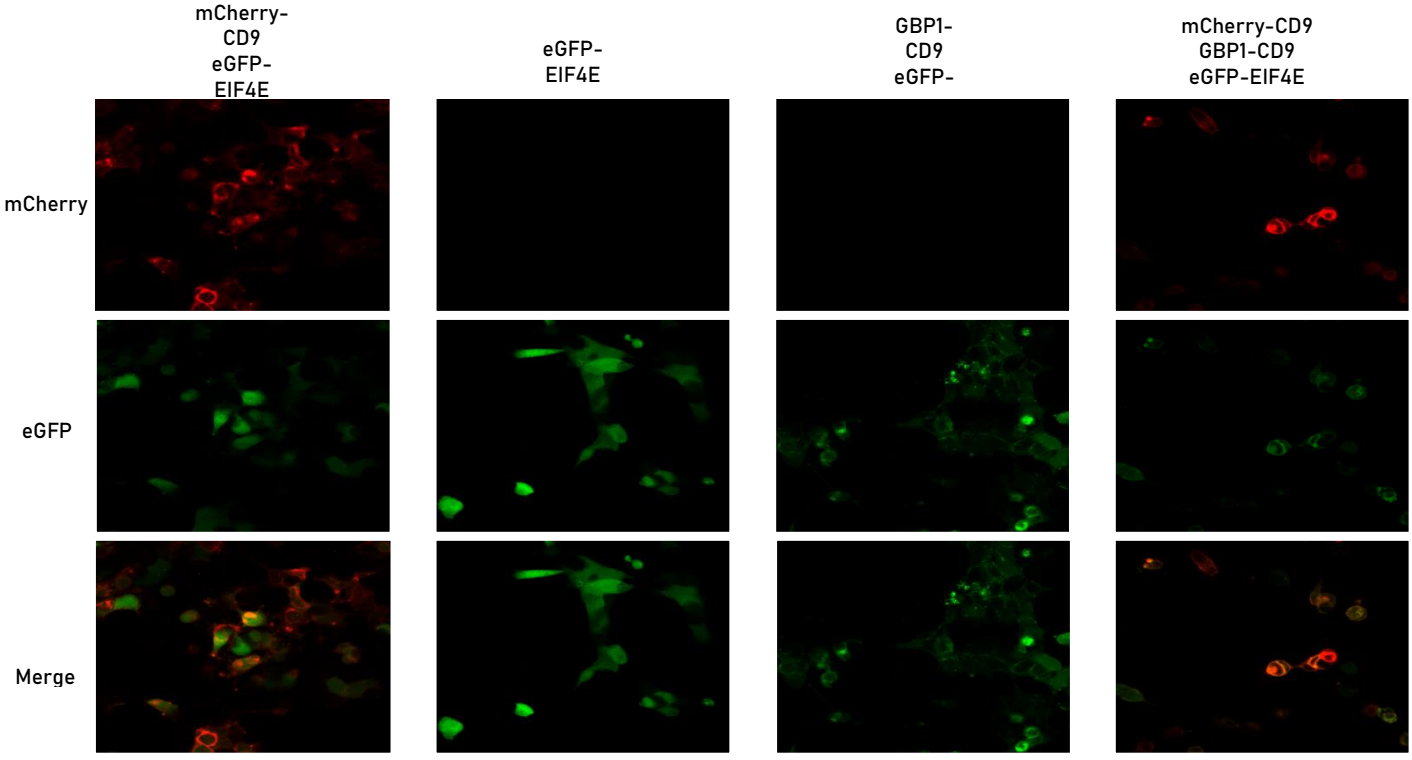
As shown in Figure 56, all RBPs seemed to trap the mRNA as expected. Indeed, all dCT values were lower than those from the None transfected cells (<6 dCT, which is our base line) meaning that the mRNA is trapped by our tested RBPs.

Figure 56. RBP test made by qPCR from different HEK293T cells population



RNA binding protein (RBP) Immunoprecipitation results of different HEK 293T populations transfected with five eGFP-RBP constructs. For each of them, cells were lysed, mRNA was purified and immunoprecipitated using an anti-GFP antibody. These trapped genomic populations were normalized against the remaining expression level of their own whole lysate.

Figure 57. Colocalization signal generated by confocal from different HEK293T cells population



Confocal imagine from Transfection of different HEK 293T populations with two constructs: eGFP-EIF4E, eGFP-EIF4E/GBP1-CD9 and mCherry-CD9. Imagine after 36h post transfection.

Surprisingly, our positive control RPL10A, the one corresponding to the polysome signal, is not the lowest. It seems that the C-YTHDF1 protein made a very good trap for the tested genes. On the side of the CD9-GBP1 plasmid check, we decided to transfect HEK293T cells with the plasmid CD9-GBP1 and the eGFP-EIF4E and see by confocal microscopy if there was a potential membranous signal proving the overall function of the CD9 and GBP1/eGFP interaction.

As presented in Figure 57, multiple kinds of transfections were realized. Thus, for the first row, two plasmids were added: the mCherry-CD9 and the eGFP-EIF4E. As expected, both signals (red and green) were different, with a red coating the membrane and a green one spreading in the cytoplasm. The second row corresponds to the eGFP-EIF4E alone and the signal is whole cytoplasmic. The third one corresponded to two plasmids, the CD9-GBP1 and the eGFP-EIF4E, and as expected, the signal switched to a green membranous signal meaning that the eGFP-EIF4E has been captured by the tetraspanin CD9 and the GBP1. Finally, the fourth one and the most interesting one is related to a three plasmids transfection: mCherry-CD9 which gave us a red membranous signal, and CD9-GBP1 with eGFP-EIF4E which combined together gave also a green membranous signal colocalized with a red signal visible in the merged signal.

All these results proved that all parts of both fusion protein constructs played their role as expected and allowed us to make a much more complicated design with the best RBP candidate we selected: C-YTHDF1.

B. Basic proof of principle

Right now, all parts of our future design are validated. We need to make our final construct, something flexible enough to be used in different configurations like *in vitro* or *in vivo.*

a. Final design and cloning strategy

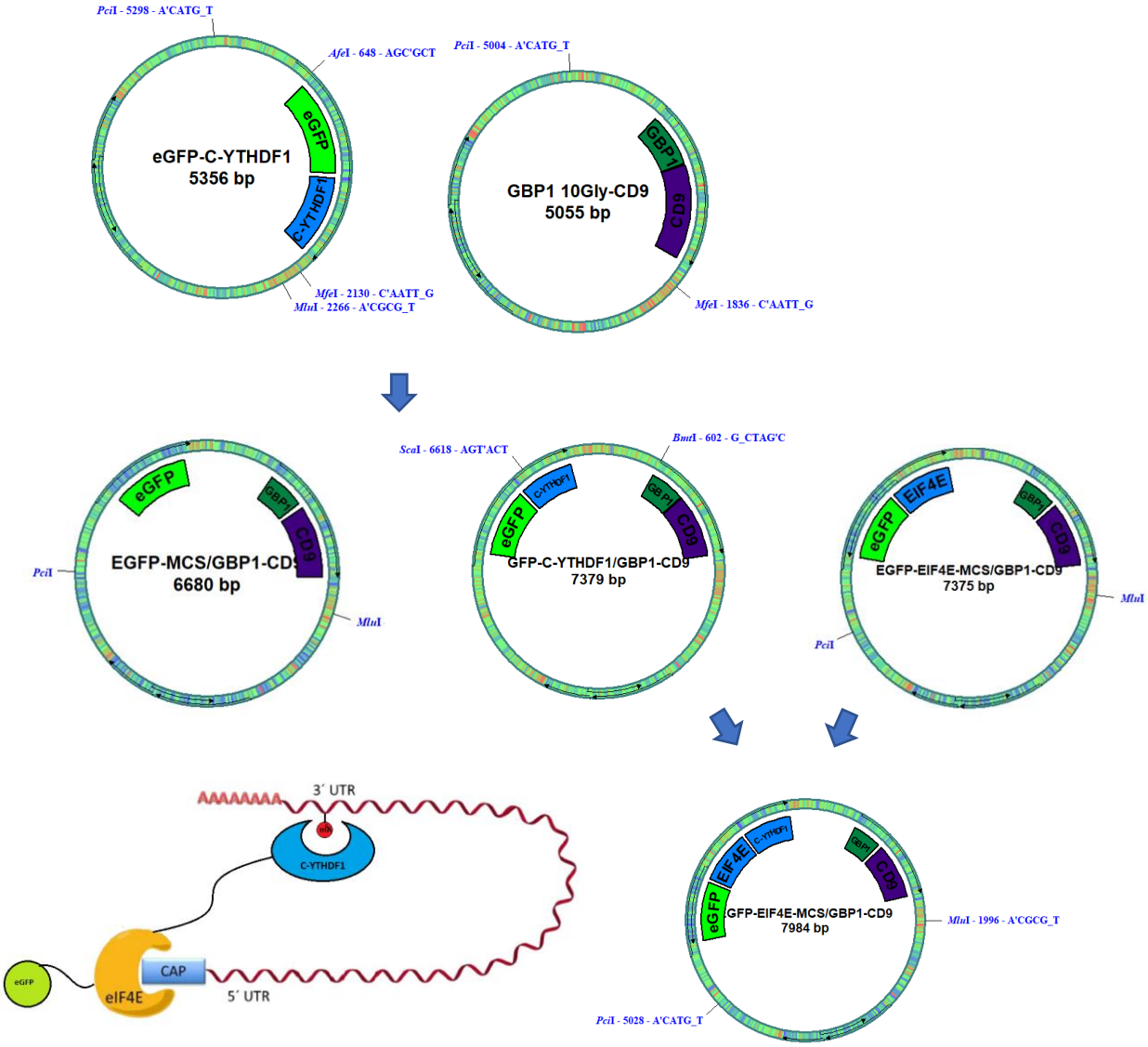
To complete our goals and get a robust and plastic design, we chose a double promoter plasmid design which should give us the best chance to get two proteins produced in two different sides of the cell without interaction.

i. Cloning overview

Indeed, as you probably noted, both of our fusion proteins do not display the same kind of production and function.

In the side of CD9-GBP1, the tetraspanin protein CD9 is related to a membranous production pathway which corresponds to a post translational modification though the endoplasmic reticulum (ER) and the Golgi organelles. On the other side, there is the eGFP-C-YTHDF1 which corresponds to a protein which needs to play its role in the whole cytosol (near the nucleus membrane) and should not interact too fast with the CD9-GBP1. If both fusion proteins interact too fast, both will be transported through the membranous secretion system. We could stay with two separated plasmids, but it would be a big challenge, and it will be almost impossible to get the same plasmid ratio in one cell.

Figure 58. Overall cloning strategy for TRACE-seq



Modified from H. Montero et al. 2015 Viruses

To reply to this issue, we decided to express both fusion proteins from one plasmid. Moreover, to be sure that both proteins do not interact with each other right after they are translated into amino acid chains, we decided to put for each fusion proteins a CMV promoter, so both proteins will be translated into specific mRNAs and will play their role separately from each other. We also decided to produce a big construct corresponding to

both fusion proteins, separated with a T2A auto cleavable peptide [248] and see if it could work or not. This design will be tested in the next chapter to illustrate different potential ideas. We now have two promoters in one plasmid, which does not prevent variabilities in terms of both fusion protein ratio. But we think it will be more critical to have both fusion proteins which can play their respective role and try to take care about the ratio after. To complete our design, we first needed to put the eGFP-C-YTHDF1 (eGFP-C1) fusion in the CD9-GBP1 plasmid. We made the plasmid pCMV-eGFP-C-YTHDF1/pCMV-GBP1-CD9 then by double digestion, we removed the C-YTHDF1 (C1) sequence and obtained the following plasmid: pCMV-eGFP/pCMV-GBP1-CD9 which we plan to use as control (Figure 58).

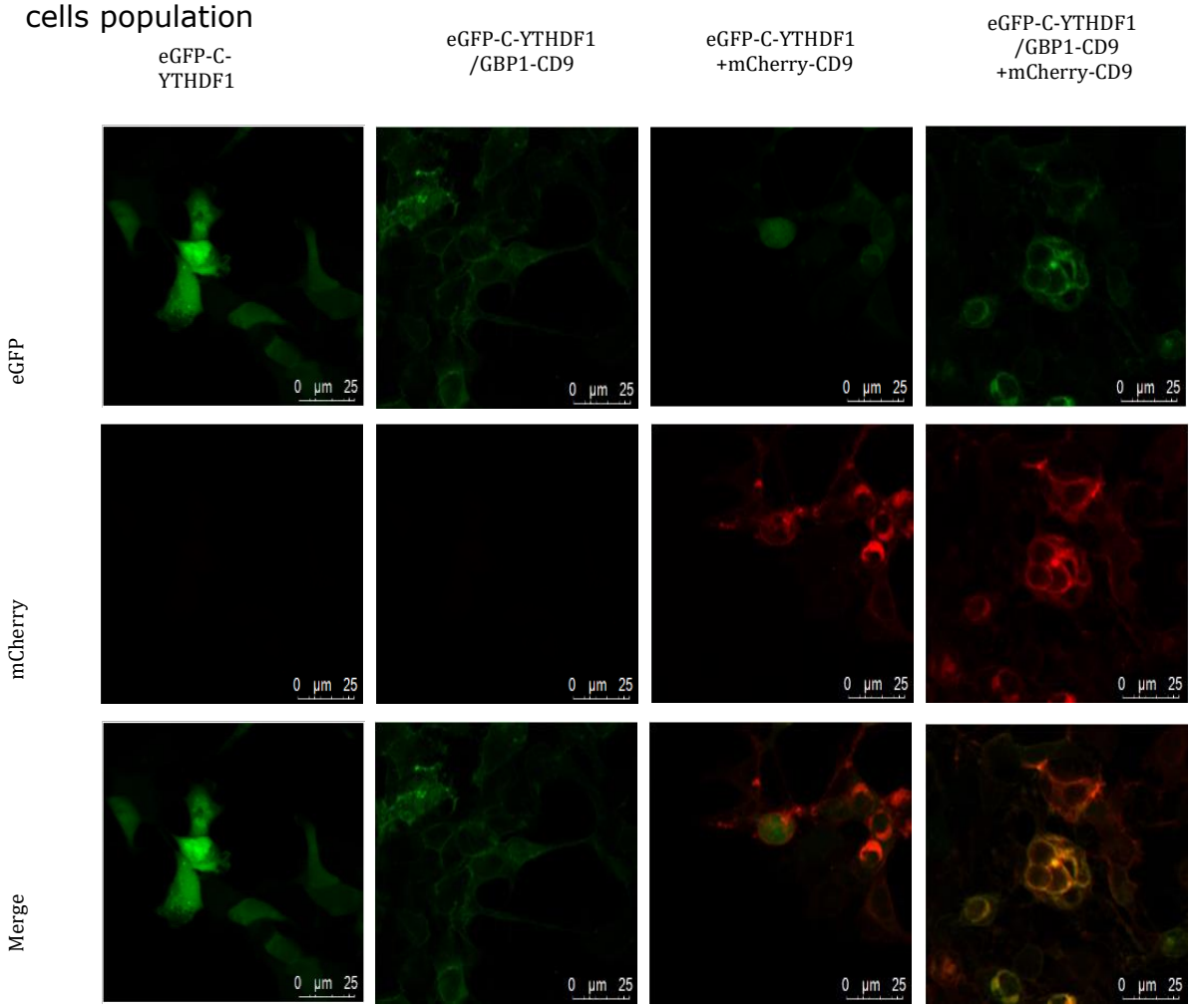
We also decided to add the cap protein EIF4E in the eGFP-C-YTHDF1/GBP1-CD9 to get the eGFP-EIF4E-C-YTHDF1/GBP1-CD9. We thought this design could bind the mRNA in two different sites, in the 3'UTR for the C-YTHDF1 (C1), and the Cap m^7G for the EIF4E, and mimic a sort of loop which could be better for the mRNA vesicles loading.

ii. Basic validation

First, we did the same kind of validation as previously. Thus, we decided to validate by transfection of HEK cells and verify the colocalization of the signal (if all the signal red and green are on the membrane). As shown in figure 59, the same kind of results are observed as the previous colocalization check, meaning that we can see a superposition of both signals on the membrane when two plasmids are transfected together (eGFP-C-YTHDF1/CD9-GBP1 + mCherry-CD9). As expected, the colocalization appears and the final construct design seems to be very efficient in regards of the membranous signal localization.

Secondly, we decided to get to the heart of the project by performing an experiment on EVs mRNA extraction and analysis. To do so, we made the protocol we previously established and transfected different HEK cells (8 million cells) with our construct eGFP-C-YTHDF1/CD9-GBP1, the negative control construct eGFP/CD9-GBP1 and a untransfected cell. After 12 hours in the transfection media, cells were put in new fresh culture media which has a specific depleted exosome free FBS for 24hours (media of EVs collection). To isolate

Figure 59. Confocal signal colocalization check on final construct for different HEK293T cells population

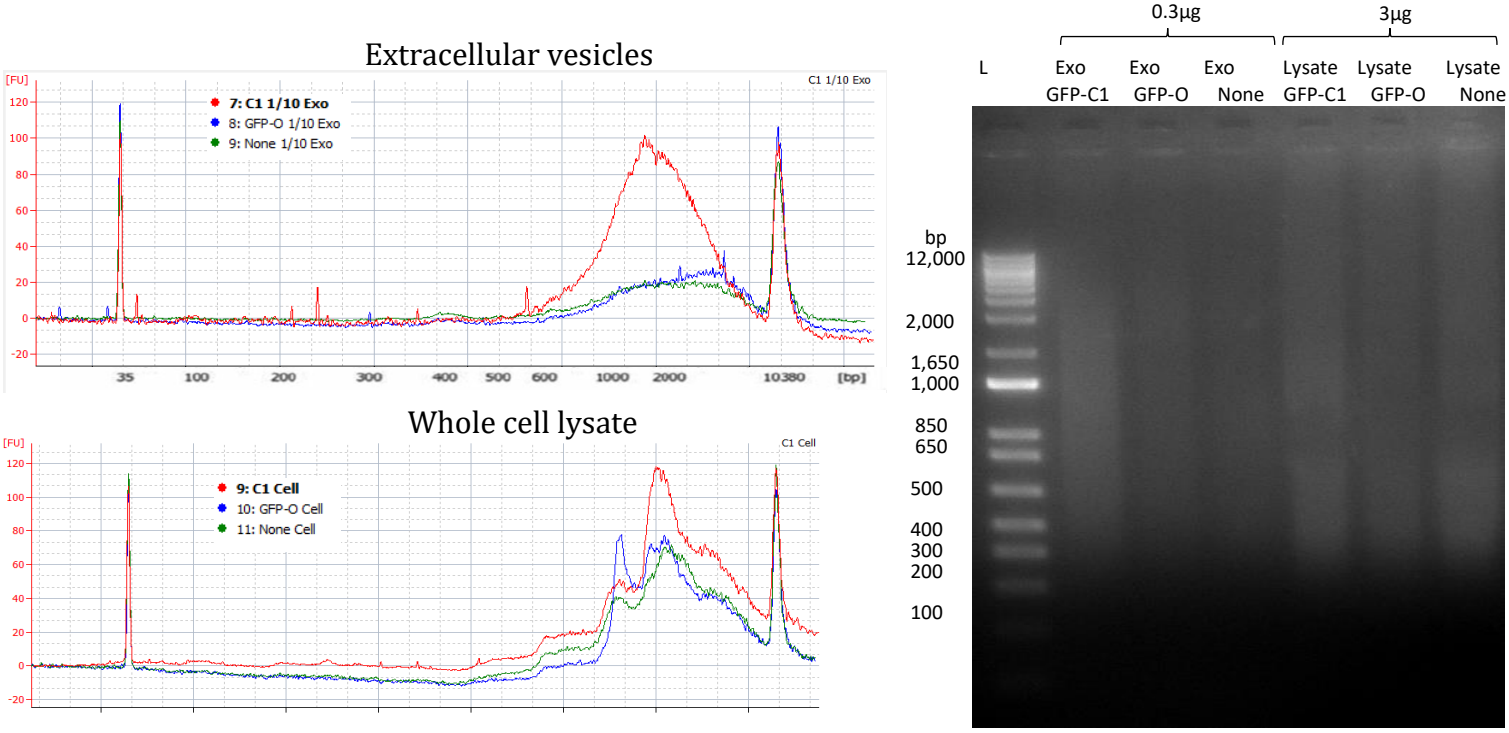


Confocal imagine from Transfection of different HEK 293T populations with two constructs: eGFP-C-YTHDF1, eGFP-C-YTHDF1/GBP1-CD9 and mCherry-CD9. Imagine after 36h post transfection.

vesicles from the cell culture media, we performed one big centrifugation without any filtration and decided to extract mRNA and proceed to the SMART-seq2 protocol to transform mRNA into cDNA and pre-amplify these populations.

We also extracted cell lysates from the corresponding samples in the same way as above and integrated it in our analysis.

Figure 60. EVs mRNA content analysis basic protocol



As presented in Figure 60, we decided to run an agarose gel on each cDNA populations obtained by the extraction and converted with the Smart-seq2 protocol. Moreover, we used a fraction of the samples that we diluted by 10 and ran a High efficiency chip on Agilent Bioanalyzer. As you can see, there was much more cDNA and also full-length mRNA in the EVs in the C1 construct sample even if there was a bit more signal for the corresponding cell

lysate compared to both controls (negative construct control and non-transfected cells). As expected, our construct seems to bring full-length mRNA inside EVs, but we need to have more precise analysis to confirm if this mRNA population is enough representative of the whole cell lysate from cells. Nevertheless, we were not completely in adequation with these data especially regarding the remaining signal we found in the negative EV control samples. We also made the same experiment with a smaller number of cells (500k) and got the same kind of results (see Annexes Figure 7 and 8). We knew by the bibliography that the signal would be very low and we thought that our protocol of EVs purification and mRNA extraction should be improved to avoid as much as possible apoptotic bodies contamination or to be able to discriminate different kind of EVs (exosome vs microvesicles). Despite this point, the results shown here were very promising and we decided to improve our protocol.

The following part is coming from the TRACE-seq manuscript paper and presents the results submitted for publication.

C. Validation and robustness of the technique

a. TRACE-seq, an mRNA translation methodology carried by Extracellular Vesicles

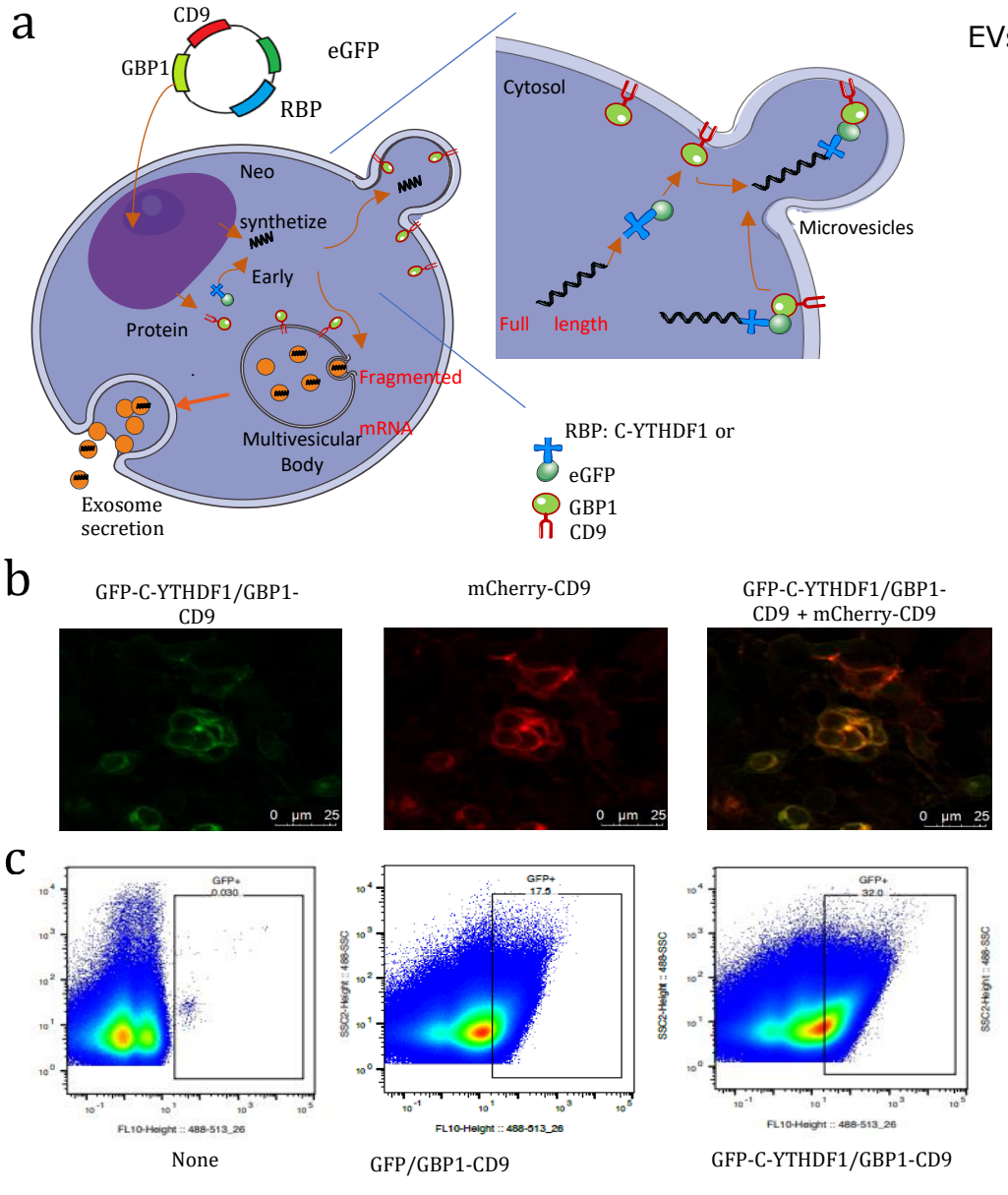


Figure 61. Overview and validation/ EVs purification

a. Overview of the TRACE methodology. A plasmid encoding to the two fusion proteins GBP1/CD9 and eGFP/RBP is transfected or transduced to the cells. Both proteins were translated and play thier role separately. The RBP "catches" m^{6}A tagged mRNA and by the link of the eGFP/GBP1 bring it into the CD9 coated EVs for an extracellular export.

b. Confocal image from Transfection of different HEK 293T populations with two constructs: eGFP-C-YTHDF1/GBP1-CD9 and mCherry-CD9.

c. NanoFACS of the whole vesicle population secreted by different HEK 293T cells. Untransfected cells (None), transfected with the Negative control constructs (eGFP/GBP1-CD9) and with the

All known cell types secrete EVs, and studies demonstrate the presence of RNAs in the cargo of EVs. However, profiling RNA in extracellular vesicles (EVs) does not always provide an

accurate representation of the original cell's transcriptome. Indeed, these RNAs consist of a large proportion of fragmented mRNA in addition to multiple small RNA species [174-178]. This presents challenges in using EV RNA as an accurate proxy for interrogating the cellular/cytosolic transcriptome. We sought to overcome this through the use of a cell-type specific transgene expression that facilitates the packaging of cellular mRNAs import to extracellular vesicles, especially microvesicles (MVs), which are used as carriers (Fig.61). A similar technology already exists, in which researchers used engineered EVs as carriers for protein transportation [179]. Based on this technology, our methodology comprises two separate components:

First, we designed an mRNA "Catcher" which corresponds to a fusion protein composed of a C terminal part of YTHDF1 protein and the fluorescent transmitter enhancer GFP. Recently reported, YTH protein domain recognizes m^6A, one of the most abundant internal modifications in eukaryotic mRNA [180, 181, 249]. Moreover, isoform 1 of the YTH protein domain enhances the translation efficiency and binds to the mRNA close to the nuclear membrane after translocation from the nucleus to the cytosol [182, 183].

Second, we designed another fusion protein that consists of a GFP binding protein 1 (GBP1), which binds to EGFP while enhancing its fluorescent signal [184, 185] and the CD9 protein known to be a canonical EV marker [186]. Thus, the mRNA "Catcher" EGFP-C-YTHDF1 can trap newly synthesized mRNA, then bind to the second fusion protein through EGFP/GBP1 affinity and transport the mRNA into extracellular vesicles via the tetraspanin CD9 protein, as shown in Fig.61 a. To validate part of the construct, we used an additional plasmid as a positive control, mCherry-CD9 (expected to be detected on EVs) and checked for membrane colocalization signal from our TRACE construct (Green) and the positive control plasmid

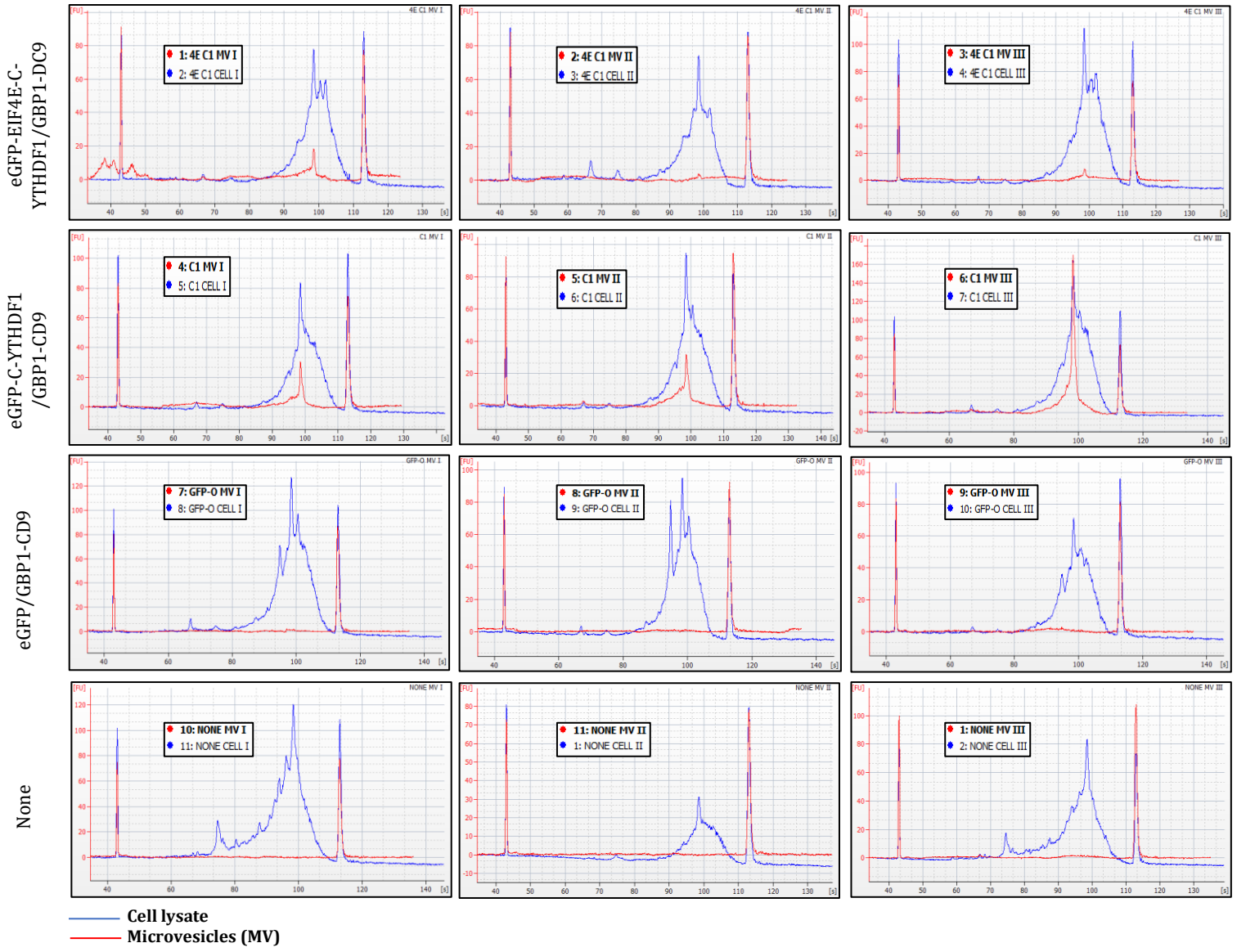
(Red). As expected, we observed this colocalization signal on the cell membrane, as shown in Fig. 61 b. We then purified MVs from the TRACE transfected cells and see if it was possible to detect GFP expressing EVs by High Dimensional Flow Cytometry (also referred as NanoFACS [250]), Fig. 61 c. A significant number of GFP EVs were detected for both cells transfected with the TRACE construct (eGFP-C-YTHDF1/GBP1-CD9) and the negative control construct (no mRNA catcher: eGFP/GBP1-CD9).

b. Detection of the mRNA inside the Extracellular vesicles and reverse transcription

A second validation consisted of analyzing the mRNA content of the TRACE transfected cell EVs. To do so, we designed a protocol for a reproducible and robust EV-mRNA profiling (Supplementary Fig 10), with mRNA library generation based on the SMART-seq2 (Switching Mechanism At 5′ end of RNA Template sequencing) technology [240, 246]. Due to the novelty of this technology, we evaluated different mRNA catchers (Supplementary Fig 11). Three constructs were generated and analyzed. Two have the following design: eGFP-C-YTHDF1/GBP1-CD9 (C1), eGFP-EIF4E-C-YTHDF1/GBP1-CD9 (4EC1), which corresponds to the addition of the cap protein EIF4E to the Cterm-YTHDF1 for a potential improvement to mRNA binding and transportation. A third, eGFP/GBP1-CD9 was also used as negative GFP control. It is important to note that each group of proteins has a specific function (one targets the mRNA and the other brings the complex to the cell membrane) so, each fusion protein is expressed under its own independent promoter. We hypothesized that premature binding of the mRNA targeted fusion protein (the ‘catcher’) to the EV-targeting fusion protein could

direct these complexes into the EV secretion pathway before they could sufficiently 'capture' the cellular mRNA. These 'empty' complexes could drastically reduce the activity of our TRACE-seq system. Using two separate promoters (one per fusion protein) would reduce the potential issue of precocious capture of the eGFP-mRNA catcher by the GBP1-CD9. To confirm this hypothesis, a one-promoter regulated construct: eGFP-C-YTHDF1-T2A-GBP1-CD9 (T2A) was also cloned and compared with the two-promoter system.

Figure 62. Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA from cell lysates (Transient transfected Cells)



Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA from cell lysates isolated from different transfected HEK 293T populations (Triplicate of 4M cells per condition): untransfected cells (None), negative control construct (eGFP/GBP1-CD9) and the two types of TRACE construct (eGFP-C-YTHDF1/GBP1-CD9 and eGFP-EIF4E-C-YTHDF1/GBP1-CD9).

For each of these three constructs and their controls (negative construct control eGFP/GBP1-CD9 and regular HEK 293T cells), RNA from ~4 million transfected cells was analyzed by Bioanalyzer. All samples only showed one peak corresponding to the vesicular miRNA signal (Supplementary Fig 13**)**, except for the T2A construct which has an additional peak in the 28s rRNA region (Supplementary Fig 14. a). All of this RNA was reverse transcribed into cDNA and pre-amplified according to our protocol and analyzed on a High efficiency DNA chip. For the two constructs regulated with distinct CMV promoters, a peak corresponding to full length mRNA was detected (average 1,200nt) which also reflected the corresponding cell lysate cDNA signal, Fig 62 and Supplementary Fig 15. Moreover, NanoFACS analysis of the EVs revealed differences in secretion between the different groups of cells (Supplementary Fig 16). The negative control group produced more EVs than the two constructs, but as full-length RNA is generally not observed in EVs, this concentration difference did not interfere with the downstream RNA analysis.

The cDNA signal from the T2A construct did not show any peak (Supplementary Fig 14. b), suggesting that the generation of the mRNA transcripts for both fusion proteins from separate promoters seems to be a critical step for their anticipated function.

As opposed to microvesicles, it has been observed that the majority of the exosomal RNA is fragmented [176, 177, 202]. We did not focus on these vesicles at first, but for control purposes, we also analyzed exosomal RNA content. As anticipated, we were not able to generate full-length cDNAs from RNAs isolated from these exosomes, as shown in

Supplementary Fig 15, suggesting that the mRNA brought into the multivesicular bodies during exosome biogenesis was previously cleaved [251]. Thus, we decided to focus all of our design and methodology on **Microvesicles** rather than exosomes. We repeated the same kind of experiment with a lower number of cells (1 million transfected cells) and observed a similar cDNA signal as before for the two TRACE constructs with a proportionally lower amplitude (Supplementary Fig 17). To assess the effectiveness of this methodology on a particular pathway, qRT-PCR analysis was performed on metabolism genes using the most promising construct (GFP-C-YTHDF1/GBP1-CD9) and the two-negative controls (GFP/GBP1-CD9 and untransfected cells). Eight million cells were transfected with the constructs using the same protocol as before and the generated cDNA was analyzed by qRT-PCR. The expression of genes in the cell lysate versus their microvesicles is much more tightly correlated in the TRACE group (ddCT close to 0), than in the two control groups (ddCT >3 or -3) (Supplementary Fig 18).

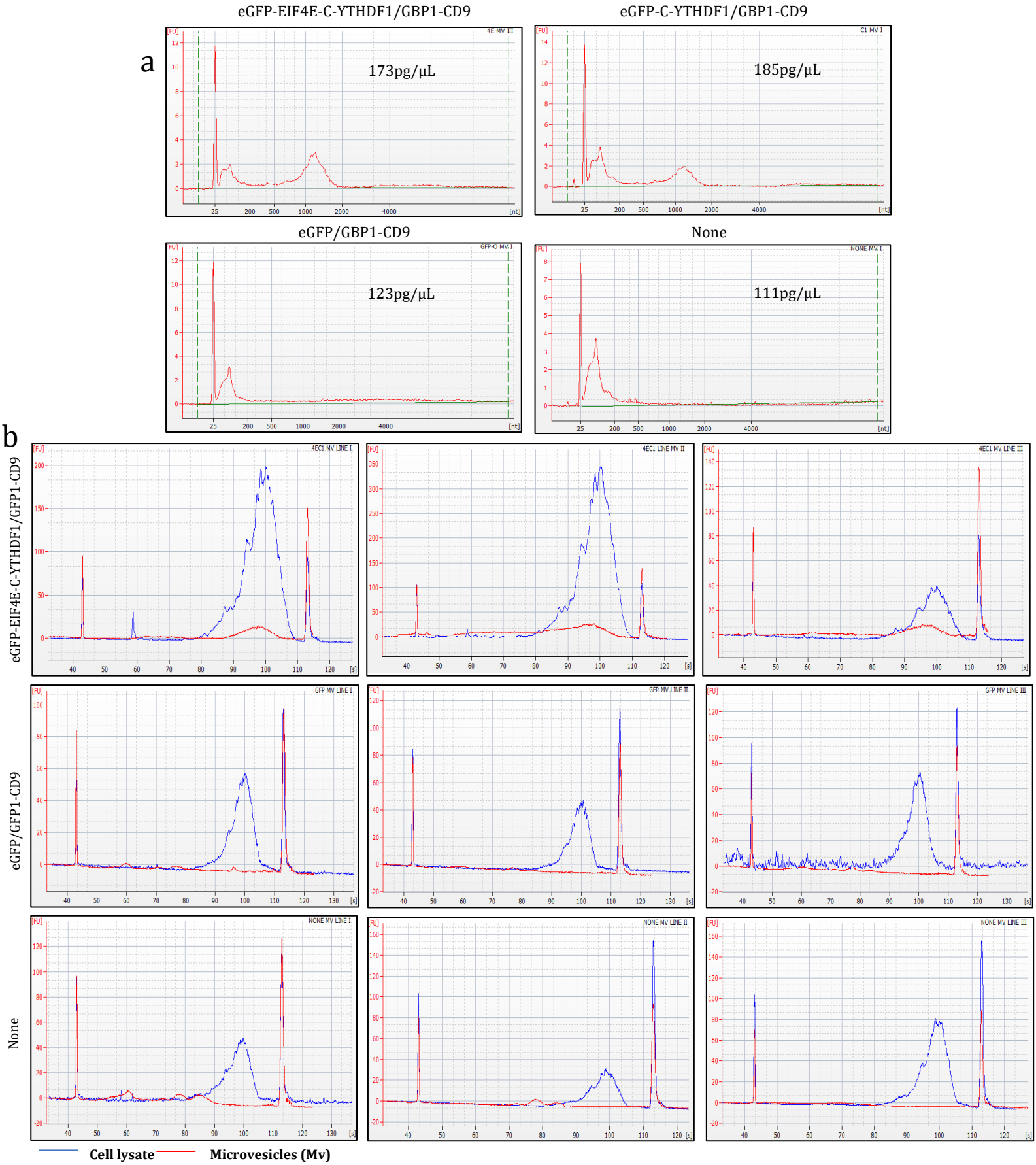
In parallel, we also generated a stable cell line by lentiviral infection for the two successful constructs (eGFP-C-YTHDF1/GBP1-CD9 and eGFP-EIF4E-C-YTHDF1/GBP1-CD9) and the negative control construct (eGFP/GBP1-CD9), inducible by doxycycline. All of the isolated MVs from the stable cell line were analyzed and compared to the regular HEK293T cells MVs and cell lysates (Fig. 63 and supplementary Fig 19). Surprisingly, the 4EC1 construct MVs group was the only one producing a robust detectable signal with amplification of full-length mRNA. We also noticed that from 10M cells, a mRNA peak ≈1,300nt was detected in the RNA pico chip for both constructs (Fig. 63 a).

In order to discriminate the two TRACE constructs, we ran a viability assay and quantified the GFP and GFP1 expression for all of the cell line populations. We found that only the 4EC1

construct induced a reduction of cell growth during long Doxycycline induction (Supplementary Fig 20 b), suggesting that the TRACE-seq technology has a long-term effect on cell physiology. This justifies the importance of having an inducible system, as also shown in Supplementary Fig 20 b; with no doxycycline induction, the doubling stage of the 4EC1 cell line remained at the same level as the control group. Moreover, because we know that the two constructs have two different promoters, the ratio between the two fusion proteins cannot be 1:1. The 4EC1 group exhibited the closest GFP and GBP1 gene expression level by qPCR, compared to the C1 construct (Supplementary Fig 20 c). The ratio of the two parts of the technology is a critical parameter especially for the protein targeting and function, as it could explain the weaker and inconsistent activity of the C1 construct (Supplementary Fig 19).

Given the effects on cellular physiology in the stable cell lines, we focused our RNA profiling efforts on transiently transfected cells. Thus, as robust proof of concept of the TRACE-seq methodology, we decided to select samples from the transient transfection batch (Fig 62) for further RNA-seq analysis.

Figure 63. Bioanalyzer analysis of RNA and cDNA generated from Microvesicles and their corresponding cDNA from cell lysates (Transduced Cells)

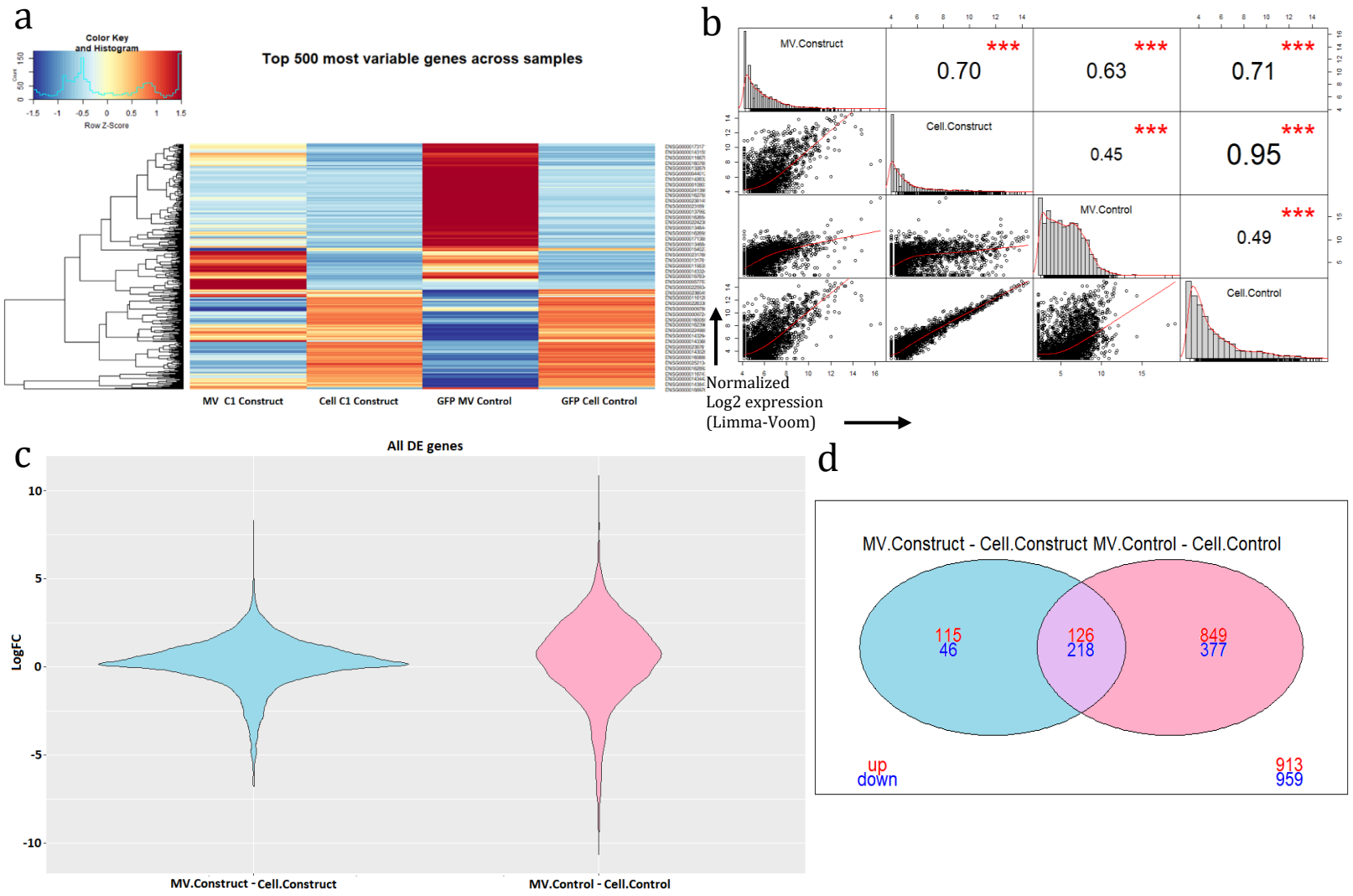


a. mRNA extract Cell line created with the construct 4EC1: eGFP-EIF4E-C-YTHDF1/GBP1-CD9, the negative control: eGFP/GBP1-CD9 and the regular HEK293T cells: None. Experiment made on 10M cells. **b.** from Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA generate from cell lysates from Cells lines (Same as above), Triplicate of 8M cells per condition.

c. The transcriptome in TRACE-Seq MVs is representative of the cellular transcriptome

To further validate the technology, we sequenced the RNA samples from the transiently transfected cells (C1 MV III, C1 MV II, C1 Cell III, C1 Cell II, GFP MV I, GFP MV II, GFP Cell I, GFP Cell II, from Fig 62), in which we focused on a selection of 3603 genes.

Figure 64. DEseq2 analysis ran on the 8 selected samples C1 construct TRACE and GFP control



DEseq2 from the 8 selected samples TRACE GFP-C-YTHDF1/GBP1-CD9 construct: duplicate MV, duplicate cell lysates and Control GFP/GBP1-CD9: duplicate MV duplicate cell lysates from transient transfection batch. Each duplicate was regrouped here, for ungrouped results see annexes fig **a.** Heat map for the most 500 variable genes across all samples (duplicates regrouped to generate the heatmap). **b.** Pearson correlation chart form the Limma-voom DEseq2 analysis between each group of samples. **c.** All genes from the Limma-voom DEseq2 analysis. **d.** Venndiagram from contrast matrix MV-Cell lysates from genes used in Limma-voom DEseq2.

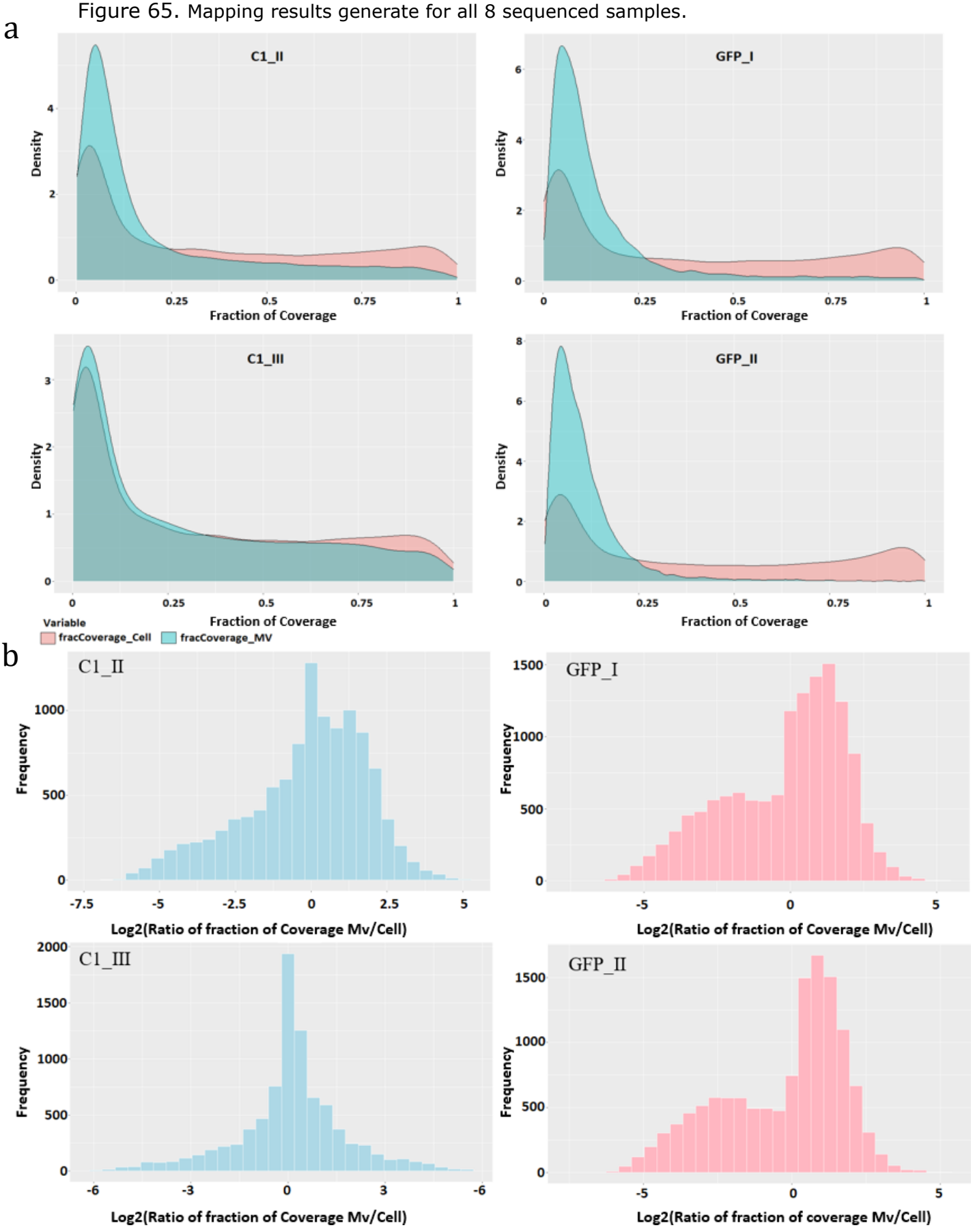
MVs from TRACE-seq C1 construct contain a non negligeable population of genes expressed at the same level as for the cell lysates (Fig 64 a). In contrast, the MV RNA content from the GFP control group showed much more of a difference with its related cell lysate. Almost the totality of the pool of the 500 most variable genes are, in terms of level of expression, different compared to cell lysate (Fig 64 a). The MV C1 construct samples are much more correlated with the cell lysates (Pearson correlation = 0.7; Fig 64 b) while expression in the GFP MV control is not well-correlated with those detected in the GFP cell lysate (Pearson correlation = 0.45). In the results in Fig 64 c, the log foldchange of all genes in DEseq2 (3603 genes) is much more correlated for MV C1 construct -C1 Cell lysates construct (with a LogFC close to 0) compared to the MV GFP control-GFP Cell lysate control (which has LogFC values above 1/-1 for half of genes). Unsurprisingly, the differential expression between MVs and their own cell lysates is much higher in the GFP control group than with the C1 construct, with 505 and 1,570 differentially expressed genes, respectively (Fig 64 d). In the GFP control 62% of genes were significantly more abundant in the EVs than in their cell lysates (975 up and 595 down), compared to 47% of genes in the C1 construct (241 up and 264 down). The significant enrichment of over-abundant genes in the control EVs (fisher exact p-value=10-11) compared to the C1 construct (fisher exact p-value=0.49) highlights the more equal/stochastic mRNA loading enabled by TRACE-seq and an EV transcriptome that is a better mirror of that of the cell.

We sought to understand whether the correlation between MVs and cell lysate would increase focusing on m^{6}A YTHDF1 IP isolated genes. Similar correlations were found here using genes identified in the POSTAR database of RNA-seq isolated by IP with C-YTHDF1 antibody on Hela cells (CLIP-seq) [252, 253], however the C1 construct showed slightly

better correlation (Supplementary Fig 21 a,b,d). These results show that even if we compare our TRACE-Seq isolated gene with a m^6A YTHDF1 IP gene population, the overall correlation TRACE-Seq MVs vs whole cell lysate or TRACE-Seq MVs vs POSTAR remains the same and comparing TRACE-Seq MVs vs their own cell lysate did not bring a lot of bias. C1 vs GFP control differential expression between both MV populations revealed many more genes that differ between both groups rather than genes in common, with 2646 different genes vs 957 genes in common (Supplementary Fig 21 c).

The coverage distribution between the MVs and their own cell lysate (Fig 65 a) is in much closer proximity with C1 Construct than the Control group. This is clearly evident in Fig 65 b, which shows on a logarithmic scale, the ratio of mRNAs in cell lysates versus MV (MV/Cell lysate). Notably, both C1 construct groups are very close to 0, meaning that both MV and cell lysate share a majority of very correlated fragments with the same coverage. In contrast, the GFP MV results are much further dispersed from 0 on this logarithmic scale coverage distribution for MV/Cell mRNA ratio, meaning that mRNAs in GFP MVs and GFP cells have disparate expression. As noticed in Fig 64, it is clear that the TRACE-seq technology transforms the mRNA EVs cargo content by the import of a non-negligible number of full length mRNA. Fig 65 shows a clear difference in terms of coverage, suggesting that the TRACE construct brings full length mRNA inside MVs with significant correlation noted between the MV and the cell lysate construct RNAs for the C1 construct. All these results taken together suggests that the TRACE technology can be used for a transcriptome analysis in order to monitor live cells.

Figure 65. Mapping results generate for all 8 sequenced samples.



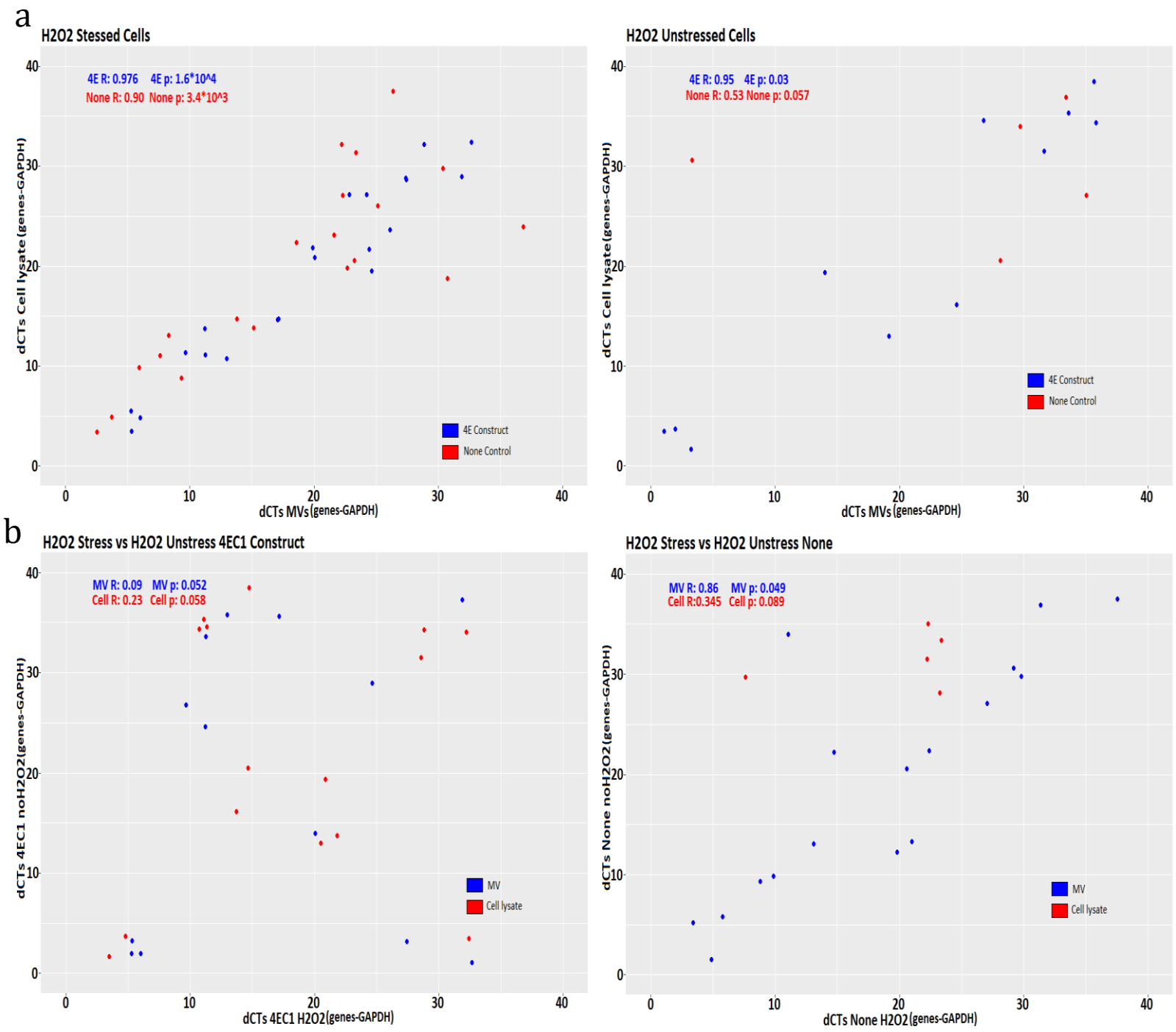
a. Mapping of %coverage distribution MV and Cell lysates on genes at ≥1000nt length. **b.** Mapping of %coverage MV/Cell vs Cell/Cell distribution on genes at ≥1000nt length.

d. Following specific gene expression from H_2O_2 stressed cells over time with TRACE

To finalize the proof of principle of the relevance of our new TRACE-seq methodology, we induced H2O2 stress to our GFP-EIF4E-C-YTHDF1/GBP1-CD9 HEK 293T cell line and a regular HEK 293T for 24 hours. In order to analyze the oxidative expression pathway, we designed a qRT-PCR analysis on traditional oxidative genes (Tnf-α, TAT, IL-6, PEPCK, GCKR, HFE). We also included mitochondrial oxidative associated genes (CYP1B1 and CYTC) as the M^6a modification in mitochondria is also associated with the YTH protein family member [254]. As presented in Fig.66**,** the 8 oxidative associated genes tested in the MVs from the GFP-EIF4E-C-YTHDF1/GBP1-CD9 cell line are more correlated to their own expression in cell lysate compared to the untransfected HEK cells in samples from both H2O2 treated or untreated cells (respectively 4E Construct; H2O2 R:0.976, noH2O2 R: 0.95 and None; H2O2 R:0.90, noH2O2 R: 0.53). As shown in Supplementary Fig 22 a,b,c, the Normalized gene expression (treated or untreated cells), appears, as expected, very close from MVs to cell lysate regarding the GFP-EIF4E-C-YTHDF1/GBP1-CD9 cell line.

Moreover, if we look at the correlation expression (Fig. 66 b) “with vs without” the H2O2 treatment effect, the variation of gene expression due to the stress induction is clearly detected and consistent in both cell lysates and MVs from the TRACE-seq group compared to the control group in which the Control MVs do not reflect the variation of expression seen within the cells (respectively 4E Construct; MV R: 0.09, Cell lysate R: 0.23 and None; MV R: 0.86, Cell lysate R: 0.345).

Figure 66. H_2O_2 stress pathway validation test. RT-qPCR results generated from MVs mRNA and their corresponding cell lysate.



H2O2 stress pathway validation test. Correlation plot obtained from QPCR results generated from MVs mRNA isolated from control cells (None), TRACE construct (eGFP-EIF4E-C-YTHDF1/GBP1-CD9) cell line on 8 genes corresponding to the oxidative stress pathway. **a.** Batch of H2O2 stressed cells and Batch of unstressed cells correlation Cell lysates vs MV. **b.** Correlation plot H2O2 vs nonH2O2, batch corresponding to 4EC1 Construct and second on to the None control cells. Results obtained from triplicate cell populations, 15M cells per condition.

This demonstrated that dynamic changes in the mRNA profiles of cells are much better reflected in the MVs in the cells with the MV TRACE-seq construct compared to the control MVs and cell lysate (Fig 66 b). These results show that the TRACE-seq technology can be used successfully to monitor the oxidative stress gene expression pathway over time with a very simple, quick and reproducible technique such as RT-qPCR.

2. Discussion

TRACE-seq enables monitoring cells in vitro by collecting media and analyzing MV mRNA content. Our methodology brings a representative part of the cellular transcriptome into MVs which is a useful and non-destructive method for a large range of analyses including monitoring drug testing, assessing cellular maturation or differentiation, or analyzing cellular response to stress. Traditional transcriptome analysis methodologies require cellular lysis; with the exception of the NEX process [129] which is designed exclusively for in vitro living cell monitoring (on a cell culture plate) and not applicable *in vivo*. Thus, our study represents the first detailed analysis of a new methodology, that is fully compatible for in vivo monitoring of the cell transcriptome.

We have been focusing on MVs populations rather than smaller vesicles like exosomes, as it gave us better signal regarding the full-length mRNA content. Our study does not allow to clearly confirm if exosomes transport or not full-length mRNA.

But as already found in other studies, the YTHDF2 protein interacts with the CCR4-Not1 complex [230], which may be one explanation for why our TRACE fusion proteins are not able to pack full length RNA into CD9 bearing exosomes. This particular point highlights characterization and discrimination of exosomes and MVs which are very different in terms of secretion pathways and cannot be considered as a single population.

V. Conclusion

Choices made for the overall strategic design of TRACE-seq were good. Indeed, only 2.5 years of work were necessary to complete this work and validate a brand-new technology in the sequencing field. It is also a completely new methodology which provides a completely innovative approach of the transcriptome analysis in living cell over time. Thus, it currently does not exist this kind of technology and we are confident to publish the TRACE-Seq manuscript in a high impact factor journal.

Moreover, a very important point remains which is the progress made by the last single cell sequencing technology. Indeed, this is with the concourse of this progress that it was possible to develop TRACE-seq. Our imported mRNA input is so low - no matter the purification technique itself - that a pre-amplification step is required. The cDNA generation is also a critical step which provides us through the SMART-Seq 2 protocol the certitude to generate full-length mRNA with the help of the LNA containing TSO (template switch oligos)[245].

Nevertheless, this is important to note that this technique requires a large quantity of starting material which is not compatible with all studies. In terms of *in vivo* experiment, it might be difficult to analyze with a best quality of signal especially for certain small animals like mice. In this context, pooling multiple experimental animals into one sample might be the best solution. Moreover, realizing the overall TRACE-seq protocol required a certain background and skills in RNA extraction and sequencing library preparation. All those steps present critical key points which might be difficult to transpose to routine or might not be compatible with a large amount of sample preparations.

Even if some groups developed approaches which are in fact quite similar by the design, a full length methylated m^6A mRNA imported through vesicles for a monitoring of living cell which could be adapted *in vivo* is a first and a very powerful tool for research groups as example in tumor monitoring.

General Conclusion

Determining the possible key factors and their time points which occur during the CM-hPSCs differentiation process is critical for the future. Because of the lack of therapeutics, patients suffering from MI usually just have the possibility to receive an allograft. In fact, the loss of CMs during the MI event is an irreversible process and the renewing of CMs is ineffective.

To circumvent the problem, researchers developed new strategies and innovative therapeutics to produce CMs *in vitro*. The most promising one came from the hPSCs. Indeed, differentiated viable CMs from hPSCs would be an inexhaustible source of CMs including patient specificity, so not or less possibility of immune rejection [255].

Nevertheless, these new therapeutics need to be clarified and more work and progress need to be made to move them towards a real competitor of the actual graft process.

This is for these reasons that my work in the Domian's Laboratory in Boston in partnership with the Larghero's laboratory in France is important. Indeed, developing new methodologies to better understand the critical key process/factors during the differentiation from hPSCs to CMs is very urgent.

Thus, I started working on the Chic-seq technology which was a very promising source of investigation regarding the TFs role during the myocardial differentiation. The overall development process of this technology was a long and treacherous process but has resulted in a usable technique which answered the objectives.

Thus, our “Chic-loop” design was very effective and through its elegant shape would have allowed us to characterize the early myocardial differentiation pathway.

But the destiny is sometimes capricious. Even if the discovery of a patent in all points similar to our Chic-seq technology was a very unpleasant moment, this was not the end of the world fortunately. PhD thesis is a long and difficult process which is crucial for young researchers to give them the arms to react in those kinds of situations. Thereby, it was at the end a fortunate event which gave me the opportunity to create and develop totally by myself a brand-new sequencing technology. TRACE-seq give the chance to researchers to have a completely new tool to analyze living myocardial stem cells differentiation without any destruction and is also a good technique to improve myocardial graft monitoring in animal models and may contribute to elucidate steps which allow the process of rejection.

This is with a very fast process of development supported by a rigorous bibliography research that a very innovative design emerged. Even if I made all the research part including the sequencing analysis alone, which was very time consuming and difficult in many ways, I came up with this satisfying thesis and I am proud of this big challenge. We hope to well publish this TRACE-seq manuscript, but for me the most important part of my work was the intellectual journey realized. Thus, it is for this reason that I decided to include the Chic-seq development as the first part of this thesis because it gave me an incredible boost and enough confidence in myself to believe in my ideas. Even if we were not the first to develop and propose this very innovative method, we saw at the end that I got a working design in a short period of time and very close to the one developed by a well-known company which is very satisfying.

At the end, there is no fail in research but just setbacks which help us to move forward and always propose new replies which open to more questions. To me, this is the most beautiful part of our work: it never ends…

Perspectives

Regarding the first thesis project, being able to offer a new technology for a multiplexing detection, characterization and correlation of TFs associated with sequencing (even if it was not developed by our team) is a big step forward for research. It could be adapted in very different contexts and aspects such as the characterization of any differentiation process as well as used to better appreciate metabolic pathways. Even though this TAM-ChIP® technique remains very expensive, its multiplexability is a precious asset for research. It opens doors to get an overall/global vision of the TFs synergy and association during stress event.

Concerning my second thesis project. Even if the TRACE-seq technology requires a large amount of starting materials, we expect TRACE-seq to have a broad applicability, to be an appreciable tool for the transcriptomic analysis field and to improve our understanding of gene expression and regulation. TRACE-seq analysis could be applied to *in vivo* studies (e.g. for monitoring of transplanted organs transfected with the TRACE-seq inducible constructs) through the analysis of EVs isolated from liquid biopsy and sorted by NanoFACS [219]. The TRACE-seq could be also useful in stem cells differentiation monitoring to permit a better understanding of the critical key parameters playing a role during the commitment of cells in specific differential pathway, or to study the response of tumor cells to treatments over time.

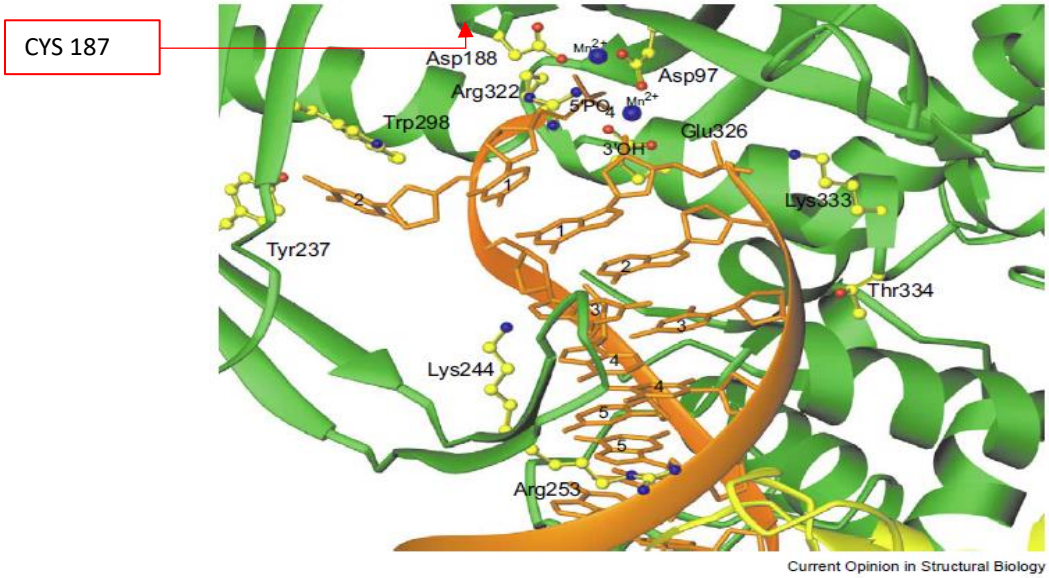
Besides, TRACE-seq technology offers possibilities to get a relative reliable window on the transcriptome of targeted cells without any major source of disruption at different time points which makes this method very interesting in a large *in vivo* animal study where performing a biopsy or sacrificing an animal will not be the most adequate manner to get such material for a transcriptome analysis.

Thus, a very remarkable study could be to monitor tumor expression in various contexts of tumor repressor treatment compatible with time scale experiment study. This kind of study could be made for large screening experiments *in vitro* before a scale up *in vivo* in large animals for a next validation step. The TRACE-seq could be also useful in stem cells differentiation monitoring to permit a better understanding of the critical key parameters playing a role during the commitment of cells in specific differential pathways such as the cardiomyocyte differentiation.

Annexes

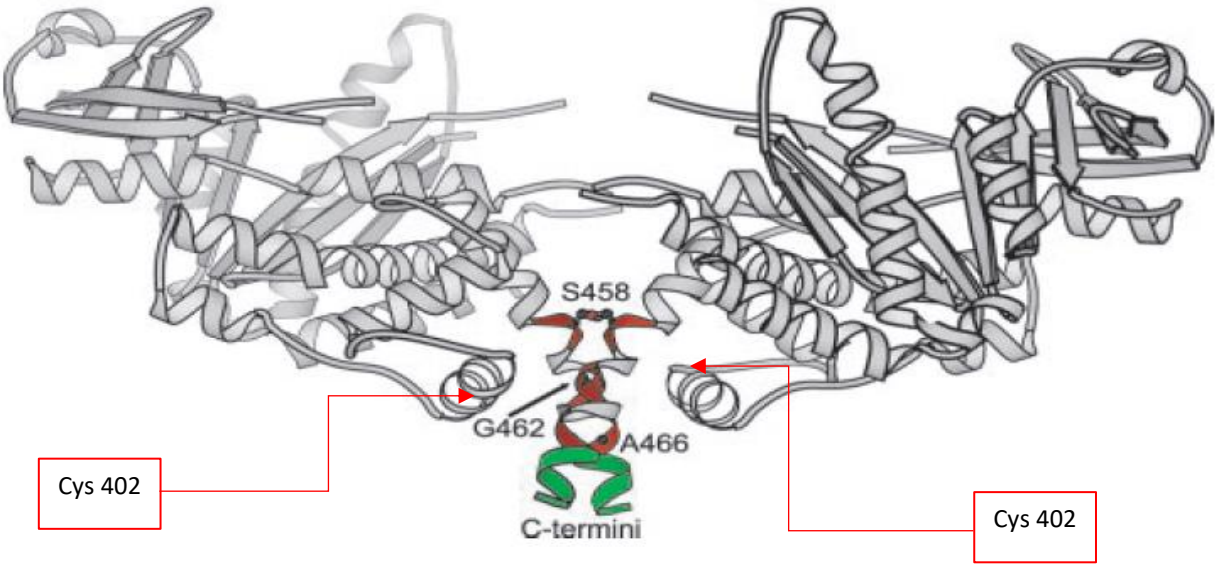
I. Annexes Tn5 Project

Annexed Figure 1. Catalytic site of the Tn5 with the Cys 187



Steiniger-White, M. et al. Curr Opin Struct Biol, 2004[168]

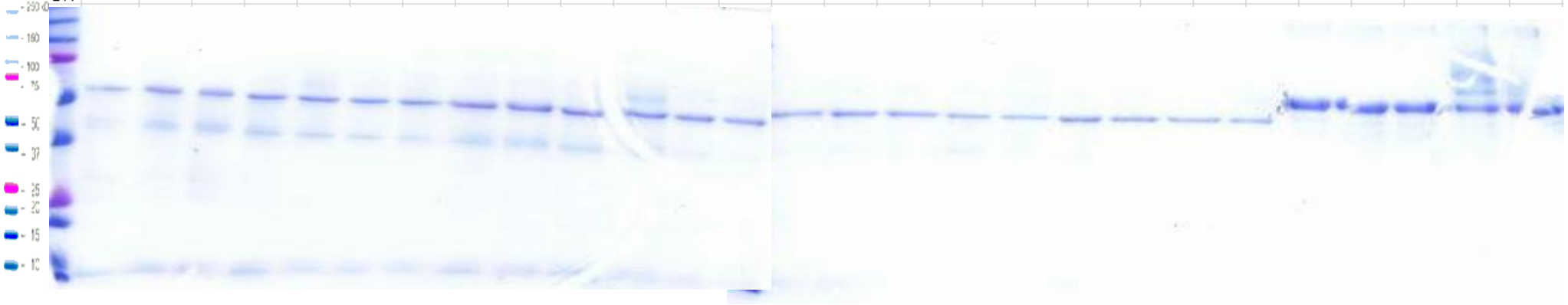
Annexed Figure 2. Dimerization site of the Tn5 with the Cys 402



Modified from: Steiniger-White, M. Reznikoff, W. S. J Biol Chem, 2000. [169]

Annexed Figure 3. Multiple binding reaction with Mal-alexa488 Mal-PEG, Cys-PEG and NHS-PEG on different Tn5 preparation SDS signal

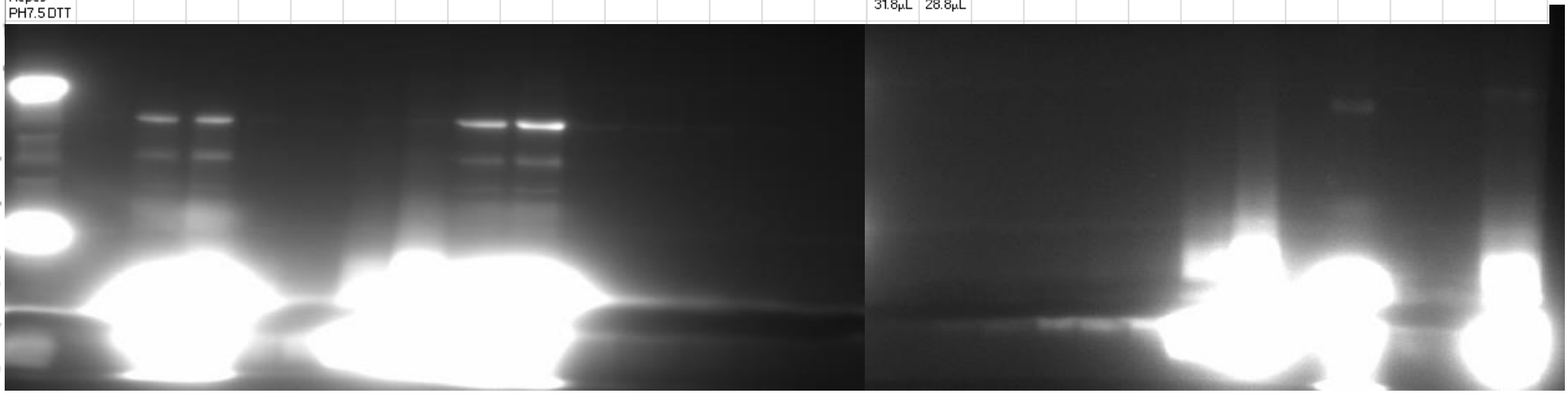
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
BSA 10 mg/ml																								3µL	3µL	3µL	3µL	3µL
Tn5 DTT 0.7 mg/ml	8µL 5.6µg														8µL 5.6µg													
TN5 Bis Tris TCEP 0.35 mg/ml		14.3µL 5µg 2µM	14.3µL 5µg 2µM	14.3µL 5µg 2µM	14.3µL 5µg 2µM	11.4µL 4µg 2µM	11.4µL 4µg 2µM																					
Tn5 Hepes TCEP 0.55 mg/ml								12µL 6µg 2.5µM	12µL 6µg 2.5µM	12µL 6µg 2.5µM	12µL 6µg 2.5µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM			7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM					
Mal Alexa 488 (10mg/ml)		0.1µL 1:50	0.3µL 1:100					0.1µL 1:50	0.3µL 1:100																0.2µL 1:50			
Mal PEG 4mM				1µl 1:50	4 µL 1:200					1µl 1:50	4 µL 1:200															3 µL 1:50		
NHS PEG 4mM												1µl 1:50	4 µL 1:200	4 µL (1:200)		1µl 1:50	4 µL 1:200	1µl 1:50	4 µL 1:200	1µl 1:50	4 µL 1:200						3 µL 1:50	
Mal PEG FITC 4mM						1µl 1:50	4 µL 1:200															1µl 1:50	4 µL 1:200					3 µL 1:50
Bis Tris PH6 TCEP		25.4µL	24.55µL	24.7µL	21.7µL	27.6µL	24.6µL																					
Hepes PH7 TCEP	32µL							27.7µL	26.85µL	27µL	24µL				32µL									37µL	39.8µL	37µL	37µL	37µL
Hepes PH7.5												31.8µL	28.8µL	36µL								31.8µL	28.8µL					
Hepes PH8.5																		31.8µL	28.8µL									
Hepes PH8.5 DTT																				31.8µL	28.8µL							
Hepes PH7.5 DTT																31.8µL	28.8µL											



Blue: Proteins tested, Red: reagents (PEGs…), Black: Buffer and pH condition

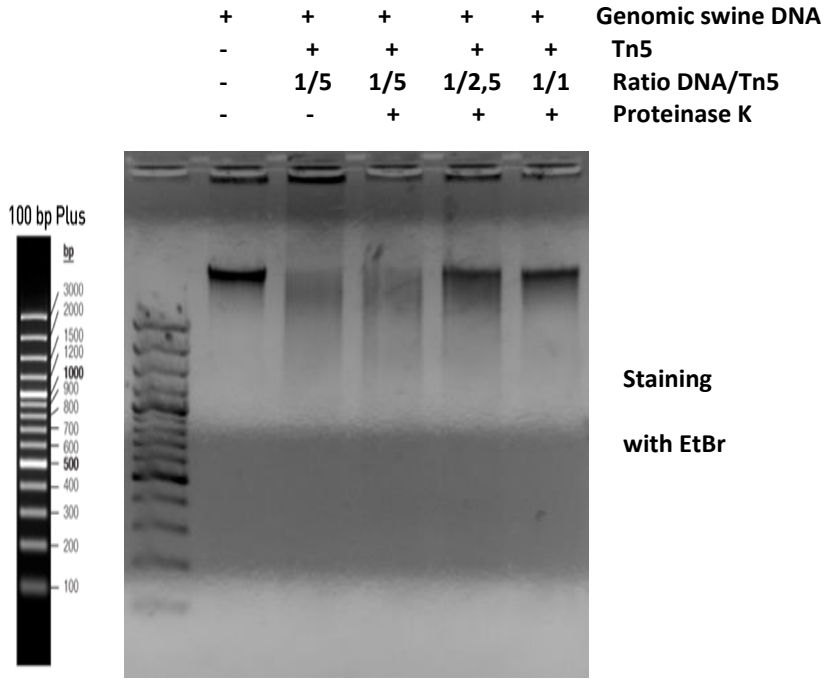
Annexed Figure 4. Multiple binding reaction with Mal-alexa488 Mal-PEG, Cys-PEG and NHS-PEG on different Tn5 preparation UV signal

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
BSA 10 mg/ml																								3µL	3µL	3µL	3µL	3µL
Tn5 DTT 0.7 mg/ml	8µL 5.6µg														8µL 5.6µg													
TN5 Bis Tris TCEP 0.35 mg/ml		14.3µL 5µg 2µM	14.3µL 5µg 2µM	14.3µL 5µg 2µM	14.3µL 5µg 2µM	11.4µL 4µg 2µM	11.4µL 4µg 2µM																					
Tn5 Hepes TCEP 0.55 mg/ml								12µL 6µg 2.5µM	12µL 6µg 2.5µM	12µL 6µg 2.5µM	12µL 6µg 2.5µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM			7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM	7.2µL 4µg 2µM					
Mal Alexa 488 (10mg/ml)		0.1µL 1:50	0.3µL 1:100					0.1µL 1:50	0.3µL 1:100																0.2µL 1:50			
Mal PEG 4mM				1µl 1:50	4µL 1:200					1µl 1:50	4µL 1:200															3µL 1:50		
NHS PEG 4mM												1µl 1:50	4µL 1:200	4µL (1:200)		1µl 1:50	4µL 1:200	1µl 1:50	4µL 1:200	1µl 1:50	4µL 1:200					3µL 1:50	
Mal PEG FITC 4mM						1µl 1:50	4µL 1:200															1µl 1:50	4µL 1:200					3µL 1:50
Bis Tris PH6 TCEP		25.4µL	24.55µL	24.7µL	21.7µL	27.6µL	24.6µL																					
Hepes PH7 TCEP	32µL							27.7µL	26.85µL	27µL	24µL				32µL									37µL	39.8µL	37µL	37µL	37µL
Hepes PH7.5												31.8µL	28.8µL	36µL								31.8µL	28.8µL					
Hepes PH8.5																		31.8µL	28.8µL									
Hepes PH8.5 DTT																				31.8µL	28.8µL							
Hepes PH7.5 DTT																31.8µL	28.8µL											

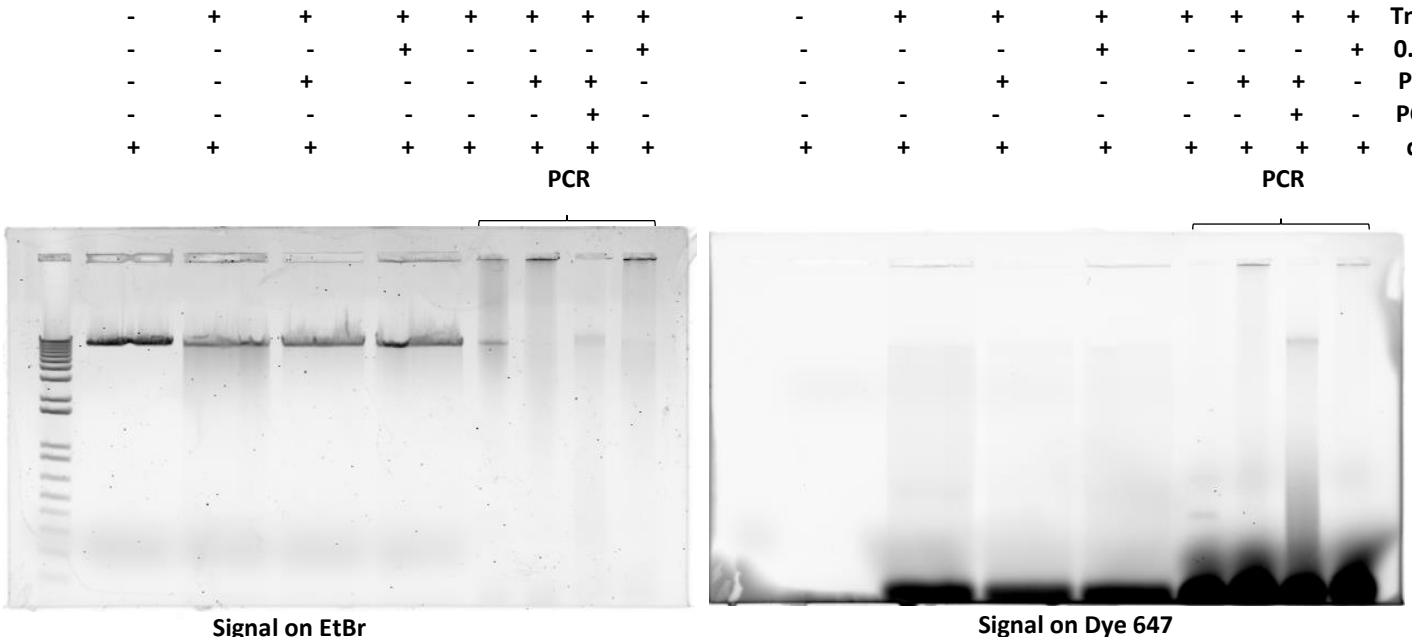


Blue: Proteins tested, Red: reagents (PEGs…), Black: Buffer and pH condition

Annexed Figure 5. Tn5 vs DNA with fixed genomic DNA quantity and increased of Tn5 quantity

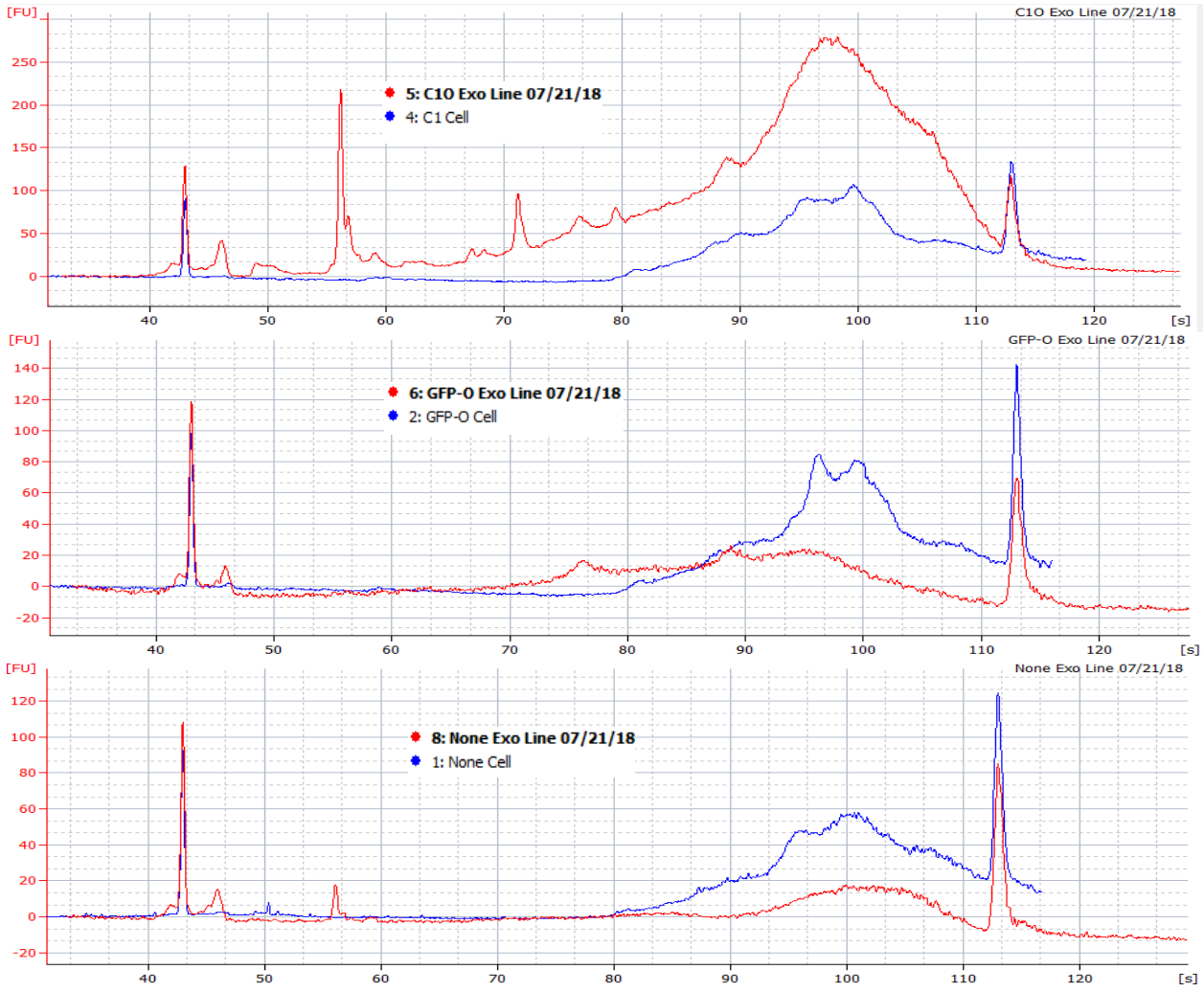


Annexed Figure 6. Tn5 vs DNA plasmid followed by a PCR amplification with fluorescent oligos alexa647

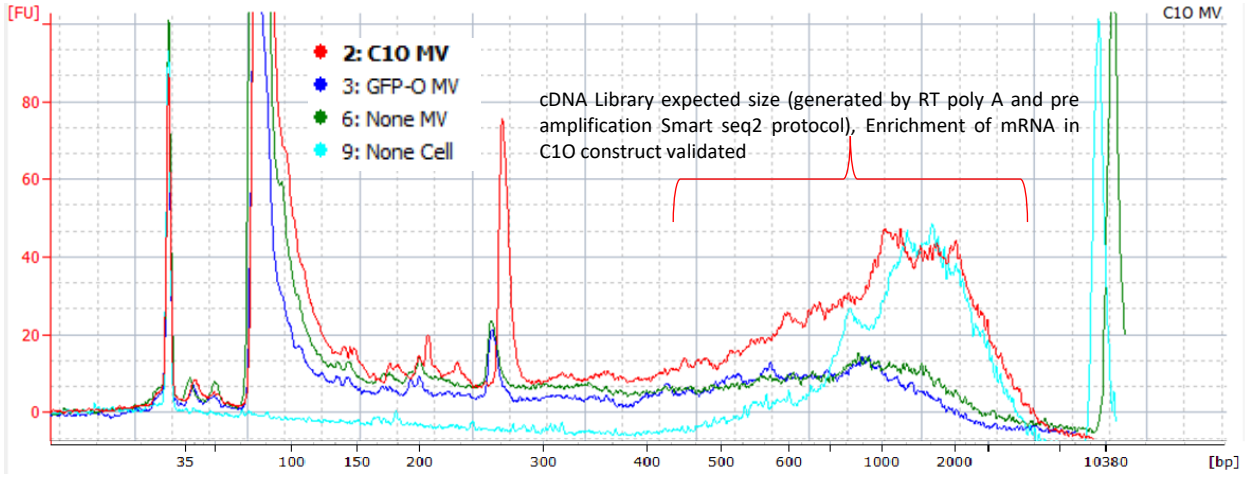


II. Annexes TRACE project

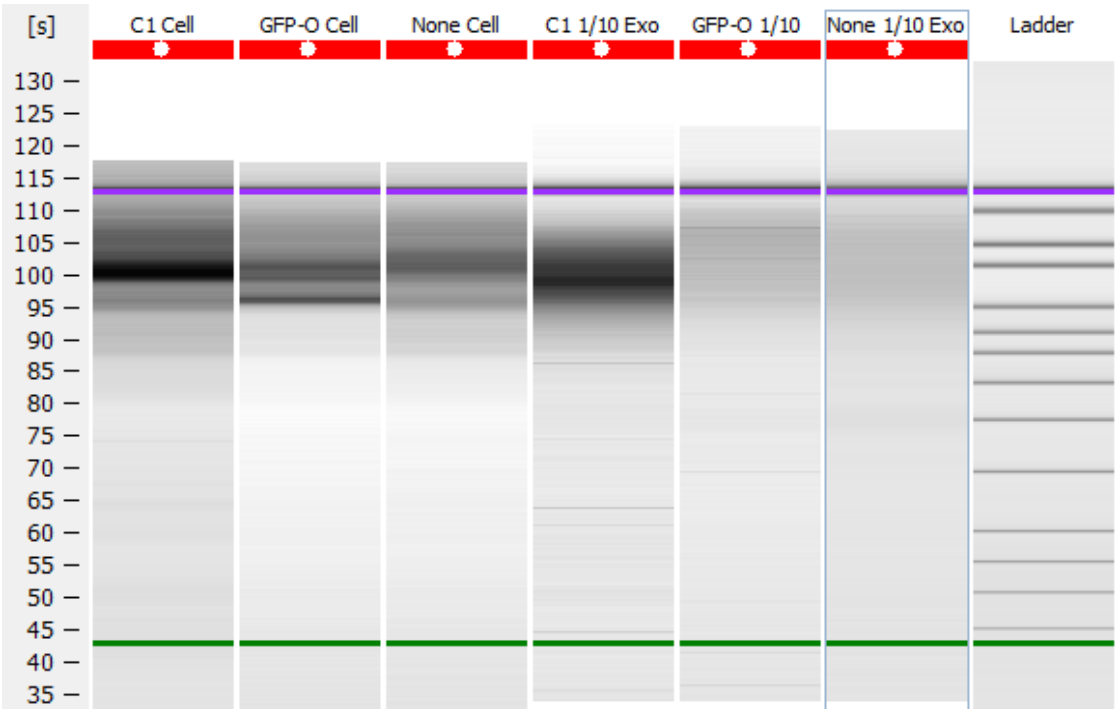
Annexed Figure 7. cDNA chip analysis from transfected HEK cells (500k cells) with the Construct C1, The negative control construct, GFP-O and untransfected cells.



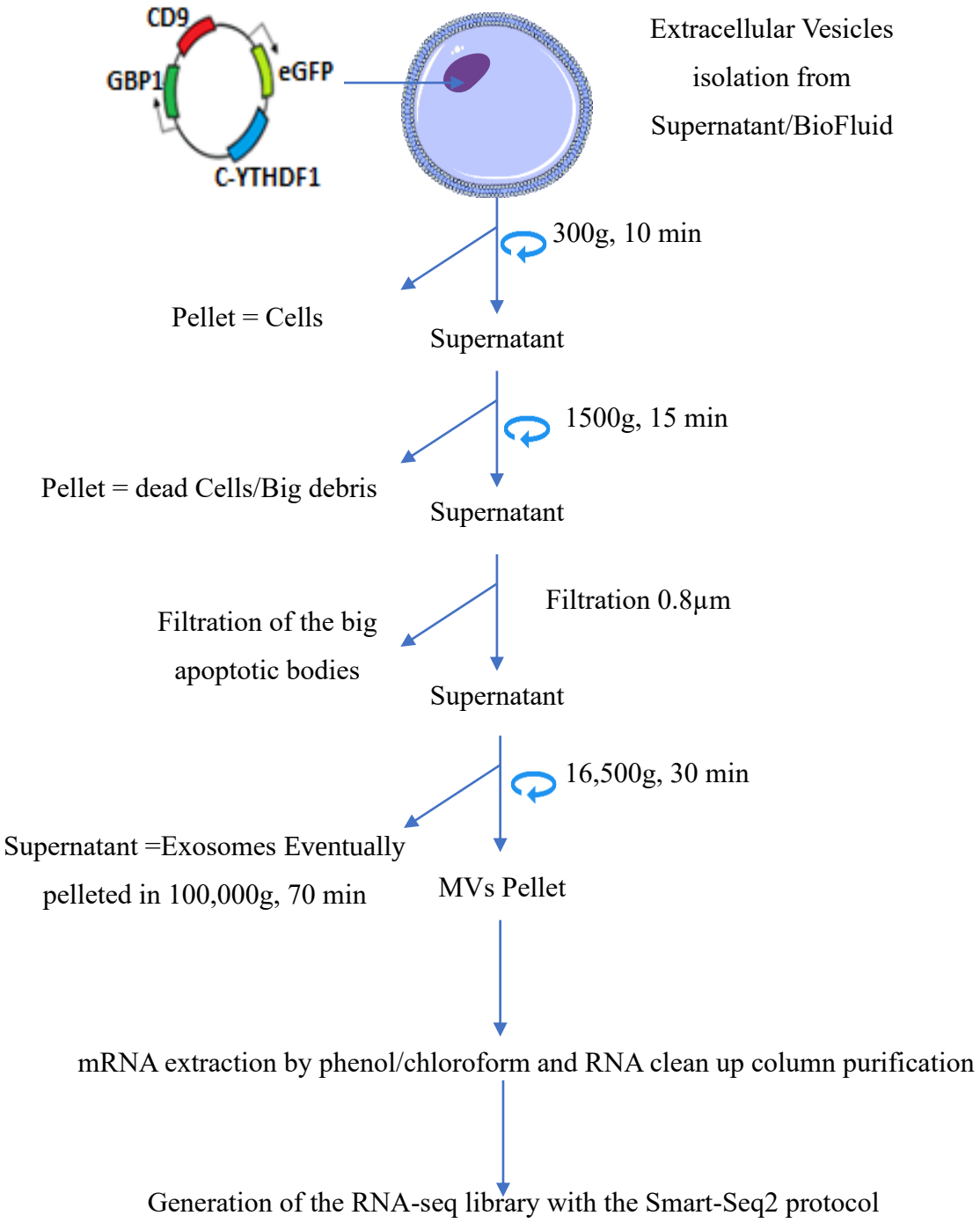
Annexed Figure 8. cDNA chip analysis from transfected HEK cells (8M cells) with the Construct C1, The negative control construct, GFP-O and untransfected cells.



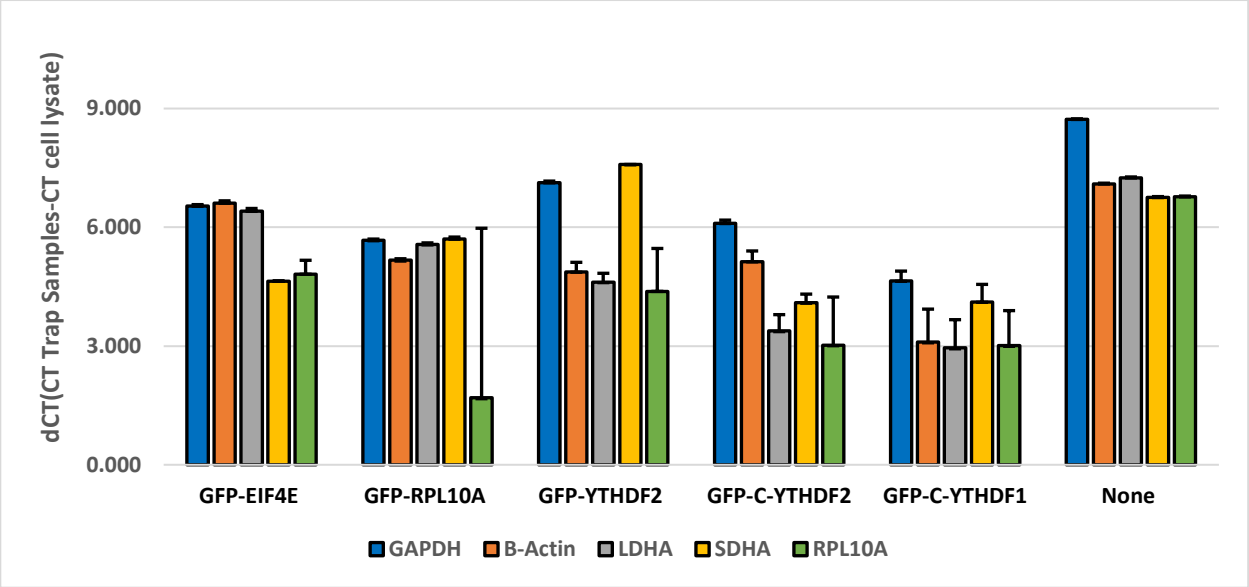
Annexed Figure 9. Gel prediction made from the bioanalyzer trace Figure 60



Annexed Figure.10 TRACE isolation method and RNA-Seq library generation steps

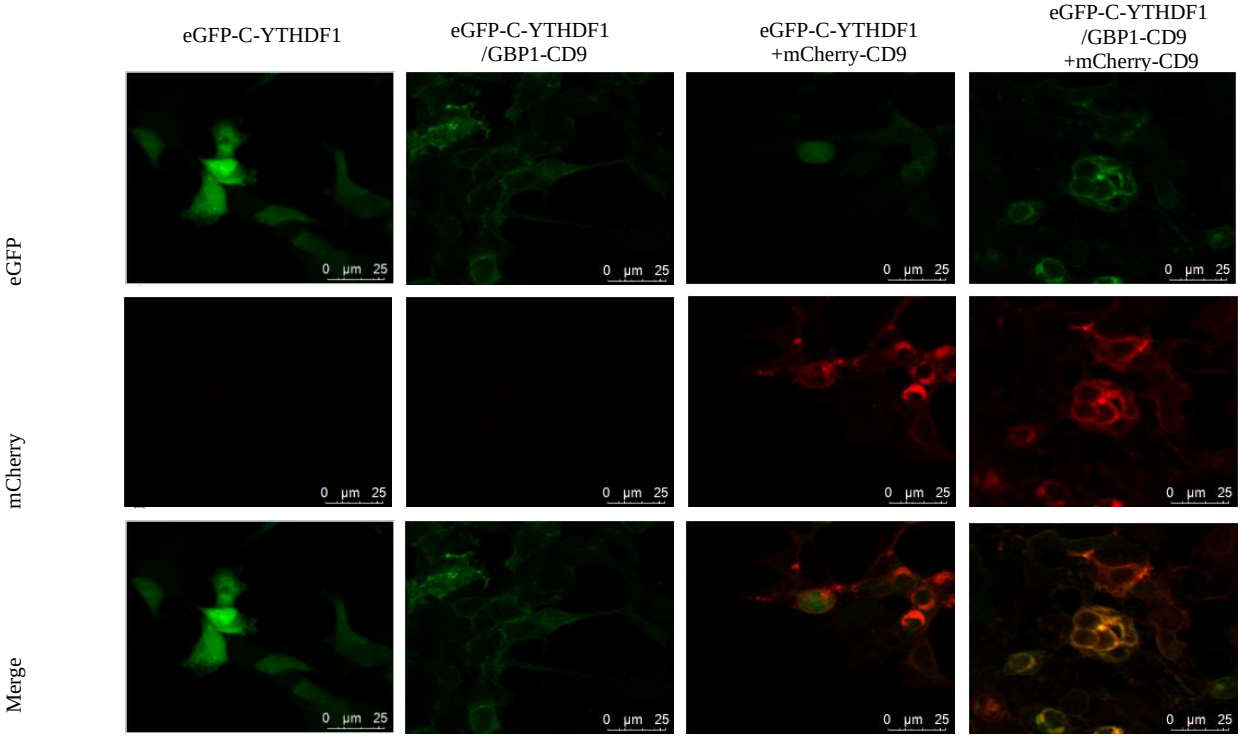


Annexed Figure.11 TRACE RBP validation RT-qPCR



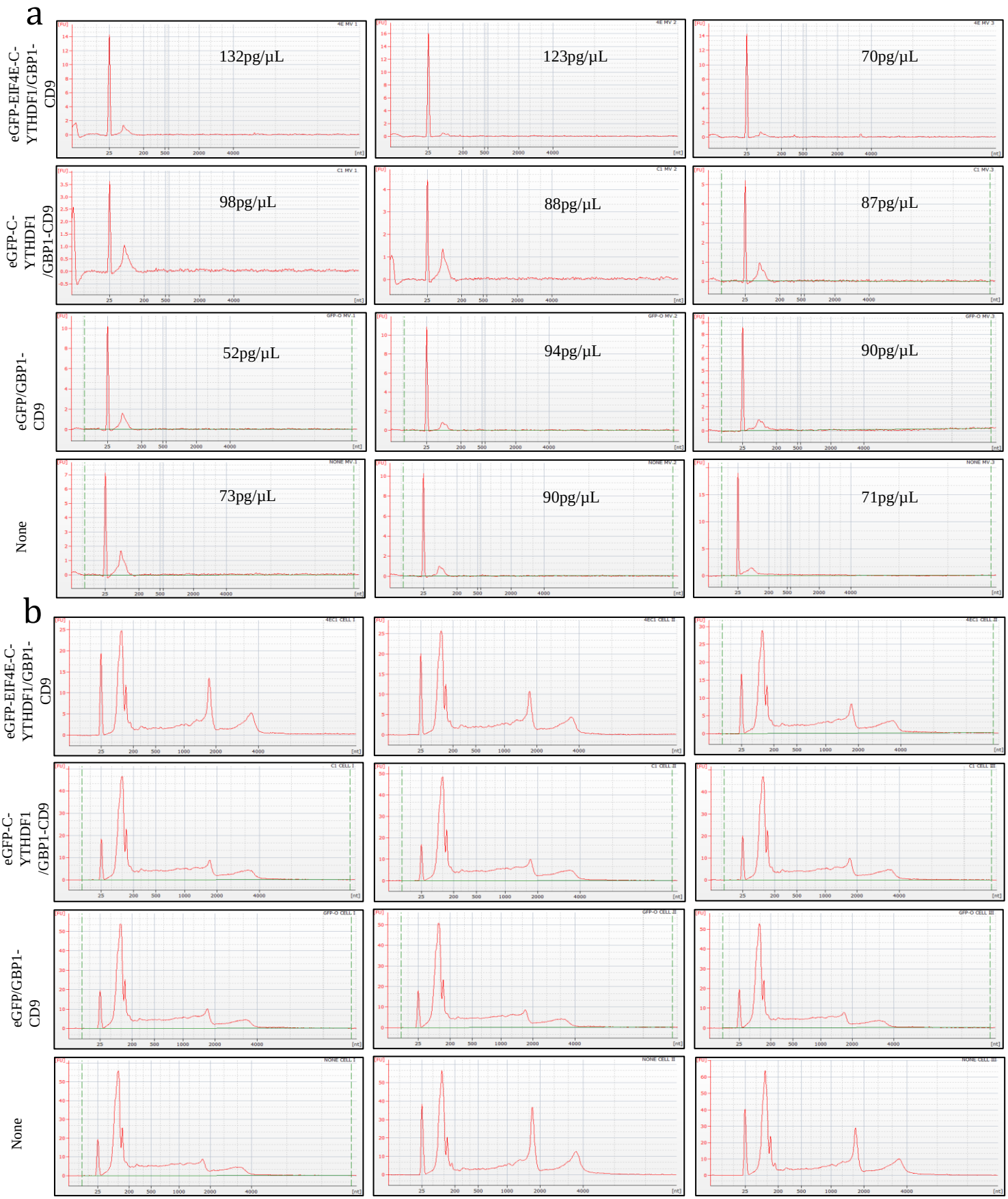
RNA binding protein (RBP) Immunoprecipitation results of different HEK 293T populations transfected with five eGFP-RBP constructs. For each of them, cells were lysed, mRNA was purified and immunoprecipitated using an anti-GFP antibody. These trapped genomic populations were normalized against the remaining expression level of their own whole lysate.

Annexed Figure.12 Confocal image from TRACE transfection



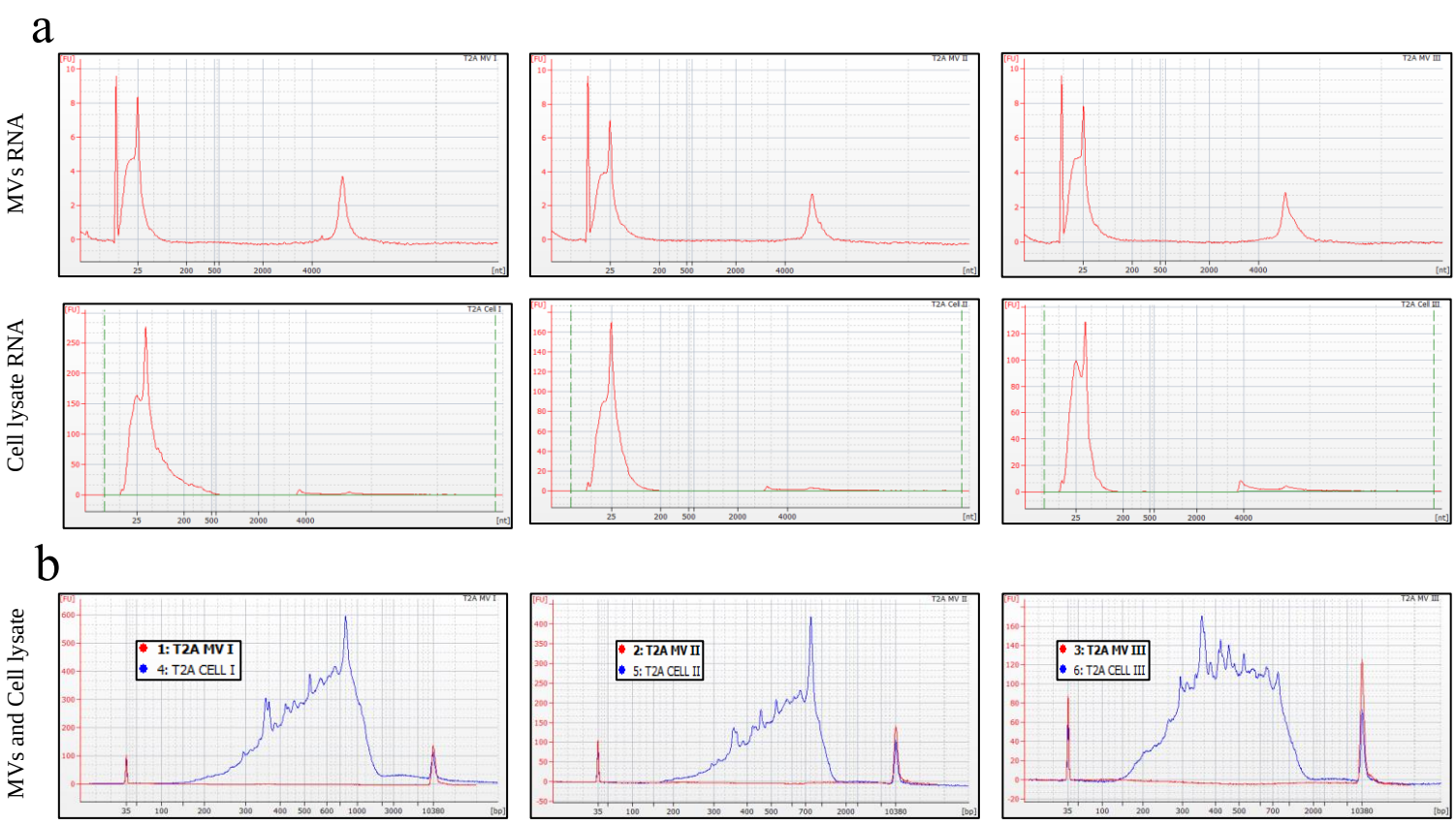
Confocal image from Transfection of different HEK 293T populations with two constructs: eGFP-C-YTHDF1, eGFP-C-YTHDF1/GBP1-CD9 and mCherry-CD9. Image taken 36h post transfection.

Annexed Figure.13 Bioanalyzer analysis of RNA content from Microvesicles and their corresponding RNA from cell



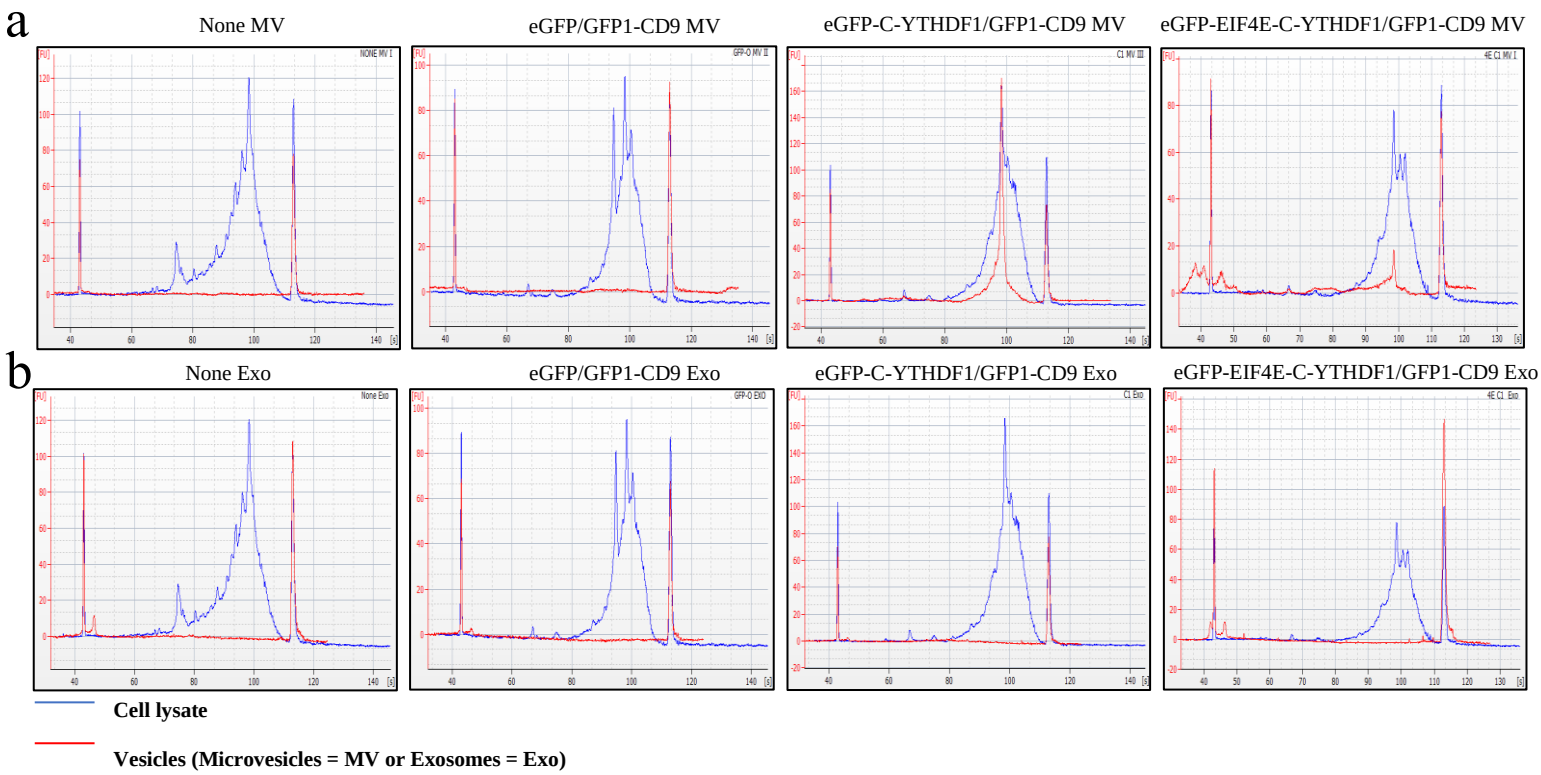
Bioanalyzer analysis of RNA content from Microvesicles and their corresponding RNA from cell lysates isolated from different HEK 293T populations (Triplicate of 4M cells per condition): untransfected cells (None), negative control construct (eGFP/GBP1-CD9) and the two types of TRACE construct (eGFP-C-YTHDF1/GBP1-CD9 and eGFP-EIF4E-C-YTHDF1/GBP1-CD9). a Microvesicles RNA content, b Cell lysate RNA content

Annexed Figure.14 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell T2A construct



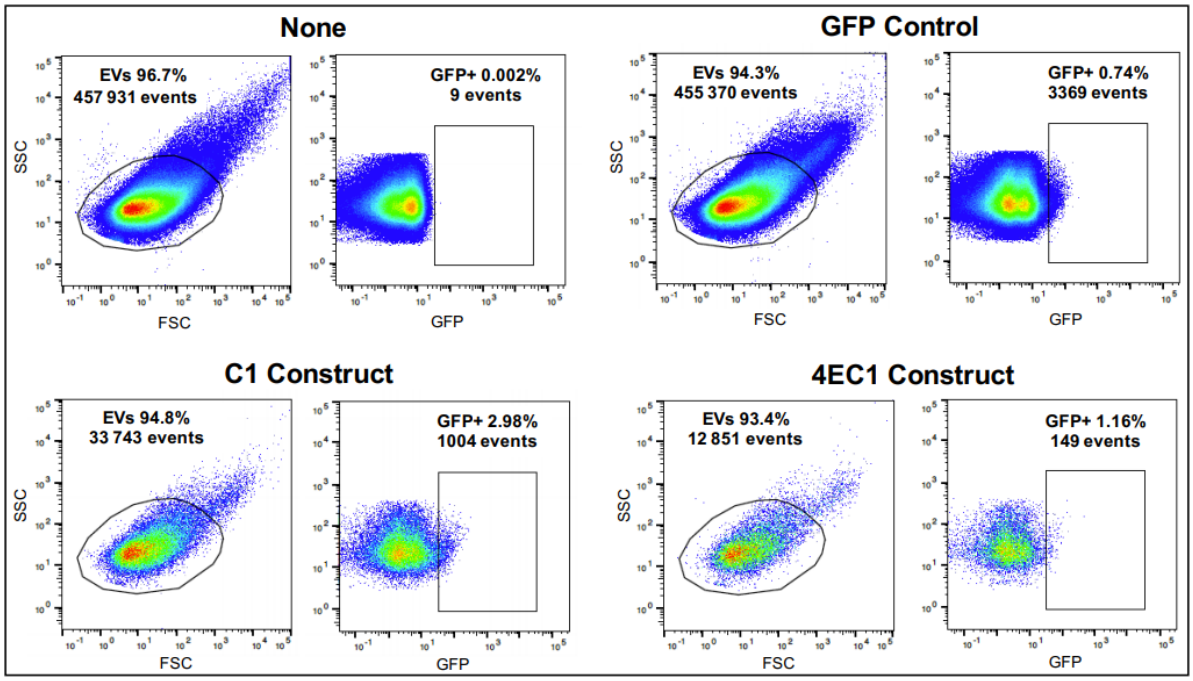
Bioanalyzer analysis of RNA content from Microvesicles and their corresponding RNA from cell lysates isolated from HEK 293T populations (Triplicate of 4M cells per condition) transfected with GBP1-CD9-T2A-eGFP-C-YTHDF1. **b** Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA generated from cell lysates, isolated from HEK 293T populations (Triplicate of 4M cells per condition) transfected with GBP1-CD9-T2A-eGFP-C-YTHDF1.

Annexed Figure.15 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell MVs vs Exosomes



a. Bioanalyzer analysis of cDNA generated from Microvesicles produced by different HEK 293T populations (4M each): untransfected cells (None), negative control construct (eGFP/GBP1-CD9) and the two types of TRACE construct (eGFP-C-YTHDF1/GBP1-CD9 and eGFP-EIF4E-C-YTHDF1/GBP1-CD9).
b. Bioanalyzer analysis of cDNA generated from Exosomes produced by the same HEK 293T populations as above.

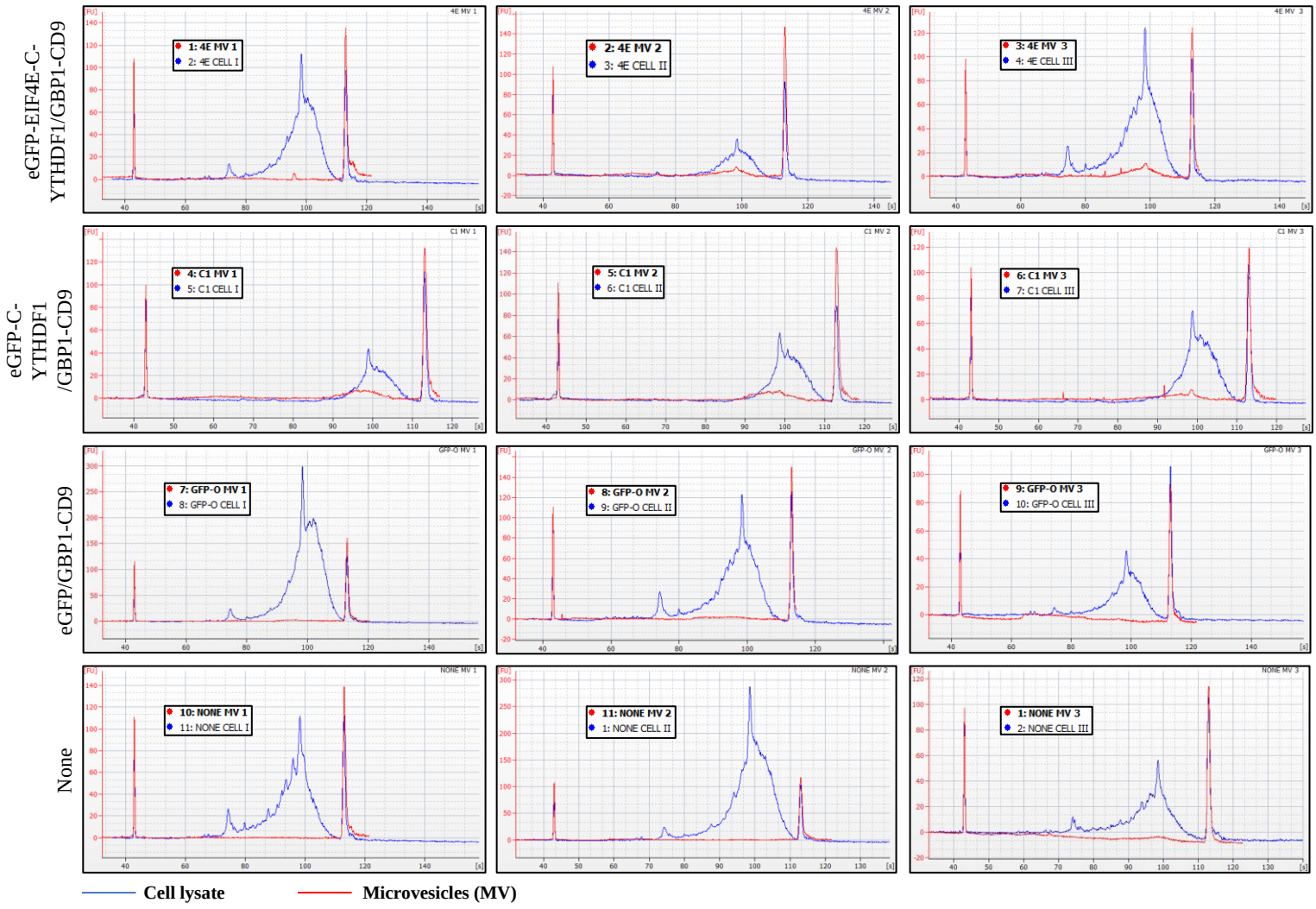
Annexed Figure.16 NanoFACS analysis of the EVs from Cells transfected with TRACE constructs and control



NanoFACS analysis on a fraction of samples (1/10) used to generate the main Fig 61 and samples used for RNAseq (triplicate of each population pooled together). Each graph focused on the EVs from HEK 293T populations (4M cells per condition): untransfected cells: None, negative control construct: eGFP/GBP1-CD9 (GFP control) and the two types of TRACE constructs: eGFP-C-YTHDF1/GBP1-CD9 (C1 construct) and eGFP-EIF4E-C-YTHDF1/GBP1-CD9 (4E construct).

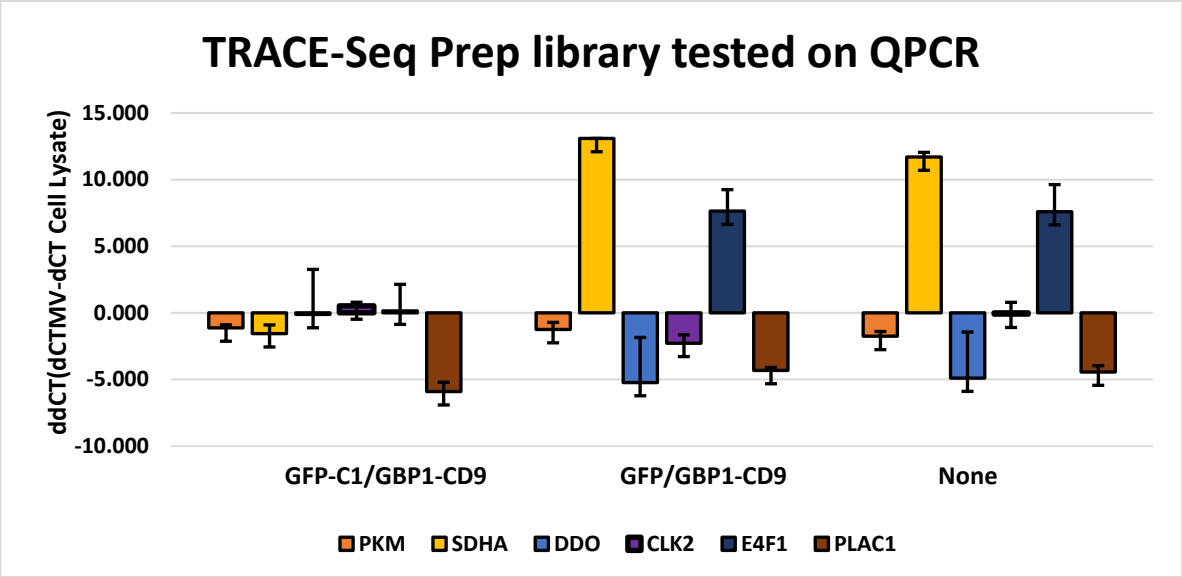
NanoFACS analysis on a fraction of samples (1/10) used to generate the main Fig 2 and samples used for RNAseq (triplicate of each population pooled together). Each graph focused on the EVs from HEK 293T populations (4M cells per condition): untransfected cells: None, negative control construct: eGFP/GBP1-CD9 (GFP control) and the two types of TRACE constructs: eGFP-C-YTHDF1/GBP1-CD9 (C1 construct) and eGFP-EIF4E-C-YTHDF1/GBP1-CD9 (4E construct).

Annexed Figure.17 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell, low amount of cells



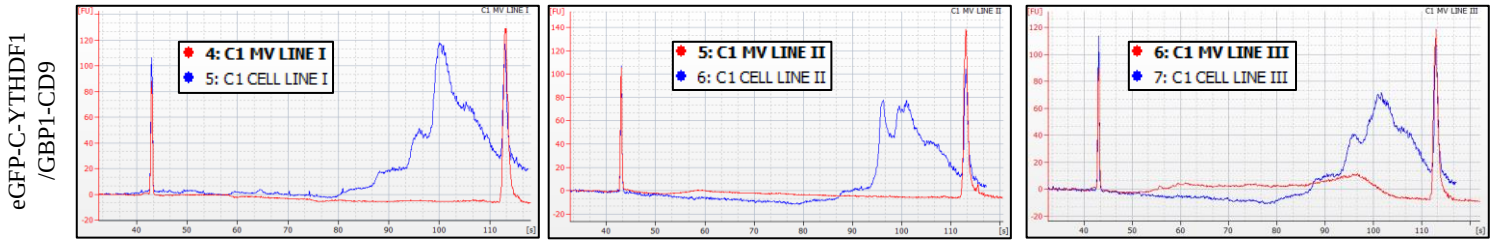
Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA from cell lysates isolated from different HEK 293T populations (Triplicate of 1M cells per condition): untransfected cells (None), negative control construct (eGFP/GBP1-CD9) and the two types of TRACE construct (eGFP-C-YTHDF1/GBP1-CD9 and eGFP-EIF4E-C-YTHDF1/GBP1-CD9).

Annexed Figure.18 TRACE-Seq library preparation validation test on RT-qPCR



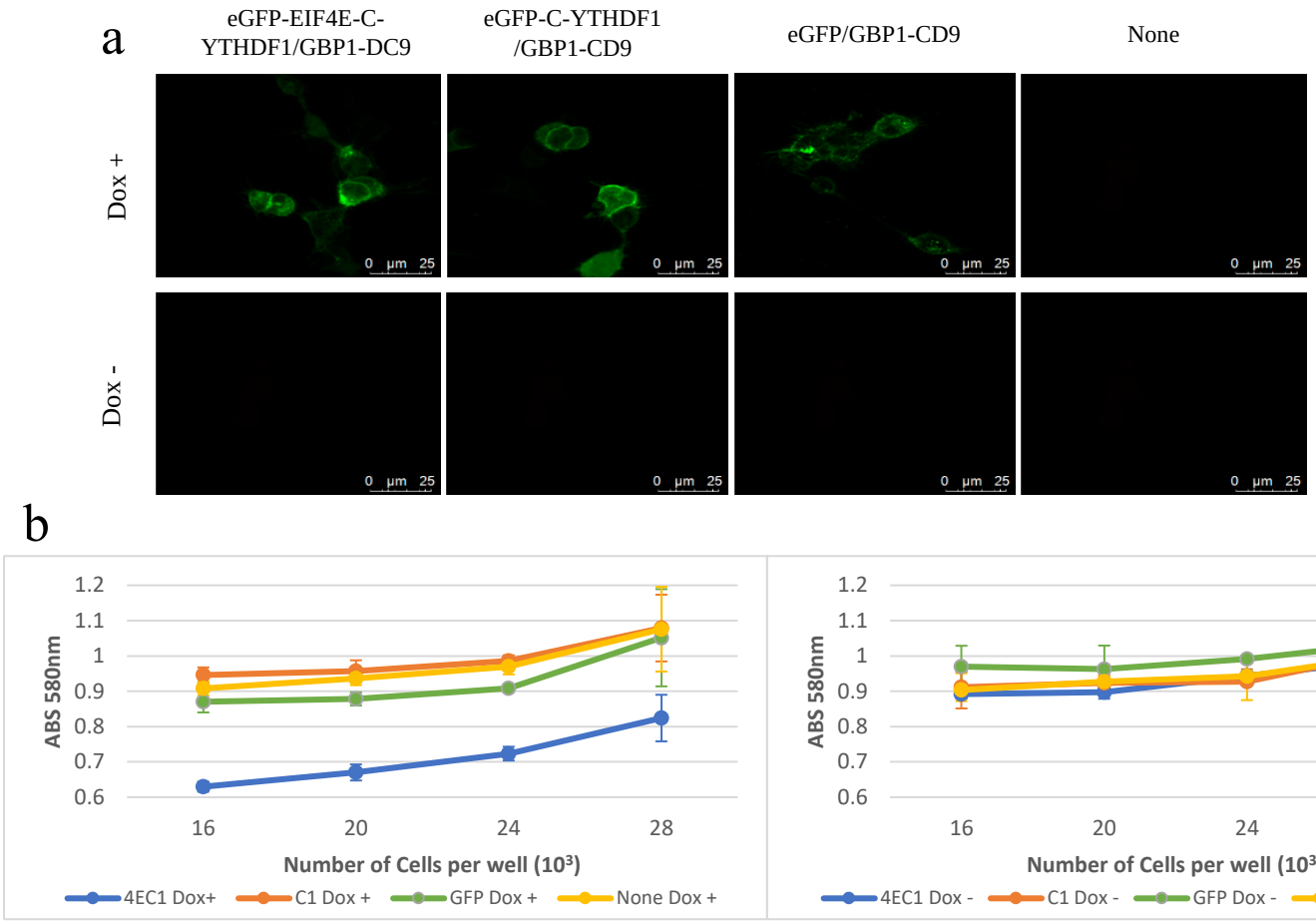
RT-qPCR results generated from MVs mRNA isolated from untransfected cells (None), negative control construct (eGFP/GBP1-CD9) and the TRACE construct (eGFP-C-YTHDF1/GBP1-CD9). These cDNA samples were normalized against the remaining expression level of their own whole lysate. All samples were made in triplicate, 8M cells per condition.

Annexed Figure.19 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell transduce with TRACE C1 constructs



Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA generate from cell lysate from Cell line generated with the construct C1: eGFP-C-YTHDF1/GBP1-CD9 (Triplicate of 8M cells per condition).

Annexed Figure.20 inductions Test of the TRACE/Control cell lines Dox vs No Dox.

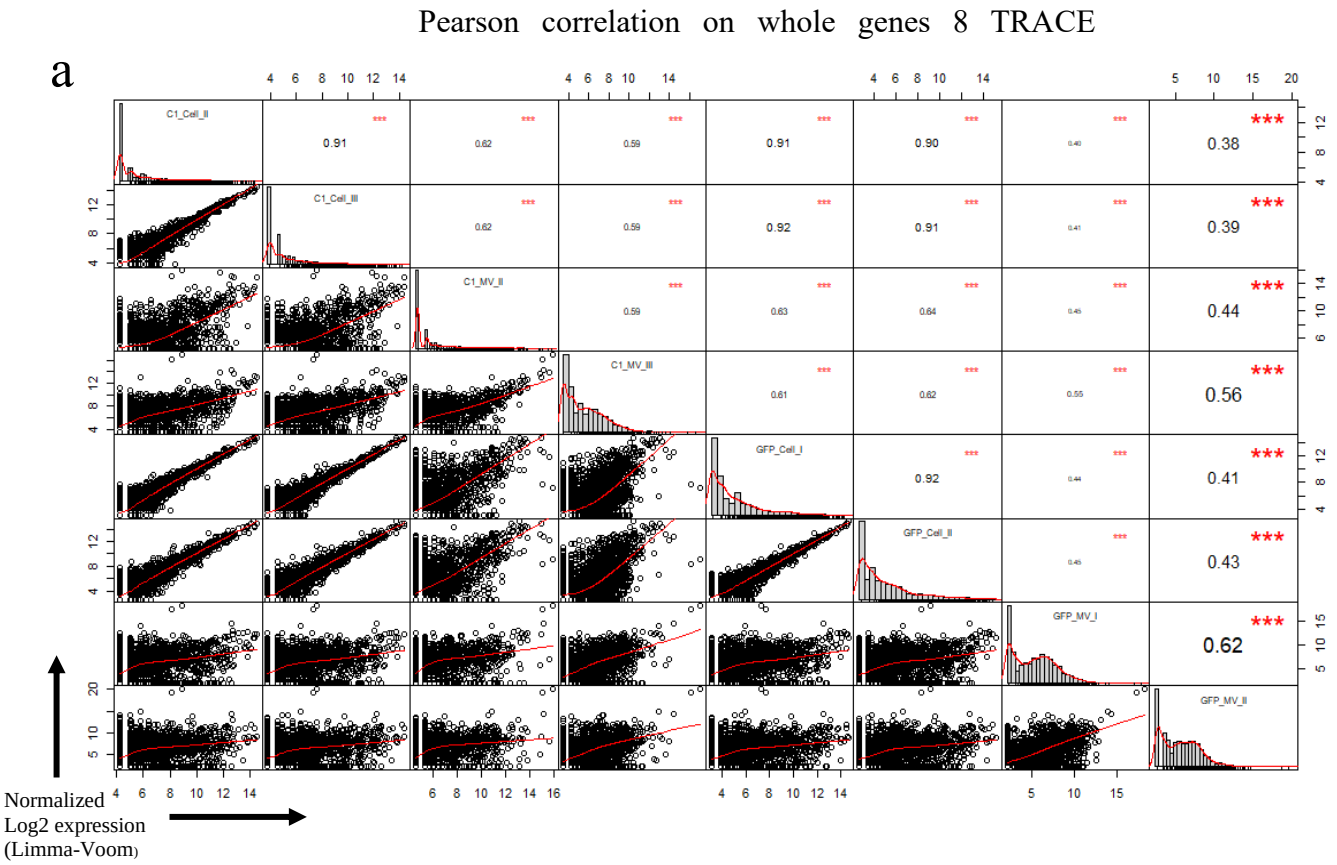


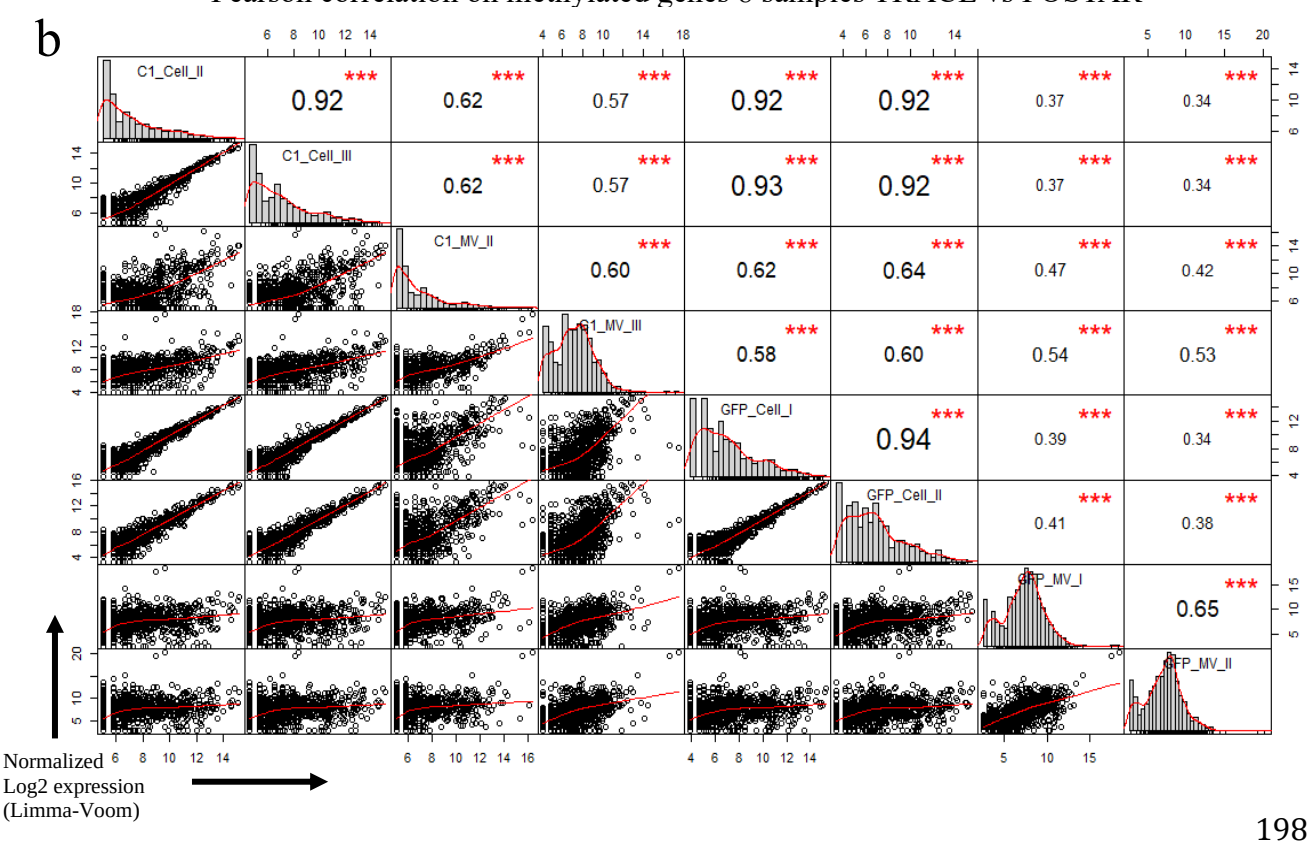
a Confocal image from the stable cell line of different HEK 293T populations treated with and without doxycycline. Cells are untreated to maintain the cell line and treated for 48h with 2µl/ml doxycycline. Cell line generated with the two constructs: eGFP-EIF4E-C-YTHDF1/GBP1-CD9 and eGFP-C-YTHDF1/GBP1-CD9, the negative control: eGFP/GBP1-CD9 and the regular HEK293T cells: None.

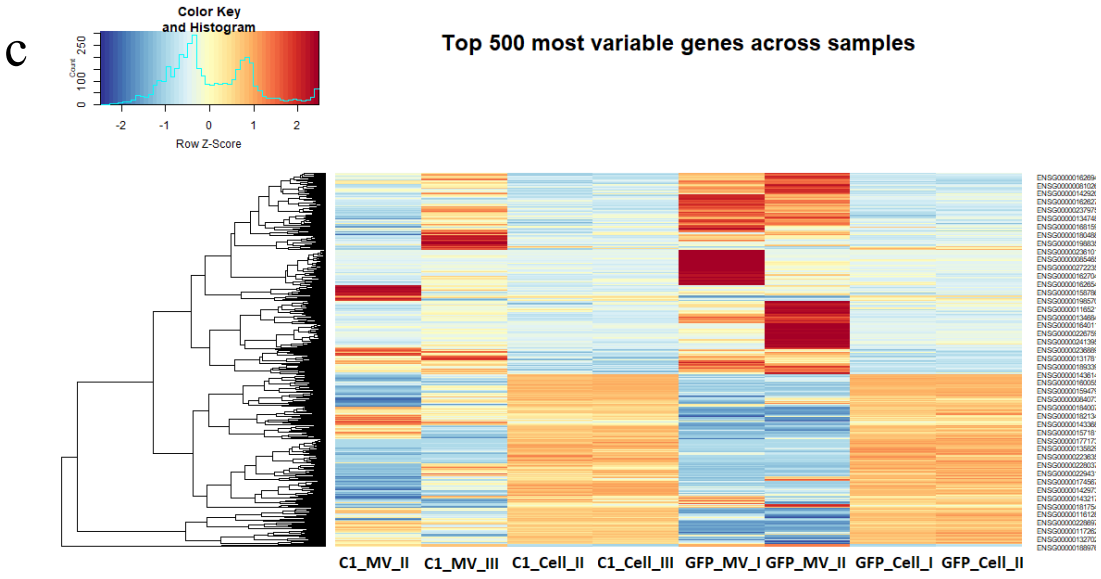
b Cell viability assay made with Reliablue on the same stable cell line as above, divided in two groups, treated or untreated with doxycycline. Cells were plated at 4 different concentrations and incubated at 37C 5% CO2 for 48 hours. After incubation, cells are incubated with Reliablue for 4 hours at 37C, 5% CO2 and absorbance were checked at 580nm.

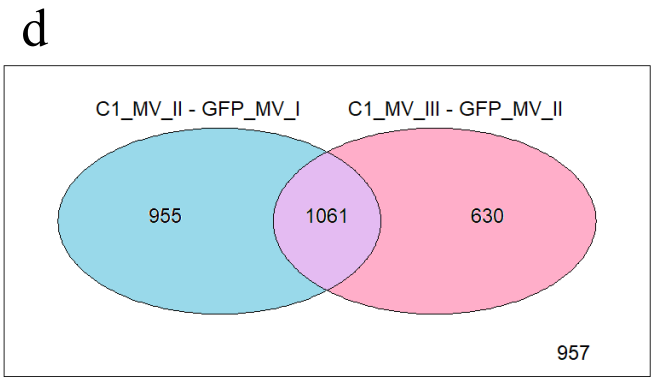
c QPCR results generated from total RNA isolated from cell lysates from all cell lines as above divided in the two groups of treated and untreated with doxycycline. Half of the plated Cell line was treated for 48h with 2µl/ml doxycycline the other not (each group made in triplicate).

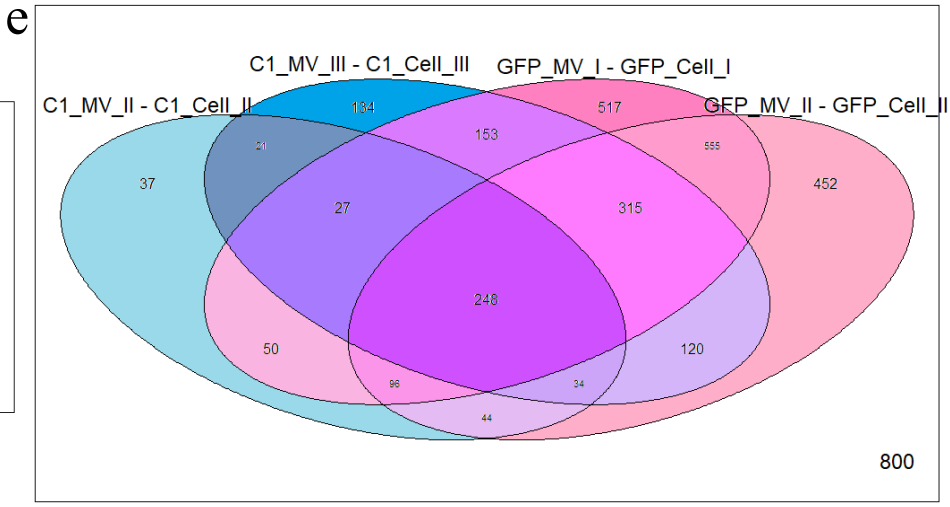
Annexed Figure.21 DEseq analysis from the 8 selected samples TRACE C1 Construct and GFP Control, Mv and Cell lysate.





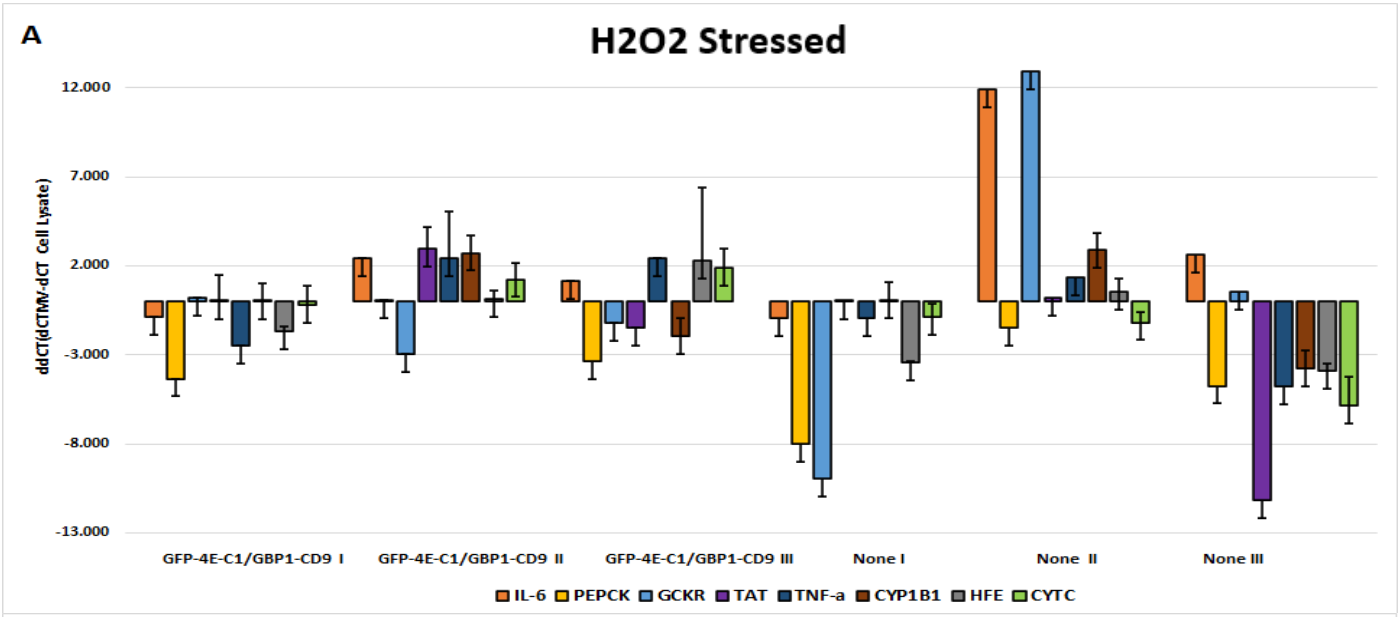


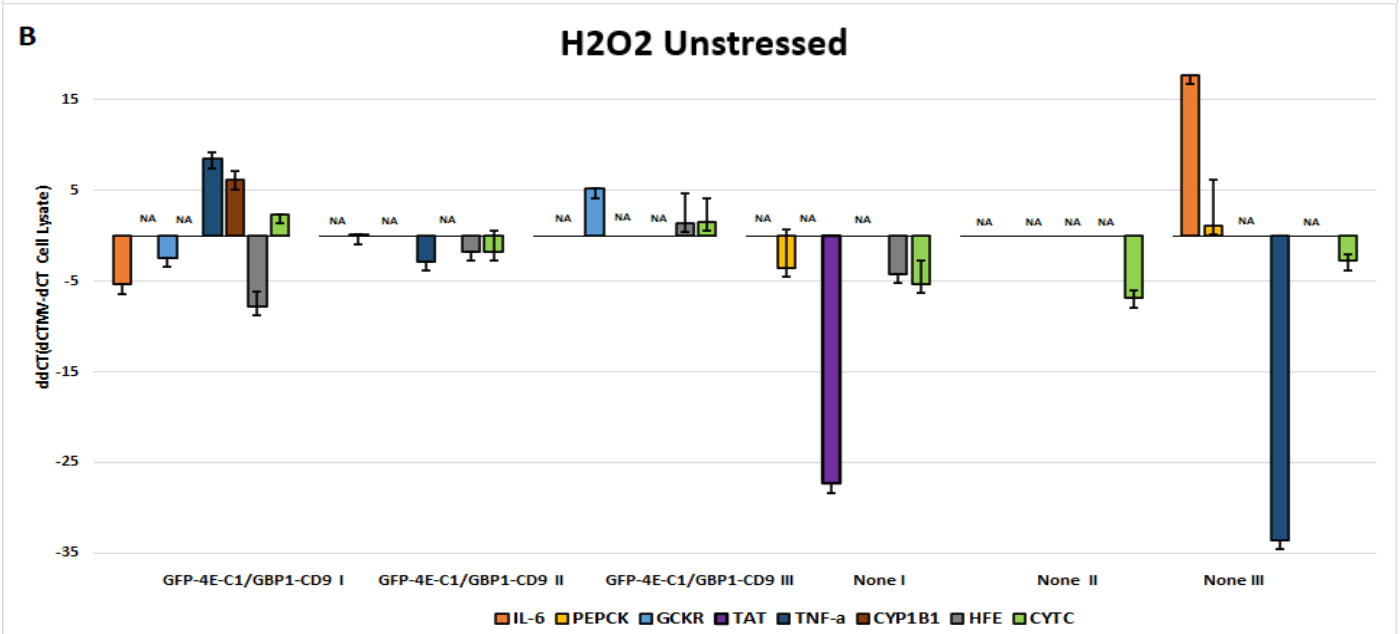


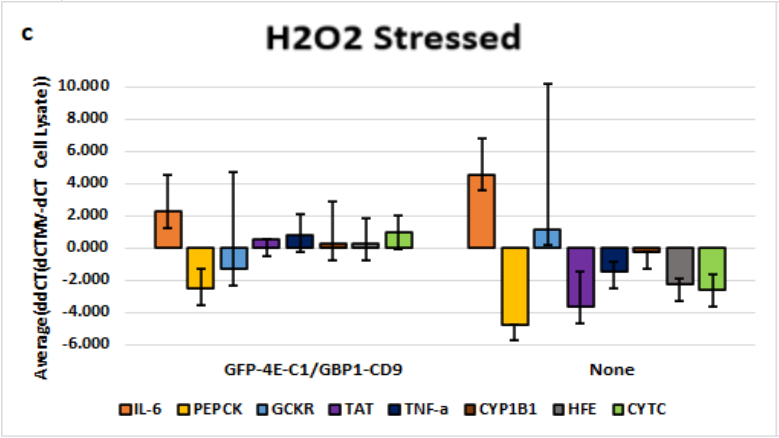


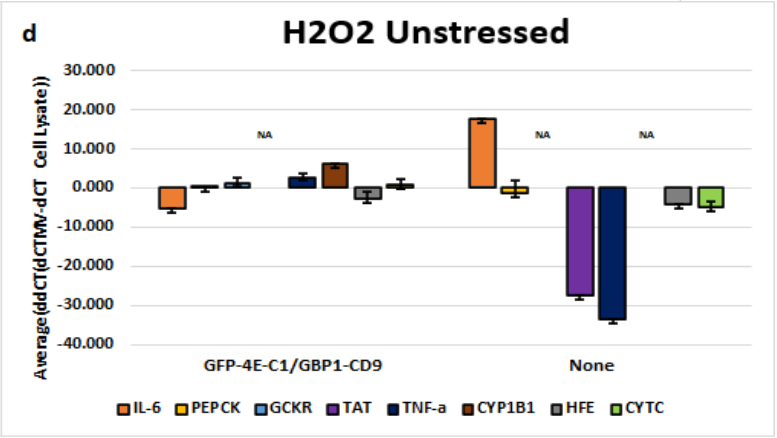
DEseq2 from the 8 selected samples TRACE GFP-C-YTHDF1/GBP1-CD9 construct: duplicate MV, duplicate cell lysates and Control GFP/GBP1-CD9: duplicate MV, duplicate cell lysates from transient transfection batch. **a.** Pearson correlation chart form the Limma-voom. DEseq analysis between each 8 samples. **b.** Pearson correlation chart from the Limma-voom. DEseq2 analysis between each 8 samples DEseq made on TRACE experiment vs POSTAR data base (methylome isolated with IP C-YTHDF1). **c.** Heat map for the most 500 variable genes across all 8 samples. **d.** Venndiagram from contrast matrix MV C1 construct- MV GFP control from genes used in Limma-voom DEseq2. **e.** Venndiagram from contrast matrix MV – cell lysate from genes used in Limma-voom DEseq with all 8 samples together. Results from duplicate of 4M cells per condition from transient transfection batch.

Annexed Figure.22 H2O2 stress pathway validation test. RT-qPCR results generated from MVs mRNA and their corresponding cell lysate









Same results of the Fig.66, H2O2 stress pathway validation test but presented in a barplot. QPCR results generated from MVs mRNA isolated from control cells (None), TRACE construct (eGFP-EIF4E-C-YTHDF1/GBP1-CD9) cell line. These cDNA samples (a,b,c), were normalized against the remaining expression level of their own cell lysate. **a.** Batch of H2O2 stressed cells. **b.** Batch of unstressed cells. **c.** Average made with stressed cells triplicate (a). **d.** Average made with unstressed cells triplicate (b). Triplicate of 15M cells per condition.

Annexed Table 1 Oligo used for cloning for Tn5 project

NAME	SEQUENCE 5' TO 3'
TN5-MSA FOR	CTCTTCACTAGTagcggcggcggcggcagcggcggcggcggcagcggcggcggcggcatggacgtttcttacctgctcga
TN5-MSA REV	CTCTTCACTAGTcaccgactccagccgggcgtat
MSA-TN5 FOR	CTCTTCTCTAGAagcggcggcggcggcagcggcggcggcggcagcggcggcggcggcatggacgtttcttacctgctcga
MSA-TN5 REV	CTCTTCCATATGcaccgactccagccgggcgtat

Annexed Table 2 Oligo used for tragmentation for the Tn5 project

NAME	SEQUENCE 5' TO 3'
TN5ME REV	[phos]CTGTCTCTTATACACATCT
TN5ME-A (ILLUMINA FC-121-1030)	NH2-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
TN5ME-B (ILLUMINA FC-121-1031)	NH2-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
PCR TN5ME-A	TGCTTAAGCAGCCGTCGCAGTCTACA
PCR TN5ME-B	TGCTTACATCTCCGAGCCCACGAGAC

“Chic-loop”

NAME	SEQUENCE 5' TO 3'
MEB TETRA POLYT V2	TTTTTTTTTTTTTTTTTTTGTATACGTTGGATTGCTGGTGGTTTGGGAGTTTTTTTTTTTTTTTTTTTTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
MEB TETRA POLYT V2 NH2	NH2-TTTTTTTTGTATACGTTGGATTGCTGGTGGTTTGGGAGTTTTTTTTTTTTTTTTTTTTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
MEA TETRA POLYT V2	CTCCCAAACCACCAGCAATCCAACGTATACTTTTTTTTTTTTTTTTTTTCTATTAAGTCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
MEB TETRA POLYT V2 BIOTIN	Biot-TTTTTTTTGTATACGTTGGATTGCTGGTGGTTTGGGAGTTTTTTTTTTTTTTTTTTTTCGCCTTAGTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Annexed Table 3 Oligo used for cloning for TRACE

NAME	SEQUENCE 5' TO 3'
GBP1 FOR	AGGAAGCTAGCGGCCCGGGATCCACCG
GBP1 REV	TCATGAAGCTTGACCTCCACCTCCTCCACCTCCTCCACCACCAGAACTAACAGTCACTTGTGTGCCCT
C-YTHDF1 FOR	CTGTTGCTCGAGTTTCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGAGAATCCCACCCCGTCCTT
C-YTHDF1 REV	CTGTGTCAATTGAACGTTTCATTGTTTGTTTCGACTCTGCC
C-YTHDF2 FOR	TACTTGTCCGGATCAGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGAACCCCACCCAGTGTT
C-YTHDF2 REV	CTGTACGGATCCTTATTTCCCACGACCTTGAC
YTHDF2 FOR	TACTTGTCCGGAACGGGTGGTGGTGGTGGTGGTGGTGGTGGAATGTCGGCCAGCAGCC
YTHDF2 REV	CTGTACGGATCCTTATTTCCCACGACCTTGACGTT
RPL10A FOR	TACAAGTCCGGAAGAACTCAGATCTCGTCCTCGTCCTTCGAATTCAGCAGCAAAGTCTCTCGCG
RPL10A REV	CTGTACGGATCCTTAATATAGGCGCTGGGGCTT
EIF4E FOR	TCTAGAAGATCTGGAGGAGGTGGTGGAGGTGGAGGTGGAGGTACTATGGCGACTGTCGAACC
EIF4E REV	CCCTCTCTCGAGTAACAACAAACCTATTTTTAGTGGTGG
CD9 T2A FOR	CATCAAAGAGGTCTTCGACAA
CD9 T2A REV	TGCATCGAATTCAGCGCTAATGGACCAGGGTTTTCTTCAACATCACCACAAGTGAGGAGAGAACCTCTACCTTCGACCATCTCGCGGTTCCT
TRE FOR	TAGTAATGCATGAGTTTACTCCCTATCAGTGATAGAGA
TRE REV	TAGTAACCGGTAGCCAATTCTCCAGGCGA
AFEI/MLUI FRAG FOR	TTGATTGCTAGCTGAGCGCTTACCTTCCTTTATGAATTGTACAGTTGGAGACGTTATTCGTTCTTTTTGTCAAGACCTACTATGCTGCATATAGTCGATCACGCGTTACCGGTTTATCT
AFEI/MLUI FRAG REV	AGATAAACCGGTAACGCGTGATCGACTATATGCAGCATAGTAGGTCTTGACAAAAAGAACGAATAACGTCTCCAACTGTACAATTCATAAAGGAAGGTAAGCGCTCAGCTAGCAAACGA

Annexed Table 4 Primers used for qPCR for TRACE

NAME	SEQUENCE 5' TO 3'	NAME	SEQUENCE 5' TO 3'
GAPDH FOR	GGAGCGAGATCCCTCCAAAAT	**DDO For**	CACAGCACGGATTGCAGTTG
GAPDH REV	GGCTGTTGTCATACTTCTCATGG	**DDO Rev**	GCCATAGTGGTGGACTACAGG
PKM FOR	ATGTCGAAGCCCCATAGTGAA	**CLK2 For**	CGAGTTGCCCTGAAGATCA
PKM REV	TGGGTGGTGAATCAATGTCCA	**CLK2 Rev**	GACTGGAGTCCCACAACTTG
B-ACTIN FOR	AGAGCTACGAGCTGCCTGAC	**E4F1 For**	CCATGTCCTCAGTGCAGTGA
B-ACTIN REV	AGCACTGTGTTGGCGTACAG	**E4F1 Rev**	CAGGATCTCGATGTCCTCTGA
LDHA FOR	ATGGCAACTCTAAAGGATCAGC	**PLAC1 For**	CCTCCTCACCTCTGCGTTT
LDHA REV	CCAACCCCAACAACTGTAATCT	**PLAC1 Rev**	CTGTGTGAAGAGACCAATCCTC
SDHA FOR	ACTGTTGCAGCACAGCTAGAA	**GFP For**	GAACGGCATCAAGGTGAACTT
SDHA REV	GCTCTGTCCACCAAATGCAC	**GFP Rev**	TCCAGCAGGACCATGTGATC
RPL10A FOR	AGCAGCAAAGTCTCTCGCG	**GBP1 For**	CATGGCCGACGTGCAGCTC
RPL10A REV	TTAATATAGGCGCTGGGGCTT	**GBP1 Rev**	AGAACTAACAGTCACTTGTGTGCCCT
TNF-A FOR	ATGAGCACTGAAAGCATGATCC	**IL-6 For**	AGACAGCCACTCACCTCTTCAG
TNF-A REV	GAGGGCTGATTAGAGAGAGGTC	**IL-6 Rev**	TTCTGCCAGTGCCTCTTTGCTG
PEPCK FOR	AAGAGACACAGTGCCCATCC	**GCKR For**	GTTGGACCTTCGGATTAGCA
PEPCK REV	ACGTAGGGTGAATCCGTCAG	**GCKR Rev**	CCCAGAAACATGGGTTCACT
TAT FOR	TGGAGTTCACAGAGCGGTTG	**CYP1B1 For**	CACCGTTTTCCGCGAATTC
TAT REV	GGTACTCGAAGCACGTTGCTG	**CYP1B1 Rev**	CCTTCTTTTCCGCAGAGAGGAT
HFE FOR	ACTGATGAAGCTGCAGAACC	**Cyt C For**	AAGGGAGGCAAGCACAAGACTG
HFE REV	GTCACCCAATTCTTTGATGG	**Cyt C Rev**	CTCCATCAGTGTATCCTCTCCC

III. Annexes manuscripts

1. The MAGIC Manuscript

The following part correspond to the abstract and the introduction from the Atmanli *et al.* elife 2019 [58] paper which I participated.

A. Abstract

A fundamental goal in the biological sciences is to determine how individual cells with varied gene expression profiles and diverse functional characteristics contribute to development, physiology, and disease. Here, we report a novel strategy to assess gene expression and cell physiology in single living cells. Our approach utilizes fluorescently labeled mRNA-specific anti-sense RNA probes and dsRNA-binding protein to identify the expression of specific genes in real-time at single-cell resolution via FRET. We use this technology to identify distinct myocardial subpopulations expressing the structural proteins myosin heavy chain α and myosin light chain 2a in real-time during early differentiation of human pluripotent stem cells. We combine this live-cell gene expression analysis with detailed physiologic phenotyping to capture the functional evolution of these early myocardial subpopulations during lineage specification and diversification. This live-cell mRNA imaging approach will have wide ranging application wherever heterogeneity plays an important biological role.

B. Introduction

A hallmark of development and disease is the cellular phenotypic diversification required for three-dimensional tissue structures. Cellular heterogeneity demonstrably contributes to the developmental dynamics of various types of stem cells (Dulken et al., 2017; Kumar et al., 2014; Wilson et al., 2015), neurons (Sandoe and Eggan, 2013) and cancer (Meacham and Morrison, 2013). In the heart, the coordinated differentiation, lineage diversification, and functional maturation of heterogeneous populations of cells is a prerequisite for the proper development of coordinated electrical and contractile function. Multiple cardiac myocyte lineages and sublineages, along with endothelial cells, smooth muscle cells and cardiac fibroblasts must interact in a cohesive program to form the mature four-chambered adult heart (Bu et al., 2009; Domian et al., 2009). Advances in pluripotent stem cell (PSC) biology open unprecedented avenues for the study of human cellular differentiation, physiology, and pathophysiology in vitro (Lan et al., 2013) and also underscore the heterogeneity of clinically important cell types (Bryant et al., 1997; Burridge et al., 2014; Cordeiro et al., 2004; Lian et al., 2012). This cellular heterogeneity along with an inherent difficulty of examining real-time gene expression of single living cells poses a major limitation in the understanding of the complex biological processes that underlie development and disease.

Single-cell transcriptional profiling initially via multiplex qPCR analysis and more recently via whole transcriptome sequencing has provided insight into how intracellular signaling is regulated at the single-cell transcriptional level during cardiac development (Cui et al., 2019; DeLaughter et al., 2016; Friedman et al., 2018; Li et al., 2016; Sahara et al., 2019). Despite this progress, whole genome expression analysis does not allow for concurrent physiological assessment of single living cells and consequently, the functional significance of single-cell transcriptomic heterogeneity remains unclear. The live-cell identification of distinct cell populations has most commonly been accomplished with gene expression assays that rely on the detection of fluorescent reporter proteins under the transcriptional control of the gene of interest. Accordingly, these approaches require the generation of transgenic animals (Domian et al., 2009; Wu et al., 2006) or embryonic stem cell lines (Elliott et al., 2011; Klug et al., 1996) to isolate and study discrete subsets of cells with specific gene expression profiles. These methods are cumbersome, time consuming, and expensive and therefore allow for only a limited number of genes to be examined at a time. Technical advances have facilitated live-cell mRNA imaging by detecting gene transcripts via nucleic acid (Santangelo et al., 2009; Tyagi and Kramer, 1996; Vargas et al., 2011) or protein probes (Bertrand et al., 1998; Nelles et al., 2016; Ozawa et al., 2007). However, several drawbacks of these existing techniques such as genetic encoding of target

mRNA and reporter protein, the necessity to target multiple binding sites, complexity of probe design and cellular delivery and low sensitivity (Armitage, 2011; Tyagi, 2009) have prevented their widespread use (Table 1).

Table 1

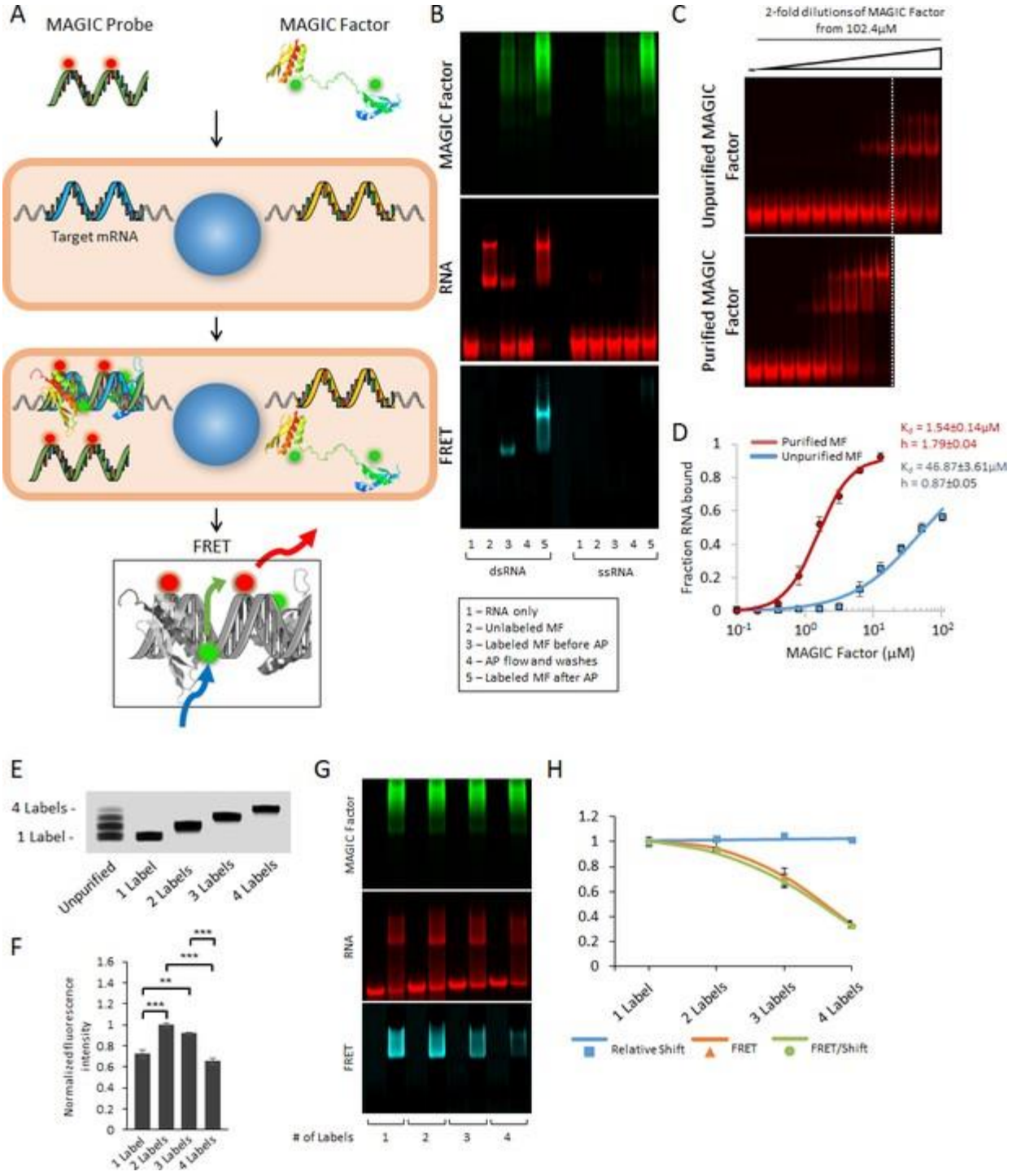
Comparison of MAGIC with other live-cell mRNA imaging technologies.

https://doi.org/10.7554/eLife.49599.002

	Advantages	**Disadvantages**
Nucleic Acid Probes	Most established approach	Complexity of probe design and cellular delivery
	Single-molecule sensitivity achievable	Need to screen many probes for specificity and sensitivity
	Cell isolation via FACS	Probe sequestration and false-positive signals
Protein Probes	Single-molecule sensitivity	Genetic encoding of target RNA and reporter protein
	Study of RNA dynamics	Multiple binding sites necessary
		Low sensitivity
MAGIC	Imaging of transcription factors	Complexity of MAGIC Probe production
	Double detection with MAGIC Factor and Probes increases specificity	Efficient transfection of MAGIC Factor and Probes into the same cell required
	Coupling with cell physiology assays	

Herein we describe a novel Förster Resonance Energy Transfer (FRET)-based technology for the *m*ultiplex *a*nalysis of gene expression in *i*ndividual living *c*ells (termed MAGIC, patent application pending; Atmanli and Domian, 2016) (Figure 1A). Our technology enables the real-time assessment of transcript expression in single living cells. We functionally characterize distinct myocardial subpopulations derived from human PSC (hPSC) in real-time and describe marked differences in cell physiology among different cardiac sublineages. Our observations reinforce the importance of live-cell lineage assignment in a heterogeneous assembly of single living cells, and we anticipate that our technology could have wide-ranging application in any biological system.

Figure 1 with 1 supplement



Strategy for the multiplex analysis of gene expression in individual living cells (MAGIC).

(**A**) MAGIC anti-sense RNA Probes are generated by in vitro transcription using T7 phage polymerase and fluorescently-labeled using aminoallyl-modified UTP nucleotides. MAGIC Factor is the recombinant dsRNA-binding domain of human protein kinase R that has been fluorescently labeled and affinity purified against dsRNA. Both the probe and the protein are delivered into living cells by transfection. Upon cellular delivery, MAGIC Probes hybridize to their target gene and generate an RNA-RNA duplex. This enables MAGIC Factor to bind to the newly formed dsRNA, thereby bringing the donor fluorophore on the protein to come into close contact with the acceptor fluorophore on the RNA probe and for FRET to occur (blue arrow: excitation of the donor; green arrow: energy transfer; red arrow: emission of the acceptor). Cells expressing the gene of interest are identified by assaying the FRET signal. (**B**) Affinity purification of fluorescent MAGIC Factor restores its binding affinity. MAGIC Factor was fluorescently-labeled with Alexa Fluor 488 and then reacted with dsRNA-coupled agarose beads to separate binding, functional protein from non-binding, non-functional protein. An electrophoretic mobility shift assay (EMSA) of dsRNA and ssRNA labeled with Alexa Fluor 647 and reacted with MAGIC Factor is shown in the RNA, protein and FRET channels. (**C**) Unpurified and purified MAGIC Factor were reacted at increasing concentrations with a fixed concentration of dsRNA (200 nM) and run on a native 12% polyacrylamide gel. The dashed lines represent the cutoff after

which increasing concentrations of only unpurified MAGIC Factor were reacted with dsRNA. (**D**) Quantification of the binding affinity (represented as dissociation constant Kd) and Hill coefficients of unpurified and affinity purified, fluorescent MAGIC Factor (MF) from the EMSA in (**C**). Affinity purification of MAGIC Factor resulted in a 30-fold increase in binding affinity to dsRNA. (**E**) In vitro transcribed and fluorescently-labeled 20-mer RNA probe was gel purified to obtain one-, two-, three- and four-labeled RNA probes. (**F**) The fluorescence intensities of equimolar concentrations of purified probes with one to four labels were measured using a spectrophotometer. (**G**) 20-mer RNA probes with one to four labels were reacted with unlabeled sense probes to generate dsRNA. Representative EMSA with MAGIC Factor is shown in the RNA, protein and FRET channels. Note that the appearance of MAGIC Factor as multiple bands is likely due to the use of an NHS-ester dye to attach Alexa Fluor 488 to the protein. Dependent on the exact location of the fluorophore molecule, each single protein molecule likely runs differently on the native polyacrylamide gel. (**H**) The relative shift of fluorescent dsRNA, the FRET intensity of shifted dsRNA and the FRET/shift ratio for one to four labeled RNA probes were quantified from the EMSA in (**G**). Quantified data are shown as mean ± s.e.m. **p<0.01 and ***p<0.001.

https://doi.org/10.7554/eLife.49599.003

2. The TRACE-seq Manuscript

The following part correspond to the abstract and the introduction from the Cherbonneau *et al.* 2020 manuscript currently in review.

A. Abstract

Changes in gene expression due to cell heterogeneity and response to stressors in specific microenvironments remains poorly understood. Current methodologies that seek to interrogate gene expression at a molecular level require sampling of cellular transcriptome and therefore lysis of the cell preventing serial analysis of cellular transcriptome. To address this area of unmet need, we have recently developed a new technology allowing transcriptomic analysis over time without cellular destruction. Our novel method, TRACE-seq (TRanscriptomic Analysis Captured in Extracellular vesicles using sequencing), is characterized by a cell-type specific transgene expression. It provides data on a representative part of the cell transcriptome inside extracellular vesicles. Thus, the transcriptome of cells expressing TRACE can be followed over time in a non-destructive manner, which is a powerful tool for many fields of fundamental and translational biology research.

B. Introduction

For several decades, mRNA profiling by microarray and RNA sequencing [123, 124] have allowed investigators to push the frontiers of knowledge by providing a better understanding of the architectural complexity that underlies cellular heterogeneity. Despite the development of many techniques with increasingly powerful resolution, a fundamental gap in methodology

persists in that the most current techniques are cell destructive [125] and are therefore incompatible with studies examining the dynamics of cellular change. While *in situ* hybridization in live cells using mRNA probes is feasible [126], it is technically challenging to use routinely, not amenable to multiplexing, and cannot provide a longitudinal measure of the whole cell transcriptome.

Substantial effort has been made to profile sub-populations of cells by transcriptomic analysis [75, 256], focusing on specific cell types [128] or using single cell sequencing methods [127, 128, 257]. Significant technological advancements have reduced experimental noise while obtaining representative results with low input starting material. Recently, development of alternative approaches for time-resolved, longitudinal extraction and quantitative measurement of intracellular mRNA have been reported [129], but are limited to *in vitro* studies and appear unsuitable for routine usage or for *in vivo* studies. Exploiting the presence of mRNAs in extracellular vesicles (EVs), we aim to develop technology to that efficiently and uniformly load cellular mRNAs into EVs in-vivo. The non-invasive transcriptome profiling that such a method affords would enable serial, non-destructive, non-interfering sampling of designated cells simply by collection of physiological fluid.

Here we present a technique, TRACE-seq (TRanscriptomic Analysis Captured in Extracellular vesicles using sequencing), that exploits the cell's existing gene expression mechanism to stochastically/uniformly secrete mRNAs into EVs. TRACE-seq provides a representative sampling of the cytosolic transcriptome allowing monitoring of live cells in physiological or pathological conditions by a serial and non-destructive transcriptomic analysis.

Bibliography

1. Mathe, G. and L. Schwarzenberg, *Bone-marrow transplantation in France, 1958-1973.* Transplant Proc, 1974. **6**(4): p. 335-43.
2. Yamanaka, S., *Strategies and new developments in the generation of patient-specific pluripotent stem cells.* Cell Stem Cell, 2007. **1**(1): p. 39-49.
3. Yu, J., et al., *Induced pluripotent stem cell lines derived from human somatic cells.* Science, 2007. **318**(5858): p. 1917-20.
4. Spivakov, M. and A.G. Fisher, *Epigenetic signatures of stem-cell identity.* Nat Rev Genet, 2007. **8**(4): p. 263-71.
5. Branco, M.A., et al., *Transcriptomic analysis of 3D Cardiac Differentiation of Human Induced Pluripotent Stem Cells Reveals Faster Cardiomyocyte Maturation Compared to 2D Culture.* Sci Rep, 2019. **9**(1): p. 9229.
6. Porrello, E.R., et al., *Transient regenerative potential of the neonatal mouse heart.* Science, 2011. **331**(6020): p. 1078-80.
7. Bergmann, O., et al., *Evidence for cardiomyocyte renewal in humans.* Science, 2009. **324**(5923): p. 98-102.
8. Laugwitz, K.L., et al., *Postnatal isl1+ cardioblasts enter fully differentiated cardiomyocyte lineages.* Nature, 2005. **433**(7026): p. 647-53.
9. Mozaffarian, D., et al., *Heart disease and stroke statistics--2015 update: a report from the American Heart Association.* Circulation, 2015. **131**(4): p. e29-322.
10. Laflamme, M.A. and C.E. Murry, *Heart regeneration.* Nature, 2011. **473**(7347): p. 326-35.
11. Barnard, C.N., *The operation. A human cardiac transplant: an interim report of a successful operation performed at Groote Schuur Hospital, Cape Town.* S Afr Med J, 1967. **41**(48): p. 1271-4.
12. Carpentier, A., et al., *First clinical use of a bioprosthetic total artificial heart: report of two cases.* Lancet, 2015.
13. van Berlo, J.H. and J.D. Molkentin, *An emerging consensus on cardiac regeneration.* Nat Med, 2014. **20**(12): p. 1386-93.
14. Kikuchi, K. and K.D. Poss, *Cardiac regenerative capacity and mechanisms.* Annu Rev Cell Dev Biol, 2012. **28**: p. 719-41.
15. Gerbin, K.A. and C.E. Murry, *The winding road to regenerating the human heart.* Cardiovasc Pathol, 2015. **24**(3): p. 133-40.
16. Li, T.S., et al., *Direct comparison of different stem cell types and subpopulations reveals superior paracrine potency and myocardial repair efficacy with cardiosphere-derived cells.* J Am Coll Cardiol, 2012. **59**(10): p. 942-53.
17. Bolli, R., et al., *Cardiac stem cells in patients with ischaemic cardiomyopathy (SCIPIO): initial results of a randomised phase 1 trial.* Lancet, 2011. **378**(9806): p. 1847-57.
18. Malliaras, K., et al., *Intracoronary cardiosphere-derived cells after myocardial infarction: evidence of therapeutic regeneration in the final 1-year results of the CADUCEUS trial (CArdiosphere-Derived aUtologous stem CElls to reverse ventricUlar dySfunction).* J Am Coll Cardiol, 2014. **63**(2): p. 110-22.

19. Takehara, N., et al., *[The cardiovascular regeneration therapy for ischemic heart disease: road to heart repair].* Nihon Rinsho, 2011. **69 Suppl 7**: p. 517-23.
20. Simari, R.D., et al., *Bone marrow mononuclear cell therapy for acute myocardial infarction: a perspective from the cardiovascular cell therapy research network.* Circ Res, 2014. **114**(10): p. 1564-8.
21. Chen, S.L., et al., *Effect on left ventricular function of intracoronary transplantation of autologous bone marrow mesenchymal stem cell in patients with acute myocardial infarction.* Am J Cardiol, 2004. **94**(1): p. 92-5.
22. Simari, R.D., et al., *Bone marrow mononuclear cell therapy for acute myocardial infarction: a perspective from the cardiovascular cell therapy research network.* Circ Res, 2014. **114**(10): p. 1564-8.
23. Dimmeler, S., J. Burchfield, and A.M. Zeiher, *Cell-based therapy of myocardial infarction.* Arterioscler Thromb Vasc Biol, 2008. **28**(2): p. 208-16.
24. Hughey, C.C., et al., *Mesenchymal stem cell transplantation for the infarcted heart: therapeutic potential for insulin resistance beyond the heart.* Cardiovasc Diabetol, 2013. **12**: p. 128.
25. Pandey, A.C., et al., *Cellular Therapeutics for Heart Failure: Focus on Mesenchymal Stem Cells.* Stem Cells Int, 2017. **2017**: p. 9640108.
26. Menasche, P., et al., *Human embryonic stem cell-derived cardiac progenitors for severe heart failure treatment: first clinical case report.* Eur Heart J, 2015.
27. Yoshida, Y. and S. Yamanaka, *iPS cells: a source of cardiac regeneration.* J Mol Cell Cardiol, 2011. **50**(2): p. 327-32.
28. Burridge, P.W., et al., *Production of de novo cardiomyocytes: human pluripotent stem cell differentiation and direct reprogramming.* Cell Stem Cell, 2012. **10**(1): p. 16-28.
29. Mignone, J.L., et al., *Cardiogenesis from human embryonic stem cells.* Circ J, 2010. **74**(12): p. 2517-26.
30. Shiba, Y., et al., *Human ES-cell-derived cardiomyocytes electrically couple and suppress arrhythmias in injured hearts.* Nature, 2012. **489**(7415): p. 322-5.
31. Fernandes, S., et al., *Human embryonic stem cell-derived cardiomyocytes engraft but do not alter cardiac remodeling after chronic infarction in rats.* J Mol Cell Cardiol, 2010. **49**(6): p. 941-9.
32. Laflamme, M.A., et al., *Cardiomyocytes derived from human embryonic stem cells in pro-survival factors enhance function of infarcted rat hearts.* Nat Biotechnol, 2007. **25**(9): p. 1015-24.
33. van Laake, L.W., et al., *Human embryonic stem cell-derived cardiomyocytes survive and mature in the mouse heart and transiently improve function after myocardial infarction.* Stem Cell Res, 2007. **1**(1): p. 9-24.
34. Shiba, Y., et al., *Electrical Integration of Human Embryonic Stem Cell-Derived Cardiomyocytes in a Guinea Pig Chronic Infarct Model.* J Cardiovasc Pharmacol Ther, 2014. **19**(4): p. 368-381.
35. Mohsin, S., D. Avitabile, and M. Khan, *Stem Cells and Cardiac Repair.* Stem Cells Int, 2015. **2015**: p. 153627.
36. Jopling, C., et al., *Zebrafish heart regeneration occurs by cardiomyocyte dedifferentiation and proliferation.* Nature, 2010. **464**(7288): p. 606-9.
37. Kikuchi, K., et al., *Primary contribution to zebrafish heart regeneration by gata4(+) cardiomyocytes.* Nature, 2010. **464**(7288): p. 601-5.

38. Szibor, M., et al., *Remodeling and dedifferentiation of adult cardiomyocytes during disease and regeneration.* Cell Mol Life Sci, 2014. **71**(10): p. 1907-16.
39. Kubin, T., et al., *Oncostatin M is a major mediator of cardiomyocyte dedifferentiation and remodeling.* Cell Stem Cell, 2011. **9**(5): p. 420-32.
40. Morrisey, E.E., *Rewind to recover: dedifferentiation after cardiac injury.* Cell Stem Cell, 2011. **9**(5): p. 387-8.
41. Wang, W.E., et al., *Dedifferentiation, Proliferation, and Redifferentiation of Adult Mammalian Cardiomyocytes After Ischemic Injury.* Circulation, 2017. **136**(9): p. 834-848.
42. Park, I.H., et al., *Generation of human-induced pluripotent stem cells.* Nat Protoc, 2008. **3**(7): p. 1180-6.
43. Yu, J., et al., *Induced pluripotent stem cell lines derived from human somatic cells.* Science, 2007. **318**(5858): p. 1917-20.
44. Abu-Issa, R. and M.L. Kirby, *Heart field: from mesoderm to heart tube.* Annu Rev Cell Dev Biol, 2007. **23**: p. 45-68.
45. Olson, E.N. and M.D. Schneider, *Sizing up the heart: development redux in disease.* Genes Dev, 2003. **17**(16): p. 1937-56.
46. Mummery, C.L., et al., *Differentiation of human embryonic stem cells and induced pluripotent stem cells to cardiomyocytes: a methods overview.* Circ Res, 2012. **111**(3): p. 344-58.
47. Verma, M.K. and N. Lenka, *Temporal and contextual orchestration of cardiac fate by WNT-BMP synergy and threshold.* J Cell Mol Med, 2010. **14**(8): p. 2094-108.
48. Schneider, V.A. and M. Mercola, *Wnt antagonism initiates cardiogenesis in Xenopus laevis.* Genes Dev, 2001. **15**(3): p. 304-15.
49. Parikh, A., et al., *Signaling Pathways and Gene Regulatory Networks in Cardiomyocyte Differentiation.* Tissue Eng Part B Rev, 2015. **21**(4): p. 377-92.
50. Itoh, N., et al., *Roles of FGF Signals in Heart Development, Health, and Disease.* Front Cell Dev Biol, 2016. **4**: p. 110.
51. Lian, X., et al., *Insulin inhibits cardiac mesoderm, not mesendoderm, formation during cardiac differentiation of human pluripotent stem cells and modulation of canonical Wnt signaling can rescue this inhibition.* Stem Cells, 2013. **31**(3): p. 447-57.
52. Sa, S. and K. McCloskey, *Activin A and BMP4 Signaling for Efficient Cardiac Differentiation of H7 and H9 Human Embryonic Stem Cells.* J Stem Cells Regen Med, 2012. **8**(3): p. 198-202.
53. Wu, S.M., *Mesp1 at the heart of mesoderm lineage specification.* Cell Stem Cell, 2008. **3**(1): p. 1-2.
54. Kuo, C.T., et al., *GATA4 transcription factor is required for ventral morphogenesis and heart tube formation.* Genes Dev, 1997. **11**(8): p. 1048-60.
55. Watt, A.J., et al., *GATA4 is essential for formation of the proepicardium and regulates cardiogenesis.* Proc Natl Acad Sci U S A, 2004. **101**(34): p. 12573-8.
56. Hu, D., et al., *Metabolic Maturation of Human Pluripotent Stem Cell-Derived Cardiomyocytes by Inhibition of HIF1α and LDHA.* Circ Res, 2018. **123**(9): p. 1066-1079.
57. Correia, C., et al., *3D aggregate culture improves metabolic maturation of human pluripotent stem cell derived cardiomyocytes.* Biotechnol Bioeng, 2018. **115**(3): p. 630-644.

58. Atmanli, A., et al., *Multiplex live single-cell transcriptional analysis demarcates cellular functional heterogeneity.* Elife, 2019. **8**.
59. Atmanli, A., D. Hu, and I.J. Domian, *Molecular etching: a novel methodology for the generation of complex micropatterned growth surfaces for human cellular assays.* Adv Healthc Mater, 2014. **3**(11): p. 1759-64.
60. Hwang, B., J.H. Lee, and D. Bang, *Single-cell RNA sequencing technologies and bioinformatics pipelines.* Exp Mol Med, 2018. **50**(8): p. 96.
61. Rotem, A., et al., *Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state.* Nat Biotechnol, 2015. **33**(11): p. 1165-72.
62. Efthymiou, S., A. Manole, and H. Houlden, *Next-generation sequencing in neuromuscular diseases.* Curr Opin Neurol, 2016. **29**(5): p. 527-36.
63. Deepak, S., et al., *Real-Time PCR: Revolutionizing Detection and Expression Analysis of Genes.* Curr Genomics, 2007. **8**(4): p. 234-51.
64. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies.* Mol Cell, 2015. **58**(4): p. 586-97.
65. Goodwin, S., J.D. McPherson, and W.R. McCombie, *Coming of age: ten years of next-generation sequencing technologies.* Nat Rev Genet, 2016. **17**(6): p. 333-51.
66. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
67. Collins, F.S., et al., *A vision for the future of genomics research.* Nature, 2003. **422**(6934): p. 835-47.
68. Pareek, C.S., R. Smoczynski, and A. Tretyn, *Sequencing technologies and genome sequencing.* J Appl Genet, 2011. **52**(4): p. 413-35.
69. Bertelli, C. and G. Greub, *Rapid bacterial genome sequencing: methods and applications in clinical microbiology.* Clin Microbiol Infect, 2013. **19**(9): p. 803-13.
70. Barba, M., H. Czosnek, and A. Hadidi, *Historical perspective, development and applications of next-generation sequencing in plant virology.* Viruses, 2014. **6**(1): p. 106-36.
71. Ekblom, R. and J. Galindo, *Applications of next generation sequencing in molecular ecology of non-model organisms.* Heredity (Edinb), 2011. **107**(1): p. 1-15.
72. Blencowe, B.J., S. Ahmad, and L.J. Lee, *Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes.* Genes Dev, 2009. **23**(12): p. 1379-86.
73. Bhargava, V., et al., *Technical variations in low-input RNA-seq methodologies.* Sci Rep, 2014. **4**: p. 3678.
74. Combs, P.A. and M.B. Eisen, *Low-cost, low-input RNA-seq protocols perform nearly as well as high-input protocols.* PeerJ, 2015. **3**: p. e869.
75. Buenrostro, J.D., et al., *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nat Methods, 2013. **10**(12): p. 1213-8.
76. Gilfillan, G.D., et al., *Limitations and possibilities of low cell number ChIP-seq.* BMC Genomics, 2012. **13**: p. 645.
77. Tang, F., et al., *mRNA-Seq whole-transcriptome analysis of a single cell.* Nat Methods, 2009. **6**(5): p. 377-82.
78. Andersson, A., et al., *Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography.* Commun Biol, 2020. **3**(1): p. 565.

79. Achim, K., et al., *High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.* Nat Biotechnol, 2015. **33**(5): p. 503-9.
80. Bolander, J., et al., *Single-cell characterization and metabolic profiling of in vitro cultured human skeletal progenitors with enhanced in vivo bone forming capacity.* Stem Cells Transl Med, 2020. **9**(3): p. 389-402.
81. Cuomo, A.S.E., et al., *Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression.* Nat Commun, 2020. **11**(1): p. 810.
82. Gaspar-Maia, A., et al., *Open chromatin in pluripotency and reprogramming.* Nat Rev Mol Cell Biol, 2011. **12**(1): p. 36-47.
83. Illingworth, R.S., et al., *Orphan CpG islands identify numerous conserved promoters in the mammalian genome.* PLoS Genet, 2010. **6**(9): p. e1001134.
84. Saxonov, S., P. Berg, and D.L. Brutlag, *A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters.* Proc Natl Acad Sci U S A, 2006. **103**(5): p. 1412-7.
85. Zentner, G.E. and S. Henikoff, *High-resolution digital profiling of the epigenome.* Nat Rev Genet, 2014. **15**(12): p. 814-27.
86. Smallwood, S.A., et al., *Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity.* Nat Methods, 2014. **11**(8): p. 817-20.
87. Angermueller, C., et al., *Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity.* Nat Methods, 2016. **13**(3): p. 229-32.
88. Billing, A.M., et al., *A Systems-level Characterization of the Differentiation of Human Embryonic Stem Cells into Mesenchymal Stem Cells.* Mol Cell Proteomics, 2019. **18**(10): p. 1950-1966.
89. Ohtani, K. and S. Dimmeler, *Epigenetic regulation of cardiovascular differentiation.* Cardiovasc Res, 2011. **90**(3): p. 404-12.
90. Zhou, Y., et al., *Epigenetic modifications of stem cells: a paradigm for the control of cardiac progenitor cells.* Circ Res, 2011. **109**(9): p. 1067-81.
91. Kodzius, R., et al., *CAGE: cap analysis of gene expression.* Nat Methods, 2006. **3**(3): p. 211-22.
92. Carninci, P., et al., *Genome-wide analysis of mammalian promoter architecture and evolution.* Nat Genet, 2006. **38**(6): p. 626-35.
93. Mikkelsen, T.S., et al., *Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.* Nature, 2007. **448**(7153): p. 553-60.
94. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
95. Calo, E. and J. Wysocka, *Modification of enhancer chromatin: what, how, and why?* Mol Cell, 2013. **49**(5): p. 825-37.
96. Harmston, N. and B. Lenhard, *Chromatin and epigenetic features of long-range gene regulation.* Nucleic Acids Res, 2013. **41**(15): p. 7185-99.
97. Zhang, M., et al., *Roles of N6-Methyladenosine (m.* Front Cell Dev Biol, 2020. **8**: p. 782.
98. Yue, Y., J. Liu, and C. He, *RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation.* Genes Dev, 2015. **29**(13): p. 1343-55.
99. Mao, Y., et al., *m.* Nat Commun, 2019. **10**(1): p. 5332.
100. Schaefer, M., et al., *RNA cytosine methylation analysis by bisulfite sequencing.* Nucleic Acids Res, 2009. **37**(2): p. e12.

101. Cong, L., et al., *Multiplex genome engineering using CRISPR/Cas systems.* Science, 2013. **339**(6121): p. 819-23.
102. Kennedy, E.M., et al., *Optimization of a multiplex CRISPR/Cas system for use as an antiviral therapeutic.* Methods, 2015. **91**: p. 82-86.
103. Zhang, F., Y. Wen, and X. Guo, *CRISPR/Cas9 for genome editing: progress, implications and challenges.* Hum Mol Genet, 2014. **23**(R1): p. R40-6.
104. Sidik, S.M., D. Huet, and S. Lourido, *CRISPR-Cas9-based genome-wide screening of Toxoplasma gondii.* Nat Protoc, 2018. **13**(1): p. 307-323.
105. Churko, J.M., et al., *Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases.* Circ Res, 2013. **112**(12): p. 1613-23.
106. Chaudhry, F., et al., *Single-Cell RNA Sequencing of the Cardiovascular System: New Looks for Old Diseases.* Front Cardiovasc Med, 2019. **6**: p. 173.
107. Chaturvedi, P. and S.C. Tyagi, *Epigenetic mechanisms underlying cardiac degeneration and regeneration.* Int J Cardiol, 2014. **173**(1): p. 1-11.
108. Friedman, C.E., et al., *Single-Cell Transcriptomic Analysis of Cardiac Differentiation from Human PSCs Reveals HOPX-Dependent Cardiomyocyte Maturation.* Cell Stem Cell, 2018. **23**(4): p. 586-598.e8.
109. McCracken, I.R., et al., *Transcriptional dynamics of pluripotent stem cell-derived endothelial cell differentiation revealed by single-cell RNA sequencing.* Eur Heart J, 2020. **41**(9): p. 1024-1036.
110. Selewa, A., et al., *Systematic Comparison of High-throughput Single-Cell and Single-Nucleus Transcriptomes during Cardiomyocyte Differentiation.* Sci Rep, 2020. **10**(1): p. 1535.
111. Govarthanan, K., et al., *DNA methylation microarray uncovers a permissive methylome for cardiomyocyte differentiation in human mesenchymal stem cells.* Genomics, 2020. **112**(2): p. 1384-1395.
112. Bhuvanalakshmi, G., et al., *Epigenetic reprogramming converts human Wharton's jelly mesenchymal stem cells into functional cardiomyocytes by differential regulation of Wnt mediators.* Stem Cell Res Ther, 2017. **8**(1): p. 185.
113. Walravens, A.S., et al., *Molecular signature of progenitor cells isolated from young and adult human hearts.* Sci Rep, 2018. **8**(1): p. 9266.
114. Quaife-Ryan, G.A., et al., *Multicellular Transcriptional Analysis of Mammalian Heart Regeneration.* Circulation, 2017. **136**(12): p. 1123-1139.
115. Tucker, N.R., et al., *Transcriptional and Cellular Diversity of the Human Heart.* Circulation, 2020.
116. Maurano, M.T., et al., *Role of DNA Methylation in Modulating Transcription Factor Occupancy.* Cell Rep, 2015. **12**(7): p. 1184-95.
117. Medvedeva, Y.A., et al., *Effects of cytosine methylation on transcription factor binding sites.* BMC Genomics, 2014. **15**: p. 119.
118. Adey, A., et al., *Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition.* Genome Biol, 2010. **11**(12): p. R119.
119. Reznikoff, W.S., *Tn5 as a model for understanding DNA transposition.* Mol Microbiol, 2003. **47**(5): p. 1199-206.
120. Wang, K., et al., *Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer.* Nat Genet, 2014. **46**(6): p. 573-82.

121. Zahir, F.R., et al., *Comprehensive whole genome sequence analyses yields novel genetic and structural insights for Intellectual Disability.* BMC Genomics, 2017. **18**(1): p. 403.
122. Picelli, S., et al., *Tn5 transposase and tagmentation procedures for massively scaled sequencing projects.* Genome Res, 2014. **24**(12): p. 2033-40.
123. Hrdlickova, R., M. Toloue, and B. Tian, *RNA-Seq methods for transcriptome analysis.* Wiley Interdiscip Rev RNA, 2017. **8**(1).
124. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.
125. Saadatpour, A., et al., *Single-Cell Analysis in Cancer Genomics.* Trends Genet, 2015. **31**(10): p. 576-86.
126. Bao, G., W.J. Rhee, and A. Tsourkas, *Fluorescent probes for live-cell RNA detection.* Annu Rev Biomed Eng, 2009. **11**: p. 25-47.
127. Sandberg, R., *Entering the era of single-cell transcriptomics in biology and medicine.* Nat Methods, 2014. **11**(1): p. 22-4.
128. Heiman, M., et al., *Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP).* Nat Protoc, 2014. **9**(6): p. 1282-91.
129. Cao, Y., et al., *Nondestructive nanostraw intracellular sampling for longitudinal cell monitoring.* Proc Natl Acad Sci U S A, 2017. **114**(10): p. E1866-E1874.
130. Demonte, D., et al., *Structure-based engineering of streptavidin monomer with a reduced biotin dissociation rate.* Proteins, 2013. **81**(9): p. 1621-33.
131. Lim, K.H., et al., *Stable, high-affinity streptavidin monomer for protein labeling and monovalent biotin detection.* Biotechnol Bioeng, 2013. **110**(1): p. 57-67.
132. Klompe, S.E., et al., *Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration.* Nature, 2019. **571**(7764): p. 219-225.
133. Reznikoff, W.S., *The Tn5 transposon.* Annu Rev Microbiol, 1993. **47**: p. 945-63.
134. Ravindran, S., *Barbara McClintock and the discovery of jumping genes.* Proc Natl Acad Sci U S A, 2012. **109**(50): p. 20198-9.
135. Ringertz, N.R., *[The discovery of "jumping genes" in corn gave the entire Nobel prize to a 81-year woman (Barbara McClintock)].* Lakartidningen, 1983. **80**(42): p. 3908-10.
136. Muñoz-López, M. and J.L. García-Pérez, *DNA transposons: nature and applications in genomics.* Curr Genomics, 2010. **11**(2): p. 115-28.
137. Berg, D.E., *Julian Davies and the discovery of kanamycin resistance transposon Tn5.* J Antibiot (Tokyo), 2017. **70**(4): p. 339-346.
138. Berg, D.E., et al., *Transposition of R factor genes to bacteriophage lambda.* Proc Natl Acad Sci U S A, 1975. **72**(9): p. 3628-32.
139. Casacuberta, E. and J. González, *The impact of transposable elements in environmental adaptation.* Mol Ecol, 2013. **22**(6): p. 1503-17.
140. Joly-Lopez, Z., et al., *Phylogenetic and Genomic Analyses Resolve the Origin of Important Plant Genes Derived from Transposable Elements.* Mol Biol Evol, 2016. **33**(8): p. 1937-56.
141. Reznikoff, W.S., *Transposon Tn5.* Annu Rev Genet, 2008. **42**: p. 269-86.
142. Goryshin, I.Y. and W.S. Reznikoff, *Tn5 in vitro transposition.* J Biol Chem, 1998. **273**(13): p. 7367-74.
143. Saint-Girons, I., et al., *Integration specificity of an artificial kanamycin transposon constructed by the in vitro insertion of an internal Tn5 fragment into IS2.* Mol Gen Genet, 1981. **183**(1): p. 45-50.

144. Wiegand, T.W. and W.S. Reznikoff, *Characterization of two hypertransposing Tn5 mutants.* J Bacteriol, 1992. **174**(4): p. 1229-39.
145. Davies, D.R., et al., *The three-dimensional structure of a Tn5 transposase-related protein determined to 2.9-A resolution.* J Biol Chem, 1999. **274**(17): p. 11904-13.
146. Lovell, S., et al., *Two-metal active site binding of a Tn5 transposase synaptic complex.* Nat Struct Biol, 2002. **9**(4): p. 278-81.
147. Gradman, R.J. and W.S. Reznikoff, *Tn5 synaptic complex formation: role of transposase residue W450.* J Bacteriol, 2008. **190**(4): p. 1484-7.
148. Miller, A.K. and F. Tausig, *Biotin-binding by parenterally-administered streptavidin or avidin.* Biochem Biophys Res Commun, 1964. **14**: p. 210-4.
149. Tausig, F. and F.J. Wolf, *Streptavidin--a substance with avidin-like properties produced by microorganisms.* Biochem Biophys Res Commun, 1964. **14**: p. 205-9.
150. Avrantinis, S.K., et al., *Dissecting the streptavidin-biotin interaction by phage-displayed shotgun scanning.* Chembiochem, 2002. **3**(12): p. 1229-34.
151. Howarth, M., et al., *A monovalent streptavidin with a single femtomolar biotin binding site.* Nat Methods, 2006. **3**(4): p. 267-73.
152. Howarth, M. and A.Y. Ting, *Imaging proteins in live mammalian cells with biotin ligase and monovalent streptavidin.* Nat Protoc, 2008. **3**(3): p. 534-45.
153. Demonte, D., C.M. Dundas, and S. Park, *Expression and purification of soluble monomeric streptavidin in Escherichia coli.* Appl Microbiol Biotechnol, 2014. **98**(14): p. 6285-95.
154. Hashimoto, N., et al., *Selective elimination of a B cell subset having acceptor site(s) for T cell-replacing factor (TRF) with biotinylated antibody to the acceptor site(s) and avidin-ricin A-chain conjugate.* J Immunol, 1984. **132**(1): p. 129-35.
155. Balakrishnan, R., et al., *A guide to best practices for Gene Ontology (GO) manual annotation.* Database (Oxford), 2013. **2013**: p. bat054.
156. Cuatrecasas, P., M. Wilchek, and C.B. Anfinsen, *Selective enzyme purification by affinity chromatography.* Proc Natl Acad Sci U S A, 1968. **61**(2): p. 636-43.
157. Levin-Kravets, O., et al., *E. coli-Based Selection and Expression Systems for Discovery, Characterization, and Purification of Ubiquitylated Proteins.* Methods Mol Biol, 2018. **1844**: p. 155-166.
158. Johansson, H., et al., *Chromatographic equipment for large-scale protein and peptide purification.* Adv Biotechnol Processes, 1988. **8**: p. 127-57.
159. Schwaninger, A.E., M.R. Meyer, and H.H. Maurer, *Gas chromatography-mass spectrometry detection of a norfluoxetine artifact in hydrolyzed urine samples may falsely indicate tranylcypromine ingestion.* J Anal Toxicol, 2010. **34**(1): p. 45-8.
160. Zareh, M.M., et al., *Gradient HPLC Method for Simultaneous Determination of Eight Sartan and Statin Drugs in Their Pure and Dosage Forms.* Pharmaceuticals (Basel), 2020. **13**(2).
161. Melcher, K., *A modular set of prokaryotic and eukaryotic expression vectors.* Anal Biochem, 2000. **277**(1): p. 109-20.
162. Schmitt, J., H. Hess, and H.G. Stunnenberg, *Affinity purification of histidine-tagged proteins.* Mol Biol Rep, 1993. **18**(3): p. 223-30.
163. Robeva, A.S., et al., *Double tagging recombinant A1- and A2A-adenosine receptors with hexahistidine and the FLAG epitope. Development of an efficient generic protein purification procedure.* Biochem Pharmacol, 1996. **51**(4): p. 545-55.

164. Di Napoli, A., et al., *Molecular cloning, expression and purification of protein 2A of hepatitis A virus.* New Microbiol, 2004. **27**(2): p. 105-12.
165. Mahnke Braam, L.A., I.Y. Goryshin, and W.S. Reznikoff, *A mechanism for Tn5 inhibition. carboxyl-terminal dimerization.* J Biol Chem, 1999. **274**(1): p. 86-92.
166. Leblond-Francillard, M., M. Dreyfus, and F. Rougeon, *Isolation of DNA-protein complexes based on streptavidin and biotin interaction.* Eur J Biochem, 1987. **166**(2): p. 351-5.
167. Jacobsen, M.T., et al., *Amine Landscaping to Maximize Protein-Dye Fluorescence and Ultrastable Protein-Ligand Interaction.* Cell Chem Biol, 2017. **24**(8): p. 1040-1047.e4.
168. Steiniger-White, M., I. Rayment, and W.S. Reznikoff, *Structure/function insights into Tn5 transposition.* Curr Opin Struct Biol, 2004. **14**(1): p. 50-7.
169. Steiniger-White, M. and W.S. Reznikoff, *The C-terminal alpha helix of Tn5 transposase is required for synaptic complex formation.* J Biol Chem, 2000. **275**(30): p. 23127-33.
170. Sharpless, N.E. and M. Flavin, *The reactions of amines and amino acids with maleimides. Structure of the reaction products deduced from infrared and nuclear magnetic resonance spectroscopy.* Biochemistry, 1966. **5**(9): p. 2963-71.
171. Ravasco, J.M.J.M., et al., *Bioconjugation with Maleimides: A Useful Tool for Chemical Biology.* Chemistry, 2019. **25**(1): p. 43-59.
172. Wang, Q., et al., *Tagmentation-based whole-genome bisulfite sequencing.* Nat Protoc, 2013. **8**(10): p. 2022-32.
173. Weichenhan, D., et al., *Tagmentation-Based Library Preparation for Low DNA Input Whole Genome Bisulfite Sequencing.* Methods Mol Biol, 2018. **1708**: p. 105-122.
174. Valadi, H., et al., *Exosome-mediated transfer of mRNAs and microRNAs is a novel mechanism of genetic exchange between cells.* Nat Cell Biol, 2007. **9**(6): p. 654-9.
175. van Balkom, B.W., et al., *Quantitative and qualitative analysis of small RNAs in human endothelial cells and exosomes provides insights into localized RNA processing, degradation and sorting.* J Extracell Vesicles, 2015. **4**: p. 26760.
176. Pérez-Boza, J., M. Lion, and I. Struman, *Exploring the RNA landscape of endothelial exosomes.* RNA, 2018. **24**(3): p. 423-435.
177. Wei, Z., et al., *Coding and noncoding landscape of extracellular RNA released by human glioma stem cells.* Nat Commun, 2017. **8**(1): p. 1145.
178. Mateescu, B., et al., *Obstacles and opportunities in the functional analysis of extracellular vesicle RNA - an ISEV position paper.* J Extracell Vesicles, 2017. **6**(1): p. 1286095.
179. Yim, N., et al., *Exosome engineering for efficient intracellular delivery of soluble proteins using optically reversible protein-protein interaction module.* Nat Commun, 2016. **7**: p. 12277.
180. Wang, X., et al., *N6-methyladenosine-dependent regulation of messenger RNA stability.* Nature, 2014. **505**(7481): p. 117-20.
181. Shi, H., et al., *YTHDF3 facilitates translation and decay of N6-methyladenosine-modified RNA.* Cell Res, 2017. **27**(3): p. 315-328.
182. Wang, X., et al., *N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency.* Cell, 2015. **161**(6): p. 1388-99.
183. Patil, D.P., B.F. Pickering, and S.R. Jaffrey, *Reading m6A in the Transcriptome:m6A-Binding Proteins.* Trends Cell Biol, 2018. **28**(2): p. 113-127.

184. Tang, J.C., et al., *A nanobody-based system using fluorescent proteins as scaffolds for cell-specific gene manipulation.* Cell, 2013. **154**(4): p. 928-39.
185. Kirchhofer, A., et al., *Modulation of protein properties in living cells using nanobodies.* Nat Struct Mol Biol, 2010. **17**(1): p. 133-8.
186. Kowal, J., et al., *Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes.* Proc Natl Acad Sci U S A, 2016. **113**(8): p. E968-77.
187. Matarredona, E.R. and A.M. Pastor, *Extracellular Vesicle-Mediated Communication between the Glioblastoma and Its Microenvironment.* Cells, 2019. **9**(1).
188. Kalluri, R. and V.S. LeBleu, *The biology.* Science, 2020. **367**(6478).
189. Kalluri, R., *The biology and function of exosomes in cancer.* J Clin Invest, 2016. **126**(4): p. 1208-15.
190. Zhang, Y., et al., *Exosomes: biogenesis, biologic function and clinical potential.* Cell Biosci, 2019. **9**: p. 19.
191. Greening, D.W., et al., *A protocol for exosome isolation and characterization: evaluation of ultracentrifugation, density-gradient separation, and immunoaffinity capture methods.* Methods Mol Biol, 2015. **1295**: p. 179-209.
192. Théry, C., et al., *Minimal information for studies of extracellular vesicles 2018 (MISEV2018): a position statement of the International Society for Extracellular Vesicles and update of the MISEV2014 guidelines.* J Extracell Vesicles, 2018. **7**(1): p. 1535750.
193. Ståhl, A.L., et al., *Exosomes and microvesicles in normal physiology, pathophysiology, and renal diseases.* Pediatr Nephrol, 2019. **34**(1): p. 11-30.
194. Muralidharan-Chari, V., et al., *Microvesicles: mediators of extracellular communication during cancer progression.* J Cell Sci, 2010. **123**(Pt 10): p. 1603-11.
195. Tricarico, C., J. Clancy, and C. D'Souza-Schorey, *Biology and biogenesis of shed microvesicles.* Small GTPases, 2017. **8**(4): p. 220-232.
196. Schmid, M. and T.H. Jensen, *The exosome: a multipurpose RNA-decay machine.* Trends Biochem Sci, 2008. **33**(10): p. 501-10.
197. Valencia, K. and F. Lecanda, *Microvesicles: Isolation, Characterization for In Vitro and In Vivo Procedures.* Methods Mol Biol, 2016. **1372**: p. 181-92.
198. Battistelli, M. and E. Falcieri, *Apoptotic Bodies: Particular Extracellular Vesicles Involved in Intercellular Communication.* Biology (Basel), 2020. **9**(1).
199. Xu, X., Y. Lai, and Z.C. Hua, *Apoptosis and apoptotic body: disease message and therapeutic target potentials.* Biosci Rep, 2019. **39**(1).
200. Caruso, S. and I.K.H. Poon, *Apoptotic Cell-Derived Extracellular Vesicles: More Than Just Debris.* Front Immunol, 2018. **9**: p. 1486.
201. Tang, Y.T., et al., *Comparison of isolation methods of exosomes and exosomal RNA from cell culture medium and serum.* Int J Mol Med, 2017. **40**(3): p. 834-844.
202. Batagov, A.O. and I.V. Kurochkin, *Exosomes secreted by human cells transport largely mRNA fragments that are enriched in the 3'-untranslated regions.* Biol Direct, 2013. **8**: p. 12.
203. Bland, C.L., et al., *Exosomes derived from B16F0 melanoma cells alter the transcriptome of cytotoxic T cells that impacts mitochondrial respiration.* FEBS J, 2018. **285**(6): p. 1033-1050.

204. Hinger, S.A., et al., *Diverse Long RNAs Are Differentially Sorted into Extracellular Vesicles Secreted by Colorectal Cancer Cells.* Cell Rep, 2018. **25**(3): p. 715-725.e4.
205. Turchinovich, A., O. Drapkina, and A. Tonevitsky, *Transcriptome of Extracellular Vesicles: State-of-the-Art.* Front Immunol, 2019. **10**: p. 202.
206. Spinelli, C., et al., *Extracellular Vesicles as Conduits of Non-Coding RNA Emission and Intercellular Transfer in Brain Tumors.* Noncoding RNA, 2018. **5**(1).
207. Margolis, L. and Y. Sadovsky, *The biology of extracellular vesicles: The known unknowns.* PLoS Biol, 2019. **17**(7): p. e3000363.
208. Das, S., et al., *The Extracellular RNA Communication Consortium: Establishing Foundational Knowledge and Technologies for Extracellular RNA Research.* Cell, 2019. **177**(2): p. 231-242.
209. Kim, H., et al., *Engineered extracellular vesicles and their mimetics for clinical translation.* Methods, 2019.
210. Zickler, A.M. and S. El Andaloussi, *Functional extracellular vesicles aplenty.* Nat Biomed Eng, 2020. **4**(1): p. 9-11.
211. Qiu, G., et al., *Functional proteins of mesenchymal stem cell-derived extracellular vesicles.* Stem Cell Res Ther, 2019. **10**(1): p. 359.
212. Jayaseelan, V.P., *Emerging role of exosomes as promising diagnostic tool for cancer.* Cancer Gene Ther, 2019.
213. Carnino, J.M., H. Lee, and Y. Jin, *Isolation and characterization of extracellular vesicles from Broncho-alveolar lavage fluid: a review and comparison of different methods.* Respir Res, 2019. **20**(1): p. 240.
214. Raposo, G. and W. Stoorvogel, *Extracellular vesicles: exosomes, microvesicles, and friends.* J Cell Biol, 2013. **200**(4): p. 373-83.
215. Borrelli, D.A., et al., *Extracellular vesicle therapeutics for liver disease.* J Control Release, 2018. **273**: p. 86-98.
216. Leidal, A.M., et al., *The LC3-conjugation machinery specifies the loading of RNA-binding proteins into extracellular vesicles.* Nat Cell Biol, 2020. **22**(2): p. 187-199.
217. Hung, M.E. and J.N. Leonard, *A platform for actively loading cargo RNA to elucidate limiting steps in EV-mediated delivery.* J Extracell Vesicles, 2016. **5**: p. 31027.
218. Yim, N. and C. Choi, *Extracellular vesicles as novel carriers for therapeutic molecules.* BMB Rep, 2016. **49**(11): p. 585-586.
219. Morales-Kastresana, A., et al., *High-fidelity detection and sorting of nanoscale vesicles in viral disease and cancer.* J Extracell Vesicles, 2019. **8**(1): p. 1597603.
220. Ariotti, N., et al., *Modular Detection of GFP-Labeled Proteins for Rapid Screening by Electron Microscopy in Cells and Organisms.* Dev Cell, 2015. **35**(4): p. 513-25.
221. Hentze, M.W., et al., *A brave new world of RNA-binding proteins.* Nat Rev Mol Cell Biol, 2018. **19**(5): p. 327-341.
222. Frye, M., et al., *RNA modifications modulate gene expression during development.* Science, 2018. **361**(6409): p. 1346-1349.
223. Zeidan, Q., et al., *Conserved mRNA-granule component Scd6 targets Dhh1 to repress translation initiation and activates Dcp2-mediated mRNA decay in vivo.* PLoS Genet, 2018. **14**(12): p. e1007806.
224. Montero, H., R. García-Román, and S.I. Mora, *eIF4E as a control target for viruses.* Viruses, 2015. **7**(2): p. 739-50.

225. Borden, K.L., *The eukaryotic translation initiation factor eIF4E wears a "cap" for many occasions.* Translation (Austin), 2016. **4**(2): p. e1220899.
226. von der Haar, T., et al., *The mRNA cap-binding protein eIF4E in post-transcriptional gene expression.* Nat Struct Mol Biol, 2004. **11**(6): p. 503-11.
227. Choi, Y.H. and C.H. Hagedorn, *Purifying mRNAs with a high-affinity eIF4E mutant identifies the short 3' poly(A) end phenotype.* Proc Natl Acad Sci U S A, 2003. **100**(12): p. 7033-8.
228. Friedland, D.E., et al., *A mutant of eukaryotic protein synthesis initiation factor eIF4E(K119A) has an increased binding affinity for both m7G cap analogues and eIF4G peptides.* Biochemistry, 2005. **44**(11): p. 4546-50.
229. O'Leary, S.E., et al., *Dynamic recognition of the mRNA cap by Saccharomyces cerevisiae eIF4E.* Structure, 2013. **21**(12): p. 2197-207.
230. Du, H., et al., *YTHDF2 destabilizes m(6)A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex.* Nat Commun, 2016. **7**: p. 12626.
231. Liao, S., H. Sun, and C. Xu, *YTH Domain: A Family of N.* Genomics Proteomics Bioinformatics, 2018. **16**(2): p. 99-107.
232. Roignant, J.Y. and M. Soller, *m.* Trends Genet, 2017. **33**(6): p. 380-390.
233. Louloupi, A., et al., *Transient N-6-Methyladenosine Transcriptome Sequencing Reveals a Regulatory Role of m6A in Splicing Efficiency.* Cell Rep, 2018. **23**(12): p. 3429-3437.
234. Dominissini, D., et al., *Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing.* Nat Protoc, 2013. **8**(1): p. 176-89.
235. Shi, H., et al., *m.* Nature, 2018. **563**(7730): p. 249-253.
236. Xu, C., et al., *Structural Basis for the Discriminative Recognition of N6-Methyladenosine RNA by the Human YT521-B Homology Domain Family of Proteins.* J Biol Chem, 2015. **290**(41): p. 24902-13.
237. Xu, C., et al., *Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain.* Nat Chem Biol, 2014. **10**(11): p. 927-9.
238. Heiman, M., et al., *A translational profiling approach for the molecular characterization of CNS cell types.* Cell, 2008. **135**(4): p. 738-48.
239. Zhou, P., et al., *Interrogating translational efficiency and lineage-specific transcriptomes using ribosome affinity purification.* Proc Natl Acad Sci U S A, 2013. **110**(38): p. 15395-400.
240. Picelli, S., et al., *Full-length RNA-seq from single cells using Smart-seq2.* Nat Protoc, 2014. **9**(1): p. 171-81.
241. Stork, C. and S. Zheng, *Genome-Wide Profiling of RNA-Protein Interactions Using CLIP-Seq.* Methods Mol Biol, 2016. **1421**: p. 137-51.
242. Park, P.J., *ChIP-seq: advantages and challenges of a maturing technology.* Nat Rev Genet, 2009. **10**(10): p. 669-80.
243. Legnini, I., et al., *FLAM-seq: full-length mRNA sequencing reveals principles of poly(A) tail length control.* Nat Methods, 2019. **16**(9): p. 879-886.
244. Ziegenhain, C., et al., *Comparative Analysis of Single-Cell RNA Sequencing Methods.* Mol Cell, 2017. **65**(4): p. 631-643.e4.
245. Picelli, S., *Single-cell RNA-sequencing: The future of genome biology is now.* RNA Biol, 2017. **14**(5): p. 637-650.

246. Picelli, S., et al., *Smart-seq2 for sensitive full-length transcriptome profiling in single cells.* Nat Methods, 2013. **10**(11): p. 1096-8.
247. Song, Y., et al., *A comparative analysis of library prep approaches for sequencing low input translatome samples.* BMC Genomics, 2018. **19**(1): p. 696.
248. Osborn, M.J., et al., *A picornaviral 2A-like sequence-based tricistronic vector allowing for high-level therapeutic gene expression coupled to a dual-reporter system.* Mol Ther, 2005. **12**(3): p. 569-74.
249. Dominissini, D., et al., *Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq.* Nature, 2012. **485**(7397): p. 201-6.
250. Morales-Kastresana, A., et al., *Labeling Extracellular Vesicles for Nanoscale Flow Cytometry.* Sci Rep, 2017. **7**(1): p. 1878.
251. Miller, J.E. and J.C. Reese, *Ccr4-Not complex: the control freak of eukaryotic cells.* Crit Rev Biochem Mol Biol, 2012. **47**(4): p. 315-33.
252. Hu, B., et al., *POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins.* Nucleic Acids Res, 2017. **45**(D1): p. D104-D114.
253. Zhu, Y., et al., *POSTAR2: deciphering the post-transcriptional regulatory logics.* Nucleic Acids Res, 2019. **47**(D1): p. D203-D211.
254. Koh, C.W.Q., et al., *Single-nucleotide-resolution sequencing of human N6-methyldeoxyadenosine reveals strand-asymmetric clusters associated with SSBP1 on the mitochondrial genome.* Nucleic Acids Res, 2018. **46**(22): p. 11659-11670.
255. Drukker, M., et al., *Human embryonic stem cells and their differentiated derivatives are less susceptible to immune rejection than adult cells.* Stem Cells, 2006. **24**(2): p. 221-9.
256. Tang, F., K. Lao, and M.A. Surani, *Development and applications of single-cell transcriptome analysis.* Nat Methods, 2011. **8**(4 Suppl): p. S6-11.
257. Saliba, A.E., et al., *Single-cell RNA-seq: advances and future challenges.* Nucleic Acids Res, 2014. **42**(14): p. 8845-60.

Illustrations Table

Figures

Figure 1. Nuclear dynamics and proliferative capacity of cardiomyocytes during growth. 23

Figure 2. Cell transplantation techniques and proposed mechanisms of cell therapy for heart regeneration 25

Figure 3. Model of Differentiation of Human PSC via Sequential Progenitors to Cardiomyocytes regeneration 31

Figure 4. Timeline and history of development of genomics and data impact. 40

Figure 5. How to sequence DNA. 40

Figure 6. Polymerase Chain Reaction steps. 41

Figure 7. Milestones in whole genome sequencing. 43

Figure 8. Single-cell isolation and library preparation. 45

Figure 9. Chromatin condensation - euchromatin and heterochromatin 48

Figure 10. Euchromatin and heterochromatin condensation process 49

Figure 11 N6-methyladenosine (m^6A) 53

Figure 12. Presentation of the new methodology 61

Figure 13. Wild type Tn5 mechanism of Transposition 65

Figure 14. EZ: Tn5® mechanism of action 67

Figure 15. Nextera® sequencing kit mechanism of action 68

Figure 16. Predicted structure of Tn5-mSA by iTasser 79

Figure 17. Predicted structure of mSA-Tn5 by iTasser 80

Figure 18. Predicted Binding site of Tn5-mSA and mSA-Tn5 by iTasser 80
Figure 19. Neb IMPACT® technology summary 84
Figure 20. Tn5 production in C3013 Cells total lysate analysis 86
Figure 21. Tn5 production optimization on [IPTG] and induction time 86
Figure 22. Tn5 production optimization on [NaCl] 87
Figure 23. Tn5 production optimization on DTT cleavage time 87
Figure 24. Gel of production of the Tn5/mSA DTT cleavage method 88
Figure 25. Gel of production of the Tn5 DTT cleavage method 89
Figure 26. Gel of production of the Tn5 MESNA (200mM) cleavage method 90
Figure 27. Tn5/mSA synaptic complex formation verify by EMSA 92
Figure 28. Tn5/mSA activity test by tagmentation reaction 93
Figure 29. mSA activity for both Tn5/mSA isoforms test by EMSA gel 95
Figure 30. Maleimide Diels-Alder reaction on a Thiols group 97
Figure 31. Maleimide-Alexa488 Diels-Alder reaction on the Tn5 98
Figure 32. Multiple binding reactions with Mal-PEG, Cys-PEG and NHS-PEG on the Tn5 100
Figure 33. Tn5 Tagmentation assay with fluoresent Alexa647-oligo 103
Figure 34. Tn5 vs DNA with fixed Tn5 quantity and increased genomic DNA quantity 104
Figure 35. Tn5 vs DNA with fixed Tn5 quantity and increased genomic DNA quantity Bioanalyzer results 104
Figure 36. Tn5 vs DNA followed by PCR amplification 105
Figure 37. Transposase sedimented to beads or well wall. 106
Figure 38. Tn5 tagmentation test on long oligo design. 107
Figure 39. NH_2 oligo/antibody binding design strategy. 108

Figure 40. Tn5 tagmentation test on MeB long oligo and NH_2 polyT custom design 109

Figure 41. Tn5 tagmentation test on MeB NH_2 polyT custom design with more or less DNA/Tn5 110

Figure 42. Final process strategy with the MeB NH_2 polyT custom 111

Figure 43. Final process strategy with the Chic-loop oligo design 112

Figure 44. Streptavidin complex formation with custom oligo 113

Figure 45. Streptavidin oligo custom complexed with Tn5 test of tagmentation 114

Followed by PCR amplification 114

Figure 46. Tam-ChiP® sequencing technology develloped by Active Motif™ 116

Figure 47. Tam-ChiP® flow chart 117

Figure 48. TRACE-seq methodology overview 122

Figure 49.Extracellular vesicles: types, sizes, content, biogenesis and uptake 123

Figure 50. EVs secretion pathway 126

Figure 51. Schematic diagram of EXPLOR technology 127

Figure 52. mRNA translation and stability 129

Figure 53. YTHDF group of proteins form an interconnected network in the cytosol. 131

Figure 54. YTHDF mRNA translation and stability 132

Figure 55. Flowchart for Smart-seq2 library preparation. 134

Figure 56. RBP test made by qPCR from different HEK293T cells population 152

Figure 57. Colocalization signal generated by confocal from different HEK293T cells population 152

Figure 58. Overall cloning strategy for TRACE-seq 155

Figure 59. Confocal signal colocalization check on final construct for different HEK293T cells population 157
Figure 60. EVs mRNA content analysis basic protocol 158
Figure 61. Overview and validation/ EVs purification 160
Figure 62. Bioanalyzer analysis of cDNA generated from Microvesicles and their corresponding cDNA from cell lysates (Transient transfected Cells) 163
Figure 63. Bioanalyzer analysis of RNA and cDNA generated from Microvesicles and their corresponding cDNA from cell lysates (Transduced Cells)..... 167
Figure 64. DEseq 2 analysis ran on the 8 selected samples C1 construct TRACE and GFP control 168
Figure 65. Mapping results generate for all 8 sequenced samples. 171
Figure 66. H2O2 stress pathway validation test. RT-qPCR results generated from MVs mRNA and their corresponding cell lysate. 173

Annexed Figure 1. Catalytic site of the Tn5 with the Cys 187..... 183
Annexed Figure 2. Dimerization site of the Tn5 with the Cys 402 183
Annexed Figure 3. Multiple binding reaction with Mal-alexa488 Mal-PEG, Cys-PEG and NHS-PEG on different Tn5 preparation SDS signal..... 184
Annexed Figure 4. Multiple binding reaction with Mal-alexa488 Mal-PEG, Cys-PEG and NHS-PEG on different Tn5 preparation UV signal 185
Annexed Figure 5. Tn5 vs DNA with fixed genomic DNA quantity and increased of Tn5 quantity 186

Annexed Figure 6. Tn5 vs DNA plasmid followed by a PCR amplification with fluorescent oligos alexa647 186
Annexed Figure 7. cDNA chip analysis from transfected HEK cells (500k cells) with the Construct C1, The negative control construct, GFP-O and untransfected cells. 187
Annexed Figure 8. cDNA chip analysis from transfected HEK cells (8M cells) with the Construct C1, The negative control construct, GFP-O and untransfected cells. 188
Annexed Figure 9. Gel prediction made from the bioanalyzer trace Figure 59 188
Annexed Figure.10 TRACE isolation method and RNA-Seq library generation steps..... 189
Annexed Figure.11 TRACE RBP validation RT-qPCR..... 190
Annexed Figure.12 Confocal image from TRACE transfection 190
Annexed Figure.13 Bioanalyzer analysis of RNA content from Microvesicles and their corresponding RNA from cell..... 191
Annexed Figure.14 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell T2A construct..... 192
Annexed Figure.15 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell MVs vs Exosomes..... 193
Annexed Figure.16 NanoFACS analysis of the EVs from Cells transfected with TRACE constructs and control 194
Annexed Figure.17 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell, low amount of cells..... 195
Annexed Figure.18 TRACE-Seq library preparation validation test on RT-qPCR..... 196
Annexed Figure.19 Bioanalyzer analysis of cDNA content from Microvesicles and their corresponding RNA from cell transduce with TRACE C1 constructs 196

Annexed Figure.20 inductions Test of the TRACE/Control cell lines Dox vs No Dox. 197

Annexed Figure.21 DEseq analysis from the 8 selected samples TRACE C1 Construct and GFP Control, Mv and Cell lysate. 198

Annexed Figure.22 H2O2 stress pathway validation test. RT-qPCR results generated from MVs mRNA and their corresponding cell lysate 200

Tables

Table 1: Mix preparation cDNA generation 144

Table 2: Mix preparation First amplification reaction 144

Table 3: Mix preparation tagmentation preparation 145

Table 4: Mix preparation for Tagmented samples followed by PCR amplification reaction 146

Annexed Table 1 Oligo used for cloning for Tn5 project 201

Annexed Table 2 Oligo used for tragmentation for the Tn5 project 201

Annexed Table 3 Oligo used for cloning for TRACE 202

Annexed Table 4 Primers used for qPCR for TRACE 202

Résumé en Français

L'objectif de ce premier projet de thèse était de développer une méthode hautement innovante pour l'analyse épigénétique multiplexée des cellules myocardiques à une résolution cellulaire unique. Après une conception de protocole complète pour générer et purifier la protéine transposase Tn5 en vue d'un travail *in vitro*, nous avons utilisé cette protéine dans une série d'expériences de mise à l'épreuve afin de démontrer sa capacité à définir le paysage épigénétique des cellules de manière hautement multiplexée à la résolution d'une seule cellule. En liant la protéine Tn5 à des anticorps ciblant des facteurs épigénétiques clés, il pourrait être théoriquement possible d'identifier les sites de liaison de facteurs de transcription spécifiques au niveau du génome. Nous avons pu ainsi développer et valider cette approche très innovante par le biais du couple : protéine Tn5 et streptavidine. Il convient de noter que notre approche ne reposait pas sur une liaison directe du Tn5 à l'anticorps cible car cela semblait bloquer l'activité de la transposase Tn5. Au lieu de cela, nous avons utilisé l'oligo transporté par la protéine Tn5 à la fois pour le processus de «tagmentation» et pour la liaison à l'anticorps cible. Avec cette nouvelle approche, il a été prouvé par une série d'expériences *in vitro* que cette nouvelle conception fonctionnerait et nous étions très confiants de l'avancement et du succès de ce projet. Néanmoins, après une lecture attentive de la littérature, nous avons trouvé un brevet récent d'une entreprise qui a développé exactement le même type de technologie. Le design proposé par cette société était en tous points égal : même choix d'utiliser un couple protéine / anticorps par le biais des oligos etc.

Ainsi, ne nous laissant pas abattre, nous avons commencé à développer un second projet de thèse : une approche innovante pour échantillonner l'ARN intracellulaire de manière non biaisée et non destructive. Ainsi, par une analyse minutieuse de la littérature scientifique et également en capitalisant sur cette expérience antérieure, nous avons pu créer cette toute nouvelle technologie permettant une analyse transcriptomique au fil du temps sans aucune destruction cellulaire. Cette approche nommée TRACE (Transcriptomic Analysis by Captured in Extracellular vesicles) repose sur l'analyse d'une partie représentative du transcriptome cellulaire capturé dans les vésicules extracellulaires (Fig.1).

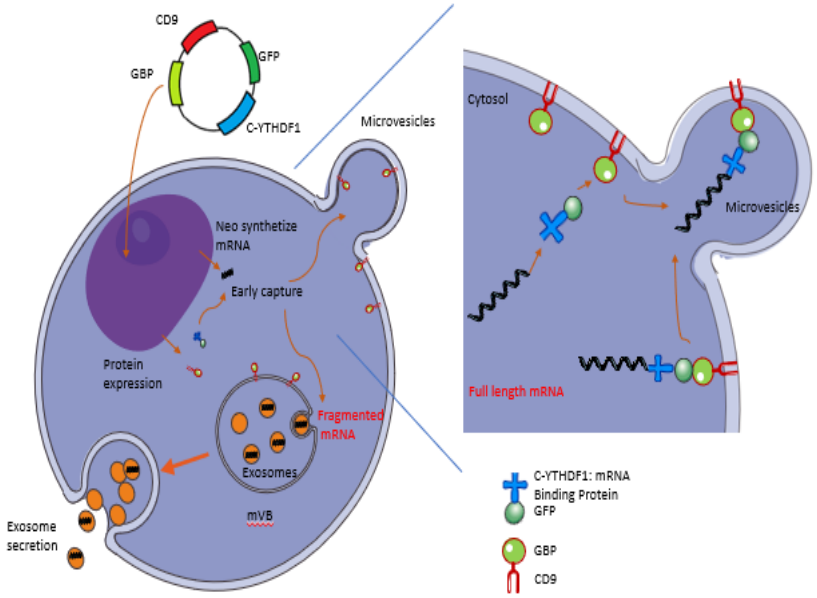


Fig. 1 Vue globale de la methode TRACEseq

Avec cette approche, le transcriptome des cellules qui expriment TRACE peut être suivi dans le temps de manière non destructive aussi bien *in vitro* qu'*in vivo*. Cela constitue un outil de recherche puissant, fondamentale comme translationnelle et a également le potentiel de réduire les coûts, le temps et pourrait être adaptée à de nombreux domaines, en particulier les investigations traitées par nos laboratoires (le développement/caractérisation de cellules souche dans le cadre de la régénération cardiaque). Comme le montre la Figure 1, la « protéine de capture » de l'ARNm consiste en une protéine de fusion composée d'une partie C terminale de la protéine YTHDF1 et de l'amplificateur d'émission fluorescent GFP. Le domaine de la protéine YTH reconnaît le site m^{6}A, l'une des modifications internes les plus abondantes dans l'ARNm eucaryotique. De plus, l'isoforme 1 du domaine de la protéine YTH, améliore l'efficacité de la traduction et se lie à l'ARNm proche de la membrane du noyau après la translocation de

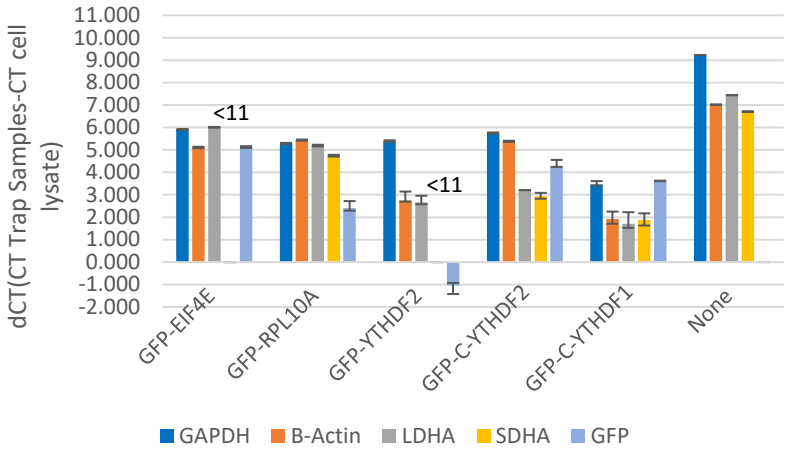


Fig.2 Résultats d'immunoprécipitation génomique de différentes populations de HEK 293T transfectées avec cinq constructions de protéine de liaison à l'ARNm couplées au GFP. Pour chacun d'eux, la cellule a été lysée et l'ARN a été purifié. L'immunoprécipitation a été réalisée contre le GFP avec un anticorps anti GFP et toutes les expressions géniques liées correspondent aux échantillons retenus. Ces populations génomiques « piégées » ont été normalisées par rapport au niveau d'expression de leur propre lysat cellulaire.

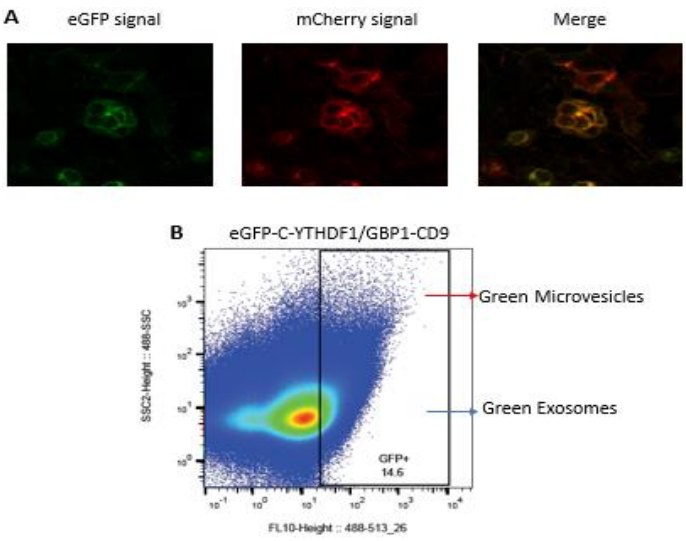


Fig. 3 A. Résultats de transfection de différentes populations HEK 293T avec deux constructions : eGFP-C-YTHDF1 / GBP1-CD9 et mCherry-CD9. Une colocalisation du signal apparaît ce qui prouve la fonction de liaison du couple GBP1 / eGFP. De plus, le signal reste comme prévu sur la membrane cytoplasmique. **B**. NanoFACS de la population de vésicules sécrétée par les HEK 293T préalablement transfectée avec la construction eGFP-C-YTHDF1 / GBP1-CD9.

celui-ci du noyau vers le cytosol. Cette protéine de capture d'ARNm est appariée (importée) aux vésicules extracellulaires (EVs) par le biais d'une autre protéine de fusion. Cette protéine de fusion sert de ciblage aux EVs et est constituée d'une protéine de liaison au GFP (GBP1), qui améliore également le signal de fluorescent vert. De plus, afin de réaliser le ciblage aux EVs, la tretraspanine CD9, marqueur transmembranaire connu des EVs a été utilisé. Ainsi, eGFP-C-YTHDF1 peut piéger l'ARNm nouvellement synthétisé, et l'ensemble du complexe peut être importée dans les EVs via la protéine de fusion GBP1-CD9 (Fig. 1). Les ARNm nouvellement capturés peuvent être ensuite purifiés à partir de milieux de culture cellulaire ou de fluides biologiques par l'isolement de vésicules extracellulaires sécrétées exprimant le GFP et analysés par des méthodes courantes de quantification transcriptomique telle que la q-PCR, microarray ou l'ARN-seq. Afin de valider et d'optimiser cette approche, nous avons testé tous les composants de cette nouvelle construction séparément et en combinaison. Ainsi, nous nous sommes d'abord concentrés sur la protéine de liaison à l'ARN pour déterminer laquelle était la meilleure (Fig. 2). Deuxièmement, il a été testé, le reste de la conception validé par microscopie confocale par le biais d'une vérification de la colocalisation de signale (Fig 3.A). Enfin par NanoFACS ou des vésicules vertes ont bien été détectées (Fig 3.B.). Une fois ces deux composants majeurs validés, l'import d'ARNm a été testée dans une série d'expériences et de preuve de concept (Fig.4,5 et 6).

Pour ce faire, Nous avons dû développer un nouveau protocole base sur les travaux SMART-seq2 (Picelli *et al* Nat Methods 2013). Il a également dû être déterminer le pourcentage et le type d'ARN présents dans les différents types de vésicules, car cela n'avait jamais été clairement traité. Dans la grande majorité des publications, l'expression de l'ARNm dans les exosomes semble être principalement fragmentée et il ne reste qu'une petite fraction (voir rien du tout) de l'ARNm complet. Comme présente sur la Figure 4 notre étude confirme la bibliographie avec une détection d'ARNm de pleine longueur a été détecté uniquement pour les microvesicules et non les exosomes. Il est possible qu'un mécanisme particulier à proximité du corps multivesiculaire interagisse avec l'ARNm lié par TRACE et le fragmente. Ainsi, nous avons décidé de concentrer notre analyse sur les plus grosses vésicules connues sous le nom de microvésicules qui ont un mécanisme différent de voie de sécrétion (Fig. 1). Nous avons répété l'expérience *in vitro* et généré des données convaincantes (Tripliquât) qui nous ont permis d'apprécier le potentiel de cette nouvelle méthodologie. Nous avons également créé une bibliothèque a RNA-seq pour séquencer l'ARNm capturé par la technique TRACE et le comparer avec le lysat cellulaire. La construction TRACE-seq a été analysé ainsi que des microvésicules de control exprimant un transgène sans protéine de capture d'ARNm

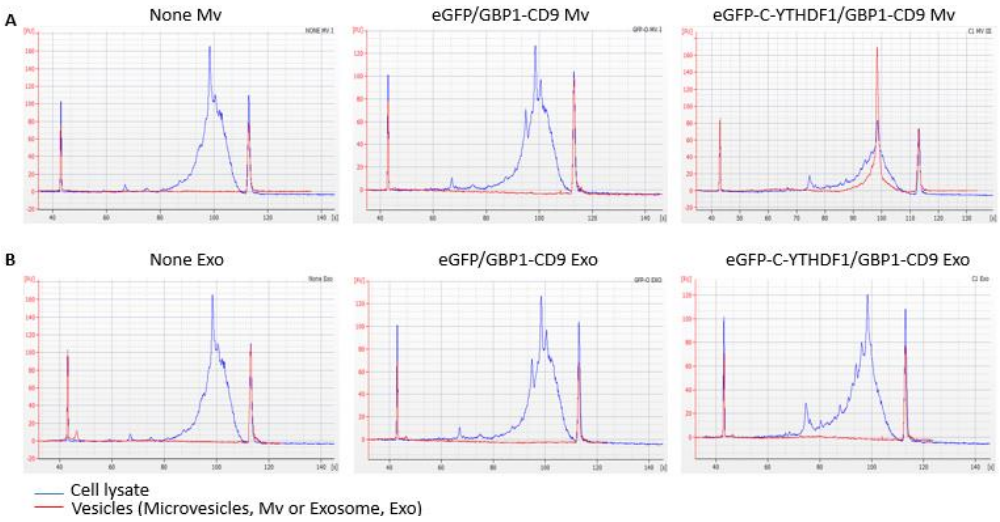


Fig.4 A. Bioanalyseur d'ADNc généré dans différentes populations HEK 293T, analyse de microvésicules : cellule non transfectée (None) Constructions de contrôle négatif (eGFP / GBP1-CD9) et la construction TRACE (eGFP-C-YTHDF1 / GBP1-CD9). **B**. Bioanalyseur d'ADNc génère dans les même populations HEK 293T mais analyse des exosomes. Mêmes échantillons que ci-dessus.

YTHDF1. Comme prévu, cette population d'ARNm extravésiculaire importée est représentative à une large majorité de l'expression d'ARNm cytosolique ce qui n'est pas le cas pour les populations contrôle (Fig 5).

Ainsi, pour valider pleinement la fonctionnalité de notre technologie, qui devrait importer une partie représentative de l'ensemble du transcriptome dans les EVs, nous avons séquencé les échantillons d'ARN à partir des cellules transfectées de manière transitoire. Nous nous sommes concentrés sur une sélection de 3603 gènes transformés via le package limma-voom et analysés par DEseq2. Comme nous pouvons le voir sur la figure 5a, les MV exprimant TRACE-seq ont en commun plus de gènes avec leurs lysats cellulaires. De plus la corrélation

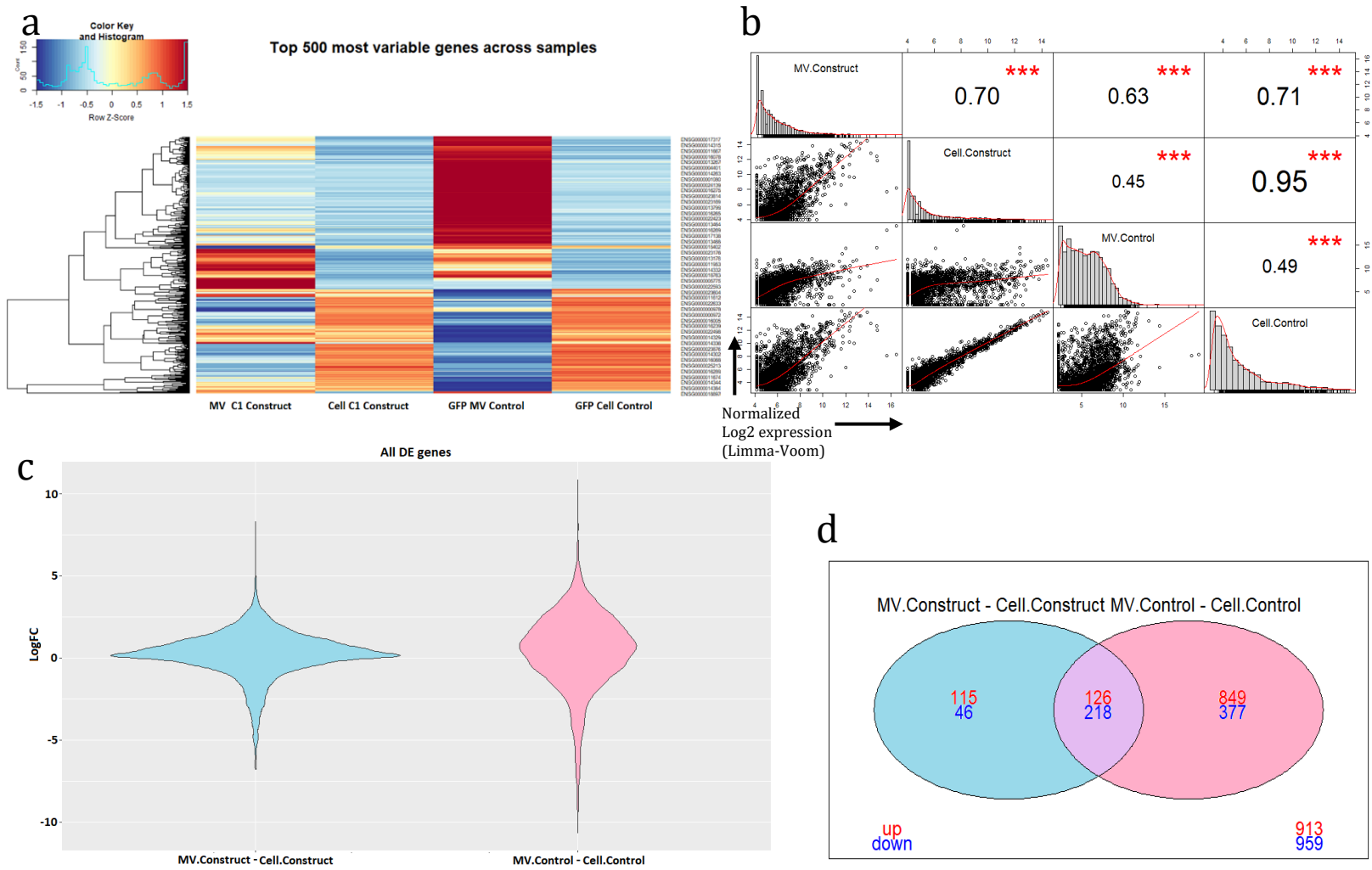


Figure.4 DEseq2 à partir des 8 échantillons sélectionnés. Construction TRACE GFP-C-YTHDF1 / GBP1-CD9: MV dupliqué, lysats cellulaires dupliqués et contrôle GFP / GBP1-CD9: lysats cellulaires dupliqués MV dupliqués. **a**. Heat map pour les 500 gènes les plus variables dans tous les échantillons (doublons regroupés pour générer la Heat map). **b**. Diagramme de corrélation de Pearson de l'analyse Limma-voom DEseq2 entre chaque groupe d'échantillons. **c**. Tous les gènes DEseq2 de l'analyse Limma-voom. **d**. Venndiagram de la matrice de contraste faite à partir de MV-lysat Cell des gènes utilisés pour le Limma-voom DEseq2.

de Pearson est plus haute pour les échantillons TRACE-seq (corrélation 0,7) par rapport aux MV de contrôle GFP (corrélation 0,45) Fig 5b. Nous avons effectué une analyse DEseq2 sur les dénombrements normalisés limma-voom et comme le montre la figure 5c, une net différence dans la variation logarithmique de tous les gènes de DEseq2 (3603 gènes) est présente. En effet, on retrouve une bien meilleure corrélation pour la construction MV C1 - C1 Cell lysats (avec un LogFC proche de 0) par rapport au contrôle MV GFP-GFP Cell lysat (qui a des valeurs LogFC supérieur à 1 / -1 pour la moitié des gènes). Il a également été réalisé un test de matrice de contraste sur chaque population de MV contre leurs propres lysats cellulaires. Comme représenté par le Venn-diagramme Fig 4 d, sans surprise, la différence d'expression génique entre les MV et les cellules est beaucoup plus élevée dans le groupe témoin GFP que dans la construction C1, avec respectivement 505 et 1570 gènes différentiellement exprimés. Dans le contrôle GFP, 62% des gènes étaient significativement plus abondants dans les EVs que dans leurs lysats cellulaires (975 haut et 595 bas), contre 47% des gènes dans la construction C1 (241 haut et 264 bas). L'enrichissement significatif des gènes surabondants dans les EVs témoins (valeur p exacte de Fisher = 10^{-11}) par rapport à la construction C1 (valeur p exacte de Fisher = 0,49) met en évidence le chargement d'ARNm de manière plus égal / stochastique activé par TRACE-seq qui a de ce faite un transcriptome d'EVs qui est un meilleur miroir de celui de la cellule.

Nous avons également effectué une analyse de Mapping pour déterminer la couverture de distribution de chaque échantillon (Fig 6). La distribution de couverture entre les MVs et leur propre lysat cellulaire est beaucoup plus proche pour la construction C1 que pour le groupe témoin. Notamment, les deux groupes de construction C1 sont très proches de 0 (échelle logarithmique pour le rapport ARNm MV / Cellule), ce qui signifie que les MVs et le lysat cellulaire partagent une majorité de fragments très corrélés avec la même couverture. En revanche, les résultats des GFP MVs de contrôle sont beaucoup plus dispersés (à partir de 0) sur cette distribution de couverture à échelle logarithmique pour le rapport ARNm MV /

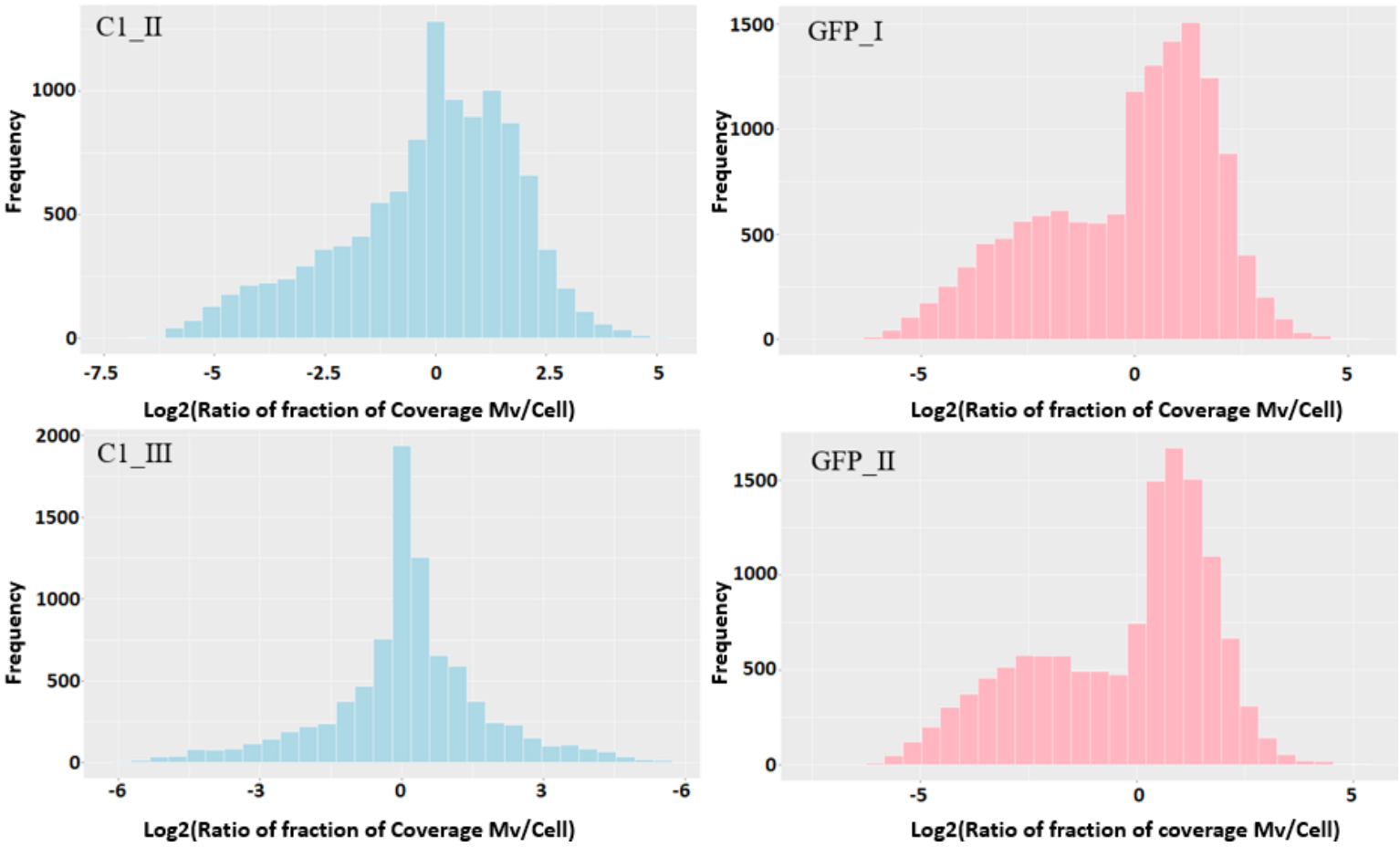


Fig.6 résultats générés pour les 8 échantillons séquencés, cartographie du pourcentage de couverture MV / Cell vs Cell / Cell de distribution des gènes à une longueur ≥1000nt.

Cellule. Cela signifie que les ARNm dans les GFP MVs et GFP lysat cellulaire ont une expression disparate entre eux.

Enfin, pour finaliser la preuve de principe de notre nouvelle méthodologie TRACE-seq, il a été décidé d'induire un stress H2O2 a la lignée cellulaires HEK293T exprimant la

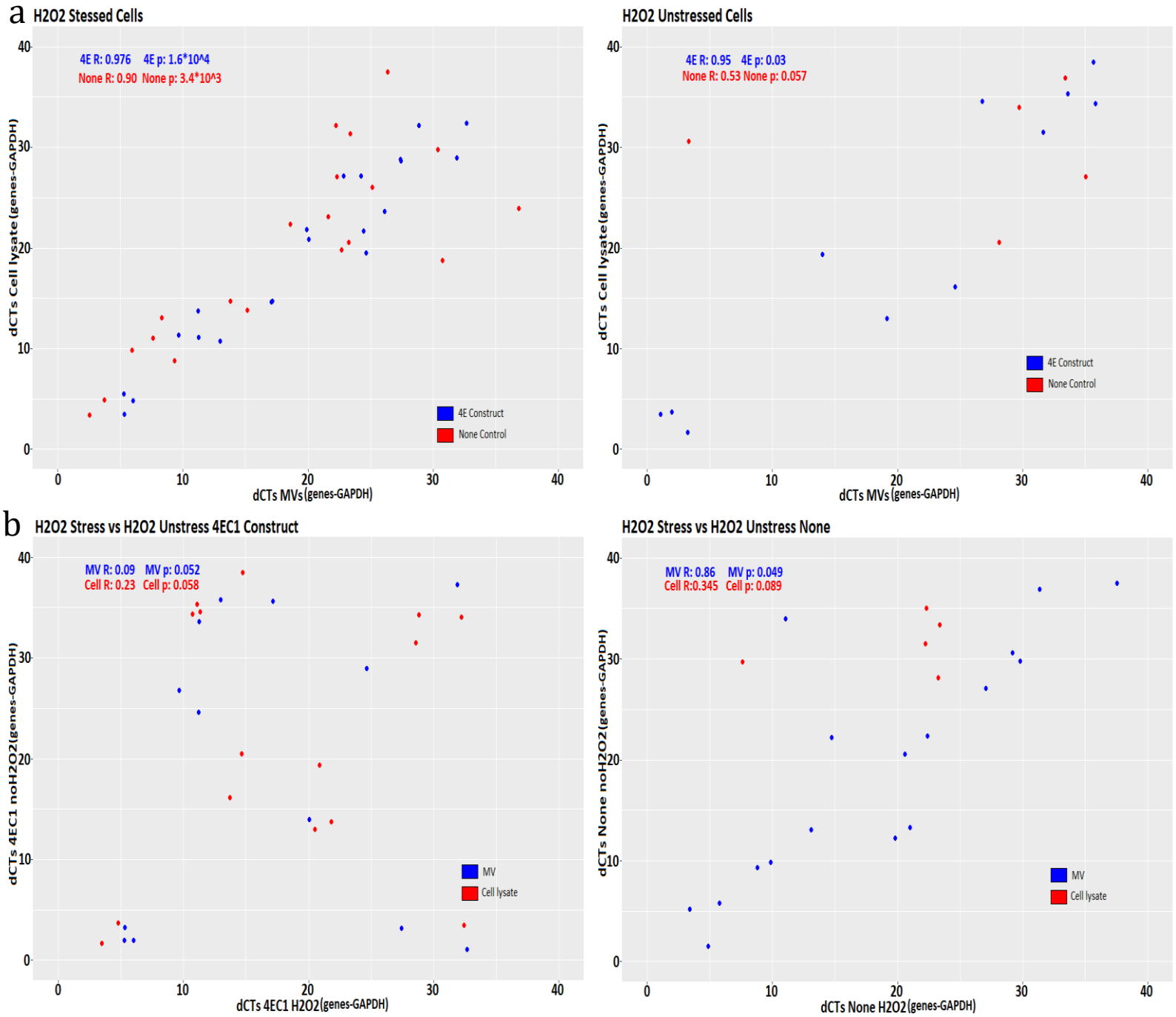


Figure.7 Test de validation de la voie de stress H2O2. Diagramme de corrélation obtenu à partir des résultats de RT-qPCR générés à partir de l'ARNm de MVs isolé à partir de cellules témoins (None), lignée cellulaire de construction TRACE (eGFP-EIF4E-C-YTHDF1 / GBP1-CD9) sur 8 gènes correspondant à la voie du stress oxydatif. **a** Lot de cellules stressées H2O2 et Lot de cellules non stressées corrélation Lysats cellulaires vs MV. **b**. Diagramme de corrélation H2O2 vs nonH2O2, lot correspondant à la construction 4EC1 et aux cellules de contrôle (None). Résultats obtenus à partir de populations cellulaires en tripliquât, 15 millions de cellules par condition

construction 4EC1 et des cellules normales afin de surveiller leur transcriptome avec TRACE-seq (Fig 7).

Les 8 gènes associés à l'oxydation testés dans les MVs de la lignée cellulaire GFP-EIF4E-C-YTHDF1 / GBP1-CD9 (4EC1 TRACE-seq) sont plus corrélés à leur propre lysat cellulaire en termes d'expression. En effet, par rapport aux cellules HEK régulières en ce qui concerne les échantillons de cellules traitées ou non traitées avec H2O2 (Fig 7 a) un meilleur indice de corrélation est détecté pour la lignée 4EC1 (respectivement la lignée 4EC1 ; H2O2 R : 0,976, no H2O2 R : 0,95 et HEK Normale ; H2O2 R : 0,90, no H2O2 R : 0,53). De plus, si l'on regarde l'expression de corrélation (Fig. 7 b) «avec vs sans» l'effet du traitement H2O2, la variation de l'expression génique due à l'induction du stress est clairement détectée et cohérente dans les lysats cellulaires et les MV de TRACE-seq par rapport au groupe témoin dans lequel les MV de contrôle ne reflètent pas cette variation de l'expression observée dans les cellules (respectivement 4EC1 ; MV R : 0,09, lysat cell R : 0,23 et HEK Normale ; MV R : 0,86, lysat Cell R : 0,345). Cela démontre que les changements dynamiques dans les profils d'ARNm des cellules sont bien mieux reflétés dans les MV pour la construction MV TRACE par rapport aux MV de contrôle (Fig 7 b). Ces résultats montrent que la technologie TRACE-seq peut être utilisée avec succès pour monitorer la voie d'expression des gènes du stress oxydatif au fil du temps avec une technique très simple, rapide et reproductible telle que RT-qPCR.

Ainsi, j'ai commencé à travailler sur la technologie Chic-seq (Transposase Tn5 - Anticorps) qui était une source d'investigation très prometteuse sur le rôle des FTs lors de la différenciation myocardique. Bien que celle-ci de soit pas finalisé en raison de la découverte d'une technologie concurrente complètement aboutie, le processus global de

développement de cette technologie a été très formateur et a débouché sur à une technique viable qui a répondu aux objectifs.

Les choix faits pour la conception de notre deuxième méthodologie de séquençage (TRACE-seq) ont été bons. En effet, seulement 2,5 ans de travail ont été nécessaires pour achever ce projet et valider une toute nouvelle technologie dans le domaine du séquençage. C'est également une méthodologie complètement nouvelle qui offre une approche totalement innovante de l'analyse du transcriptome dans une cellule vivante au fil du temps.

Par ailleurs, un point clef de développement pour ce projet fut les progrès réalisés par les dernières technologies de séquençage unicellulaire. En effet, c'est grâce à ces progrès qu'il a été possible de développer TRACE-seq. Notre matériel de base d'ARNm importé est si faible - quelle que soit la technique de purification - qu'une étape de pré-amplification est nécessaire. La génération d'ADNc est donc une étape critique qui grâce au protocole SMART-Seq 2 nous donne une relative certitude de générer un ARNm de pleine longueur à l'aide des oligos LNA contenant le site TSO (template switch oligos).

Même si certains groupes ont développé des approches similaires, un ARNm méthylé de pleine longueur m^6A importé par vésicules pour un suivi de cellule vivante qui pourrait être adapté *in vivo* est une première et un outil très puissant dans le suivi de tumeurs par exemple.

Université de Paris

LISTE DES ÉLÉMENTS SOUS DROITS

Liste de **tous les éléments retirés** de la version complète de la thèse
faute d'en détenir les droits

Document à intégrer dans la version partielle de la thèse

Illustrations, figures, images...

Légende de l'image	N° de l'image	Page(s) dans la thèse
How to sequence DNA.	Figure 5	40
Chromatin condensation - euchromatin and heterochromatin	Figure 9	48
EZ:Tn5®mechanism of action	Figure 14	67
Nextera®sequencing kit mechanism of action	Figure 15	68
Neb IMPACT®technology summary	Figure 19	84
Maleimide Diels-Alder reaction on a Thiols group	Figure 30	97
Transposase sedimented to beads or well wall.	Figure 37	106
Tam-ChiP® sequencing technology develloped by Active Motif™	Figure 46	116
Tam-ChiP® flow chart	Figure 47	117

Articles, chapitres, entretiens cliniques…

Titre du document	N° (si numéroté)	Page(s) dans la thèse