Ph.D. degree in Systems Medicine

Curriculum in Human Genetics

European School of Molecular Medicine (SEMM)

University of Milan and University of Naples "Federico II"

Disciplinary sector: BIO/11

DISSECTING THE REGULATORY LOGIC OF CELL FATE REPROGRAMMING THROUGH A MULTI-OMIC APPROACH

Francesco Panariello

Telethon Institute of Genetics and Medicine (TIGEM)

University ID No.: R12777

Supervisor: Prof. Davide Cacchiarelli

Internal advisor: Prof. Diego Di Bernardo

External advisor: Prof. Michael J. Ziller

Internal examiner: Prof. Sandro Banfi

External examiner: Prof. Matteo Cereda

PhD coordinator: Prof. Saverio Minucci

Academic Year 2022-2023

TABLE OF CONTENTS

ABSTRACT **4**

FIGURES INDEX **5**

1 INTRODUCTION **10**

1.1 ROAD TO CELL REPROGRAMMING 10

1.1.1 Stemness *10*

1.1.2 Cell programming *12*

1.1.3 The role of transcription factors in lineage switching *13*

1.1.4 Cell reprogramming *14*

1.2 NEXT GENERATION SEQUENCING 20

1.2.1 History of DNA sequencing *20*

1.2.2 NGS workflow and state of the art *22*

1.2.3 Single-cell RNA sequencing *28*

2 MATERIALS AND METHODS **32**

2.1 MICROFLUIDIC DEVICE 32

2.2 CELL CULTURE 32

2.3 REPROGRAMMING IN MICROFLUIDICS 32

2.4 REPROGRAMMING OF HUMAN FIBROBLASTS TO HIPSC COLONIES 32

2.5 SAMPLE PREPARATION FOR LC-MS/MS 33

2.6 MASS SPECTROMETRY ANALYSIS 34

2.7 PROTEOMIC BIOINFORMATIC ANALYSIS 34

2.8 SAMPLE PREPARATION FOR SINGLE-CELL RNA-SEQ 35

2.9 SINGLE-CELL RNA-SEQ DATA PRE-PROCESSING 35

2.10 SINGLE-CELL RNA-SEQ DATA VISUALIZATION AND CLUSTERING 36

2.11 SINGLE-CELL RNA-SEQ DIFFERENTIAL GENE EXPRESSION AND GENE SETS ENRICHMENT 36

2.12 SINGLE-CELL RNA-SEQ TRAJECTORY INFERENCE 37

2.13 SINGLE-CELL RNA-SEQ INTERACTION ANALYSES 38

2.14 STAT3 TARGETS EXPRESSION 38

2.15 BULK RNA-SEQ ANALYSIS OF REPROGRAMMING DATA 38

2.16 IMMUNOFLUORESCENCE STAINING 39

2.17 ASSESSMENT OF REPROGRAMMING EFFICIENCY 39

2.18 SECONDARY REPROGRAMMING EXPERIMENTS 39

2.19 SAMPLE PREPARATION FOR MULTIOME ANALYSIS 40

2.20 MULTIOME DATA PRE-PROCESSING 40

2.21 MERGED SCRNA-SEQ DATA VISUALIZATION AND CLUSTERING 41

2.22 MULTIOME DATA PROCESSING 41

2.23 Statistics and reproducibility 41

3 RESULTS 42

3.1 Human cell reprogramming in microfluidic system 42

3.1.1 Experimental design 42

3.1.2 Quality assessment 44

3.2 Development of a temporal multi-omic approach 44

3.2.1 Quality controls 44

3.2.2 Transcriptional waves contribute to the secretion of specific proteins 47

3.3 A rich extracellular signaling environment is shaped during human cell reprogramming 48

3.3.1 Accumulation of embryonic extracellular matrix 48

3.3.2 Dynamics of extrinsic regulatory signals 49

3.4 Extracellular environment and cell heterogeneity are interconnected 52

3.4.1 Resolving cell population heterogeneity 52

3.4.2 SR clusters contribute to the establishment of a specific environment 53

3.5 Trajectory inference reveals different fates 55

3.5.1 An early reprogramming fate is characterized by a secretory phenotype 56

3.6 Reprogramming fates interact through different ligand-receptor pairs 57

3.6.1 HGF-MET 59

3.6.2 NRG1-ERBB3 60

3.7 HGF-MET crosstalk functionally sustains the acquisition of pluripotency through STAT3 62

3.7.1 STAT3 pathway is active in reprogramming cells 63

3.7.2 Perturbation of STAT3 pathway components affect the efficiency of reprogramming 64

3.8 Future perspectives 66

3.8.1 Multiome approach to dissect the regulatory logic behind different fates 66

4 DISCUSSION 70

4.1 Identification and characterization of an unproductive cell fate during reprogramming 70

4.2 Role of the somatic fate in subpopulation crosstalk 71

4.3 An early embryonic stage is recapitulated during reprogramming 72

5 CONCLUSION 73

6 REFERENCES 74

ABSTRACT

Human cellular reprogramming to induced pluripotency is still an inefficient process, which has hindered studying the role of critical intermediate stages. Here we take advantage of high efficiency reprogramming in microfluidics and temporal multi-omics to identify and resolve distinct sub-populations and their interactions. We perform secretome analysis and single-cell transcriptomics to show functional extrinsic pathways of protein communication between reprogramming sub-populations and the re-shaping of a permissive extracellular environment. We pinpoint the HGF/MET/STAT3 axis as a potent enhancer of reprogramming, which acts via HGF accumulation within the confined system of microfluidics, and in conventional dishes needs to be supplied exogenously to enhance efficiency. Our data suggests that human cellular reprogramming is a transcription factor-driven process that is deeply dependent on extracellular context and cell population determinants. To further investigate the relationship between the fates that characterize this process, we aim to implement the results produced herein with chromatin accessibility data from the same cells. We hypothesize that different epigenetic states of cells at early time-points may sustain or repress the ability of cells to undergo one trajectory or the other.

FIGURES INDEX

FIGURE 1: THE STEM CELLS BIOLOGY. UPON THE FUSION OF AN EGG AND A SPERM, A TOTIPOTENT ZYGOTE EMERGES, CAPABLE OF FORMING BOTH THE INNER CELL MASS (ICM) AND THE EXTRA-EMBRYONIC (EE) TISSUE WITHIN THE BLASTOCYST. WHEN EXTRACTED FROM THE BLASTOCYST IN VITRO, ICM CELLS CAN BE SUSTAINED IN CULTURE, EVOLVING INTO PLURIPOTENT EMBRYONIC STEM CELL (ESC) LINES. AS THE EMBRYO DEVELOPS, THESE PLURIPOTENT STEM CELLS WITHIN THE ICM GRADUALLY LOSE THEIR POTENCY AND TRANSFORM INTO TISSUE-SPECIFIC, MULTIPOTENT STEM CELLS, MARKING A PROGRESSION TOWARDS INCREASED LINEAGE SPECIALIZATION. ADAPTED FROM ECKFELDT ET AL., 2005[185]. 12

FIGURE 2: EXAMPLES OF TRANSCRIPTION FACTORS-INDUCED DIRECT REPROGRAMMING. ADAPTED FROM GRAF., 2011[13]. 14

FIGURE 3: HIPSCS APPLICATIONS. HIPS CELLS CAN BE DERIVED FROM SOMATIC CELLS OF THE TISSUE REPERTOIRE OF A PATIENT. AFTER SPECIFIC TREATMENTS IN VITRO, HIPSCS CAN BE INDUCED TO FORM CELLS FORM SPECIALIZED CELLS THAT HAVE SEVERAL APPLICATIONS, SUCH AS DISEASE MODELING, DRUG SCREENING AND TESTING OF CELLULAR TOXICITY RESPONSE. ADAPTED FROM BELLIN ET AL., 2012[44] 16

FIGURE 4: ILLUMINA NGS CHEMISTRY OVERVIEW. EXPERIMENTAL WORKFLOW OF ILLUMINA TECHNOLOGY[85]. 24

FIGURE 5: **LIBRARY DEMULTIPLEXING OVERVIEW**. DNA (OR CDNA) FRAGMENTS ARE SEQUENCED IN POOL (LEFT) AND THEN ASSIGNED TO THE CORRECT SAMPLE USING THEIR SPECIFIC INDEX (RIGHT). ADAPTED FROM ILLUMINA[85]. 26

FIGURE 6: NGS COMPUTATIONAL WORKFLOW. THE STEPS THAT CHARACTERIZE A TYPICAL NGS COMPUTATIONAL WORKFLOW ARE DIVIDED IN PRIMARY (GREEN), SECONDARY (PURPLE) AND TERTIARY (RED) ANALYSES. COMMON SOFTWARE ARE REPORTED IN GREY. 27

FIGURE 7: NGS STATE OF THE ART. NUMBER OF PUBLICATION ON PUBMED WITH THE PATTERN "NEXT GENERATION SEQUENCING" PER YEAR. 28

FIGURE 8: GEMS GENERATION. THE CHROMIUM PLATFORM USES OIL TO CREATE MICRO-CHAMBERS WHERE REACTIONS CAN OCCUR. IN EACH CHAMBER (GEM), A BEAD COVERED WITH PRIMERS, CELLS AND ENZYMES ARE STORED. ADAPTED FROM 10X GENOMICS[186]. 29

FIGURE 9: MULTIOME WORKFLOW. TRANSPOSED NUCLEI AND ENZYME ENTER THE GEM WITH BEADS COVERED IN 10X PRIMERS. AFTERWARDS, DIFFERENT ADAPTERS ARE LIGATED TO ACCESSIBLE DNA FRAGMENTS AND RETRO-TRANSCRIBED RNA. ADAPTED FROM 10X GENOMICS[123]. 31

FIGURE 10: COMPARISON OF EFFICIENCY BETWEEN REPROGRAMMING SYSTEMS. LEFT: SCHEMATICS OF THE IN-SCALE CONVENTIONAL (WELL) AND MICROFLUIDIC (μF) SETUP. RIGHT: COMPARISON OF REPROGRAMMING EFFICIENCY BETWEEN THE TWO SYSTEMS. TWO-SIDED WILCOXON'S TEST WAS USED TO ASSESS DIFFERENCES AMONG THE CONDITIONS. N = 8 FOR WELL AND N = 15 FOR μF (***P = 0.0001593). 42

FIGURE 11: EXPERIMENTAL DESIGN. SCRNA-SEQ DATA WERE COLLECTED BY STOPPING PARALLEL EXPERIMENTS AT DAY 0, 3 AND EVERY 48 H. PROTEOMIC DATA WERE OBTAINED BY TANDEM MASS SPECTROMETRY ANALYSIS OF CONDITIONED MEDIA ALONG THE SAME REPROGRAMMING EXPERIMENTS. THE DIFFERENCES IN MEDIUM USAGE AND REPROGRAMMING FACTORS SUPPLIANCE (RNA TRANSFECTIONS) ARE REPORTED. 43

FIGURE 12: REPROGRAMMING EFFICIENCY OF THE PROTOCOL USED FOR SECRETOME DATA. EFFICIENCY EVALUATED WITHIN THE SAME MICROFLUIDIC CHANNELS USED FOR PROTEOMIC ANALYSES (N=40 FOR REPL.1, N=35 FOR REPL.2 AND N=39 FOR REPL.3). DATA ARE PRESENTED AS MEAN VALUES +/- SD. 43

FIGURE 13: MORPHOLOGICAL EVALUATION. TOP: MORPHOLOGICAL CHANGES OCCURRING DURING REPROGRAMMING, SAMPLED AT DAY 0 (D0), DAY 5 (D5), DAY 9 (D9) AND DAY 13 (D13). BOTTOM: IMMUNOSTAINING OF A SINGLE MICROFLUIDIC CHANNEL AT DAY 14 FOR PLURIPOTENCY MARKERS (NANOG AND TRA-1-60). SCALE BAR: 100NM. 44

Figure 14: Secretome data consistency. Left: Visualization of proteomic data correlation between replicates. Each dot represents an identified protein. Log_2 relative quantification is shown on the axes (a.u.). Right: PCA plot proteomic data, shown as the distribution of the proteomic pattern for sampled conditioned medium over a 48-hour period.....45

Figure 15: Sequencing data quality control and filtering. Left: Scatter plot representing the number of reads (x axis) over the number of detected genes (y axis) for each cell. Color gradient shows the percentage of reads associated with mitochondrial genes. The dotted line has been put at 1,000 detected genes, used for filtering. Right: Schematic representation of cells/genes filtering from raw data to the final dataset. ..... **Error! Bookmark not defined.**

Figure 16: scRNA-seq data consistency. Left: Heatmaps of Pearson correlation coefficient for each replicate, divided by each time-point. Correlation has been evaluated by comparing the distribution of each replicate in the clusters identified in Figure 21. Right: FLE plot for scRNA-seq data. Sequencing data is shown as the distribution of transcriptional patterns for single cells across sampled time-points.. **Error! Bookmark not defined.**

Figure 17: Comparative dynamics of features. Absolute number of differential features for each -omics data, both up- (Up - red) and down-regulated (Down - blue). Each value refers to the differential analysis between subsequent time-points. Peaks of deregulation are highlighted (arrows). 47

Figure 18: Definition of a relationship between transcription and secreted proteins. Median z-score of the 155 proteins found up-regulated from day 5 (D5) to day 7 (D7) in proteomic data. Trends have been evaluated along the time-course for both -omics data. 48

Figure 19: Accumulation of embryonic extracellular matrix. Top: ECM-related results from the enrichment analysis within the Reactome database of the 555 proteins identified as secreted. Edges connecting different categories reproduce Reactome hierarchy relationships. Bottom: Hierarchical clustering of proteins identified in this study and belonging to the core ECM components[137] at specific stages of embryo development[136]. 50

Figure 20: Dynamics of extrinsic regulatory signals. Top: Signaling-related results from the enrichment analysis within the Reactome database of the 555 proteins identified as secreted. Edges connecting different categories reproduce Reactome hierarchy relationships. Bottom: Hierarchical clustering of proteins identified in this study and belonging to signaling pathways.....51

Figure 21: Cell clustering and annotation. Left and middle: Somatic and Developmental signatures enrichment scores shown along the FLE map. Right: FLE map showing the distribution of cells across identified clusters. 52

Figure 22: Characterization of cell heterogeneity. Left: Time-points and cell-cycle phase distribution for each cluster. Right: Heatmap of Z-scored normalized counts, averaged by clusters, for key reprogramming related genes (right). NA cluster not shown. 53

Figure 23: Contribution of SR clusters to the secretome. Heatmaps of highly dynamic proteins from the secretome analysis. The colors display log2 fold change protein concentration with respect to D1-D2 (Secretome - left) and Z-scored log2 counts per million (scRNA-seq - right). Hierarchical clustering was performed on scRNA-seq data according to each separate cluster of cells. Proteins involved in primitive node formation are highlighted (red names).....54

Figure 24: SR clusters profile is enriched by secreted pathways. GSEA results for each cluster. Only significant results are shown. The gene set made of the secreted proteins found in this work is written in bold. NES, Normalized enrichment score. 55

Figure 25: Trajectory inference analysis. Monocle3 (black line) and WOT (colored dots) trajectory inferences are displayed on the FLE graph. Arrows point to the starting point (blue) and 4 end points (red) of the inferred trajectories. A representative scheme of the trajectories is shown on the top-right.....56

Figure 26: SR2 cluster shows a secretory phenotype. Left: GSEA has been performed on SR2 cluster using signaling-related genesets used in Fig. 17. The results are shown as a barplot, displaying FDR (x axis) and NES (colors). Right: Enrichment Score graph relative to the GSEA of SR2 cluster for senescence-associated secreted proteins geneset (SASP). Black lines on the x axis represent a match between the ranked list and the geneset analyzed. NES, Normalized enrichment score. FDR, False Discovery Rate. **Error! Bookmark not defined.**

Figure 27: SASP genes are specific to SR2 cluster. SR2 cluster marker genes relative expression, shown in a heatmap of Z-scored normalized counts, averaged by clusters. Genes with (*) have been detected in secretome analysis. ..... **Error! Bookmark not defined.**

Figure 28: Interaction Score analysis. Left: Schematic representation of ligand-receptor interactions hypothesized during reprogramming. Fibroblasts (D0, left) develop two fates: a somatic secretory phenotype (bottom) and induced pluripotency (top). Black arrows show the directionality of the examined interaction. Right: Schematic representation of ligand-receptor pairs selection for interaction score analyses, as described in Methods. 58

Figure 29: Interaction Score results. Heatmap of z-scored standardized interaction scores for all the ligand-receptor pairs analyzed. 58

Figure 30: HGF-MET dynamics in microfluidic system. HGF and MET gene expression profiles (log_2 CPM) displayed on the FLE map as fold change relative to HGF and averaged across the time course (bottom-left). 59

Figure 31: HGF-MET dynamics in other systems. HGF and MET gene expression profiles shown in Liu et al., 2020[170], as averaged across their identified clusters (top) and in Cacchiarelli et al., 2015[64], shown as mouse and human mean normalized expression at sampling day 8 (bottom-left; ** BH-adjusted p-value <0.01). Bottom-right: Representative pictures of HiF-T DOX secondary reprogramming performed with or without depletion of MEFs in standard 12-well plates, assessed by immunostaining of TRA-1-60. 60

Figure 32: NGR1-ERBB3 dynamics in microfluidic system. NGR1 and ERBB3 gene expression profiles (log_2 CPM) displayed on the FLE map as fold change relative to NRG1 and averaged across the time course (bottom-left). 61

Figure 33: NRG1-ERBB3 dynamics in other systems. NRG1 and ERBB3 gene expression profiles shown in Liu et al., 2020[170], as averaged across their identified clusters (top) and in Cacchiarelli et al., 2015[64], shown as mean FPKM across the time-course (bottom). 61

Figure 34: HGF/c-MET/STAT3 axis. A schematic representation of HGF/c-MET/STAT3 signaling pathway (Created with BioRender.com). 62

Figure 35: STAT3 nuclear activation. Top: In the FLE graph, green dots represent cells with positive enrichment scores for STAT3 target genes (Methods). Bigger circles summarize averaged HGF (left) and MET (right)

gene expression in identified clusters. Significant inter-cluster HGF-MET interactions are displayed (arrows). Arrow thickness relates to the strength of the interaction. Bottom: Time course of STAT3 activation during the microfluidic reprogramming process. Representative images showing a first wave around day 4 to day 7, and then a second wave at the end of the process, when it is active just in the hiPSC colonies. 63

Figure 36: MET co-localizes with nuclear STAT3. Top: Representative images of expression of nuclear STAT3 and c-MET during reprogramming performed in microfluidics at day 6. Bottom: Correlation between the expression intensity of nuclear STAT3, c-MET, and cell size obtained from experimental data shown on top. Data from n = 61 cells (n = 3 independent experiments).....64

Figure 37: STAT3 inhibition impairs reprogramming. Left: Reprogramming efficiency in microfluidics measured as the relative area occupied by NANOG+ colonies in cells upon inhibition of c-Met and JAK1 kinases using small molecules at day 12, compared to the ones treated with the vehicle (n = 6 for vehicle, n = 12 for JAKi and n = 7 for c-METi); ANOVA followed by two-sided Dunnett's multiple comparisons test was used to assess differences among the conditions (JAKi – *** FDR = 0.0001; cMETi – *** FDR = 0.0001). Representative quantification pictures in microfluidic channels assessed by immunostaining of NANOG are shown. Right: Reprogramming efficiency in microfluidics upon knock-down of STAT3 using siRNAs at day 12 (n = 8 for scramble siRNA, n = 11 for siSTAT3); two-sided unpaired t-test was used to assess differences among the conditions (***P < 0.0001). Representative quantification pictures in microfluidic channels assessed by immunostaining of NANOG are shown.....65

Figure 38: Exogenous signals improve reprogramming. Left: Reprogramming efficiency in standard 24-well plates upon addition of HGF, IL-6 and soluble IL6 receptor (sIL6R), or NRG1 at day 9 (n = 14 for control, n = 19 for HGF, n = 5 for IL6 + sIL6R, n = 16 for NRG1); ANOVA followed by two-sided Dunnett's multiple comparisons test was used to assess differences among the conditions (HGF - *** FDR = 0.0001; IL6 + sILR – **FDR = 0.0083; NRG1 – ** FDR = 0.0021). Representative quantification pictures in standard 24-well plates assessed by immunostaining of NANOG and TRA-1-60 are shown. Right: Reprogramming efficiency in standard 24-well plates upon temporally modulate addition of HGF, IL6 and soluble IL6 receptor (sIL6R), and NRG1 at day 9 (n = 14 for control, n = 6 for HGF in the early phase and NRG1 in the late phase, n = 4 for HGF + IL6 + sIL6R in the early phase and NRG1 in the late phase, n = 4 for HGF + IL6 and sIL6R + NRG1 for the entire process); ANOVA followed by two-sided Dunnett's multiple comparisons test was used to assess differences among the conditions (E HGF + L NRG1 - ** FDR = 0.0038; E HGF/IL6 + sILR + L NRG1 – *** FDR = 0.0001; ALL – *** FDR = 0.0001). Representative quantification pictures in standard 24-well plates assessed by immunostaining of NANOG and TRA-1-60 are shown.....66

Figure 39: Multiome approach. Cells from the whole time-course are pooled together (Pooling) and sequenced for both mRNA and accessible DNA (Same cell Multiomics). The transcriptional profile of scRNA-seq data from this work is used to characterize the gene expression phenotype of pooled cells (Anchoring). 67

Figure 40: RPCA enables correct anchoring. UMAP visualization of scRNA-seq data from this work after re-analysis with (right) or without (left) pooled cells from Multiome approach (black dots).....68

Figure 41: Correlation of gene expression and chromatin accessibility at marker genes loci. Reads distribution from accessible DNA at SOX2 (top) and ELN (bottom) loci, divided by clusters. Violin plots of gene

expression (log_2 CPM) are reported on the right. Links are generated by correlating gene expression with chromatin accessibility at called peaks. 69

Figure 42: Primitive node formation and reprogramming. Left: Heatmap of Z-scored log2 counts per million, averaged by day, of genes encoding for primitive node components. Right: Schematic representation of primitive node formation (Adapted from Boccaccio and Comoglio, 2006) and primitive node components (Created with BioRender.com). Snapshot of early and late events are reported according to the expression dynamics on the left. Black arrows show the contribution of SR and DR cells based on their average gene expression. 72

1 Introduction

1.1 Road to Cell Reprogramming

1.1.1 Stemness

1.1.1.1 Stem cells potency

Stem cells, essential components of the human body, are undifferentiated cells capable of transforming into various cell types and renewing themselves. These cells exist in both embryos and adult tissues. The process of specialization occurs through several stages, each diminishing the developmental potential. As cells progress from pluripotent to unipotent states, their ability to differentiate into different cell types becomes progressively limited. Unipotent stem cells possess restricted differentiation capacity compared to pluripotent ones, showcasing the gradual reduction in developmental potency. Due to these features, they are becoming a very powerful tool in a plethora of field, including regenerative and precision medicine, toxicology testing, disease modeling, and so forth.

Their classification is based on the level of "potency", namely their potential to specialize in a multitude of different cell types:

- **Totipotent stem cells:** exemplified by the zygote formed after fertilization, they possess the highest differentiation potential, able to form all cells in an organism, including both embryo and extra-embryonic structures.
- **Pluripotent stem cells (PSCs):** examples are embryonic stem cells (derived from preimplantation embryos) and induced pluripotent stem cells (generated from somatic cells through artificial means). Pluripotent stem cells can form cells of all germ layers but not extraembryonic structures.
- **Multipotent stem cells**: like hematopoietic stem cells, they have a narrower differentiation spectrum, specializing in specific cell lineages. These cells can differentiate into several types within their lineage. After a few divisions, multipotent cells become **oligopotent**, capable of differentiating into a limited number of related cell types within their lineage.
- **Unipotent stem cells**: they are the most specialized, with the ability to differentiate into only one cell type, yet they possess a unique property of repeated division. For instance, myeloid stem cells can give rise to white blood cells but not red blood cells. Unipotent stem cells, due to their restricted differentiation and robust division capabilities, hold significant promise in therapeutic applications within regenerative medicine.

1.1.1.2 Stem cells biology

Stem cells with the greatest differentiation potential reside in the early stages of embryo development. Indeed, they must guarantee the whole organism's development. Their appearance starts from the zygote, after the fusion of a spermatocyte with an oocyte (fertilization)[1]. After some divisions, a blastocyst is formed, housing short-lived embryonic stem cells along its inner wall. These embryonic stem cells, or ES cells, are pluripotent, possessing the capacity to develop into any cell type in the organism. The blastocyst comprises two distinct cell types: the inner cell mass (ICM), which gives rise to the epiblasts, initiating the development of the fetus, and the trophectoderm (TE), responsible for forming extraembryonic support structures, including the placenta. While TE specializes, ICM cells maintain their undifferentiated, fully pluripotent, and proliferative state[2]. Human embryonic stem cells (hESCs) are derived from the ICM[3,4].

During embryogenesis, cells organize into germ layers - endoderm, mesoderm, and ectoderm - each ultimately leading to differentiated cells and tissues in the fetus and later in the adult organism[5]. Upon differentiating into one of these germ layers, hESCs become multipotent stem cells, with their potential limited to cells within the specific germ layer. This differentiation process is quick in human development. Subsequently, pluripotent stem cells are distributed throughout the organism as undifferentiated cells, capable of both proliferating by generating the next generation of stem cells and differentiating into specialized cells under particular physiological conditions.

The specialization of stem cells is influenced by signals, categorized as external (such as physical cell contact or chemical secretions from surrounding tissues) and internal (signals regulated by genes within DNA). Stem cells also serve as internal repair systems within the body. They continuously replenish and generate new cells, a process that remains unlimited as long as the organism is alive. The activity of stem cells varies depending on the organ they are located in; for example, in bone marrow, they undergo constant division, whereas

in organs like the pancreas, division occurs only under specific physiological conditions (Figure 1).

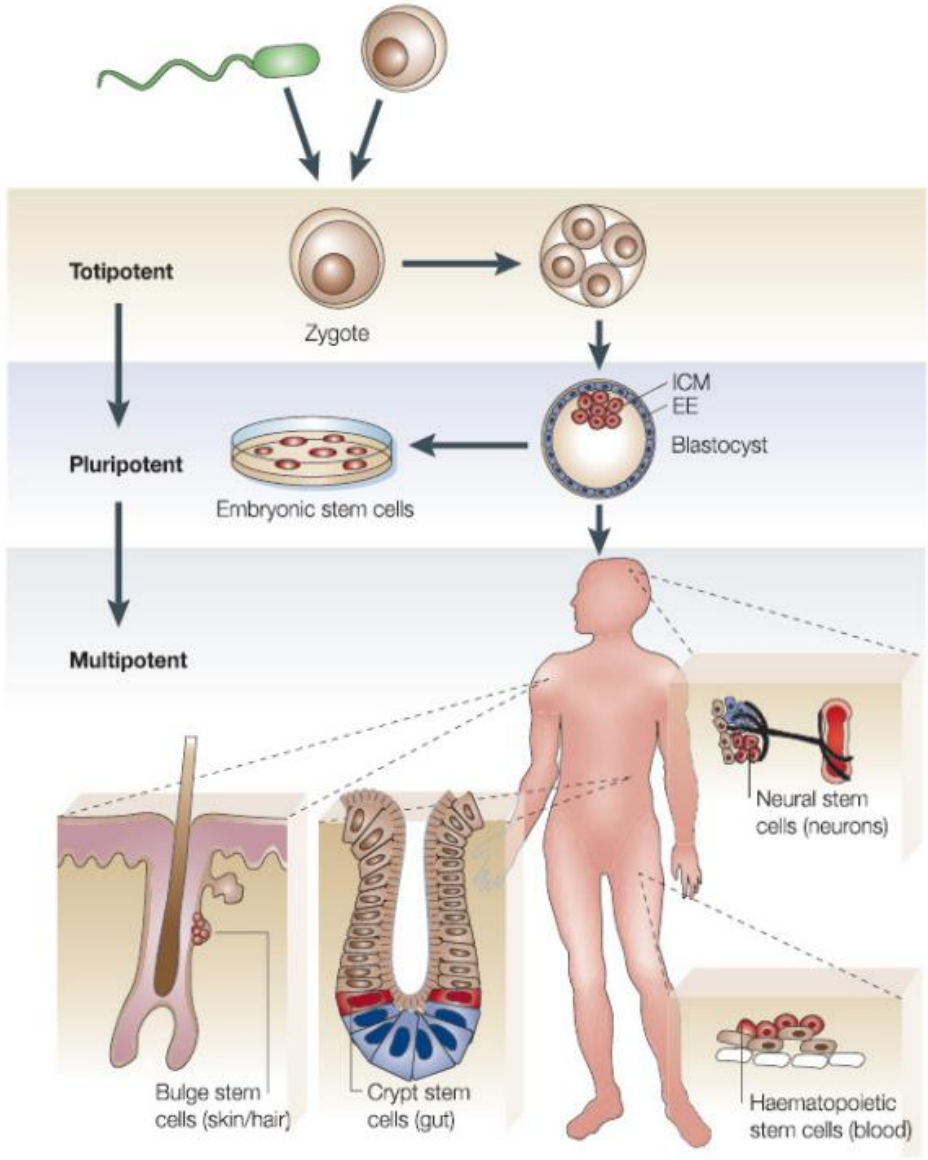

Figure 1: The stem cells biology. Upon the fusion of an egg and a sperm, a totipotent zygote emerges, capable of forming both the inner cell mass (ICM) and the extra-embryonic (EE) tissue within the blastocyst. When extracted from the blastocyst in vitro, ICM cells can be sustained in culture, evolving into pluripotent embryonic stem cell (ESC) lines. As the embryo develops, these pluripotent stem cells within the ICM gradually lose their potency and transform into tissue-specific, multipotent stem cells, marking a progression towards increased lineage specialization. Adapted from Eckfeldt et al., 2005[185].

1.1.2 Cell programming

For an extended period, the process of differentiation was perceived as a unidirectional journey, where cell states followed a predetermined path within the 'epigenetic landscape' conceptualized by Conrad Waddington in 1957.

In subsequent classic studies, the notion that 'committed' cells irreversibly follow a predetermined path was challenged. These studies indicated that committed cells, while retaining their genetic information, can alter their fate in response to specific stimuli. For instance, experiments with Drosophila melanogaster pupae's imaginal disc cells showcased 'transdetermination', i.e., cells destined for specific structures transformed into different tissues upon transplantation, displaying remarkable plasticity[6,7]. Although these fate switches occurred infrequently, they demonstrated the surprising adaptability of explanted cells.

Another significant study involved transplanting cells between quails and chickens. Despite their resemblance, these cells had distinct nuclei, allowing tracking of their fate. The research by Le Lievre and Le Douarin[8] revealed that neural crest cells, when transplanted, adopted new fates influenced by their new cellular surroundings, generating diverse tissues such as bone, cartilage, and connective tissue in the avian embryo.

Parallel experiments across various species, including both embryonic and adult somatic cells transferred into enucleated oocytes, led to the formation of all three germ layers and even the development of entire organisms. These experiments, conducted by researchers like Gurdon, Byrne, Melton, Hochedlinger, and Jaenisch, provided unequivocal evidence that the identity of differentiated cells can be completely reversed[9–12]. These findings challenged the traditional belief in the irreversibility of cell differentiation.

1.1.3 The role of transcription factors in lineage switching

A significant principle contributing to the discovery of induced pluripotency involved the observation that lineage-associated transcription factors, which regulate cellular identity during development by activating specific genes and suppressing inappropriate ones, can alter cell fate when expressed in different cell types[13,14]. In 1986, Lassar and colleagues demonstrated 'transdetermination' by converting fibroblasts into myoblast-like cells using the demethylating agent 5-azacytidine and the expression of specific cDNAs[15]. Subsequent research identified MyoD as the key transcription factor driving this conversion[16], illustrating how transcription factors could reprogram differentiated somatic cells into different phenotypes.

Further studies in the hematopoietic system, a well-defined mammalian cellular differentiation system, provided fundamental insights. Experiments revealed that forced expression of GATA1 induced erythroid and megakaryocytic markers in monocytic cell lines[17], demonstrating that transcription factors not only activate new gene programs but also suppress existing ones, a characteristic of transdifferentiation. Additionally, it was found that

the expression of PU.1 in megakaryocytic and erythrocyte precursors converted these cells into myeloblasts[18]. Later, primary B and T cells were efficiently converted into functional macrophages by overexpressing the myeloid transcription factor C/EBPα[19,20], demonstrating the remarkable plasticity of mature hematopoietic cells. In vivo experiments in mice illustrated the conversion of exocrine cells into insulin-producing cells by overexpressing specific transcription factors, alleviating diabetes symptoms[21]. Additionally, the discovery that specific combinations of transcription factors, such as Gata4, Mef2c, and Tbx5, could

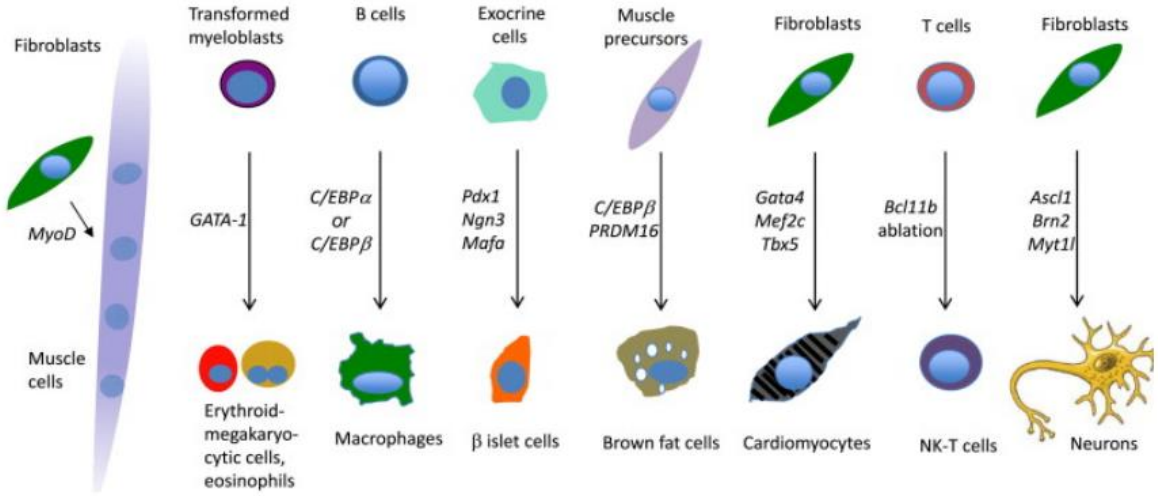

Figure 2: Examples of transcription factors-induced direct reprogramming. Adapted from Graf., 2011[13].

convert fibroblasts into cardiomyocytes further expanded the possibilities of lineage conversion[22,23]. Neural factors like Ascl1, Brn2, and Myt1l were shown to transform fibroblasts into induced neuron-like cells[24] (Figure 2).

Remarkably, these experiments demonstrated the feasibility of converting cells not only within the same tissue or germ layer but also between different tissues and embryonic origin. These pioneering direct programming experiments laid the groundwork for the systematic exploration of transcription factors capable of inducing the conversion of differentiated cells into a pluripotent state.

1.1.4 Cell reprogramming

Prior to achieving somatic cell reprogramming to pluripotency mediated by transcription factors, other approaches were utilized to achieve the same goal. These methods include nuclear transfer to eggs or oocytes, cell fusion and extract treatment. Nevertheless, from a cost-efficacy point of view, the obtainment of human induced pluripotent stem cells (hiPSCs) via the use of transcription factors remains the best option.

Yamanaka and Takahashi employed a meticulous screening approach to identify factors crucial for reprogramming somatic cells into pluripotent cells. They initially tested 24 candidate genes associated with pluripotency, focusing on the activation of the ES-specific gene Fbx15[25]. Through a systematic reduction strategy, they determined the minimal set of factors necessary for reprogramming fibroblasts into pluripotent-like cells. This minimal combination, consisting of Oct4, Sox2, Klf4, and Myc (OSKM), was defined as the Yamanaka factors. The cells generated using OSKM expressed pluripotency markers such as Nanog and SSEA-1 and formed teratomas when injected into immunocompromised mice[25].

However, these "first generation" induced pluripotent stem (iPS) cells were only partially reprogrammed; they expressed lower levels of key pluripotency genes compared to embryonic stem (ES) cells, displayed incomplete demethylation of the Oct4 promoter, and failed to generate live chimeras[25]. To address these limitations, subsequent efforts focused on more stringent criteria, using endogenous Nanog or Oct4 activation as markers for pluripotency. The resulting Oct4-iPS or Nanog-iPS cells, termed "second generation" iPS cells, were fully reprogrammed. They exhibited global gene expression and chromatin configuration identical to ES cells, complete demethylation of Nanog and Oct4 loci, reactivation of the X-chromosome in female lines, contribution to germline-competent chimeras, and correct expression of all well-known pluripotency markers[26–28]. Inducible lentiviral vectors demonstrated that the four factors needed to be expressed for at least 12 days to obtain iPS cells, and the frequency of reprogramming increased with time, with up to 0.5% of input mouse embryonic fibroblasts giving rise to iPS cells 3 to 4 weeks after infection[29].

Moreover, iPS cells capable of generating "all-iPS" mice upon injection into tetraploid blastocysts were developed[30,31]. These advancements were not limited to mice; iPS cells were derived from various species, including humans[32–34], rats[35], rhesus monkeys[36], and endangered species[37], showcasing the conservation of pluripotency transcriptional networks across evolution. Additionally, iPS cells were successfully generated from diverse somatic cell types, underscoring the universal applicability of induced pluripotency. Examples of differentiated cells that were reprogrammed include keratinocytes[38], hematopoietic cells[39], and so forth. Stem cells and certain progenitor cells were found to be more readily reprogrammed, likely due to their expression of specific pluripotent stem cell genes[40,41].

1.1.4.1 hiPSCs application in biology and medicine

Cell reprogramming has revolutionized regenerative medicine by demonstrating that the earliest stages of embryonic development can be regained by a process considered

irreversible until then. This discovery has revolutionized the fields of pharmaceuticals, clinics, and research laboratories. The ability to derive self-renewing hiPSCs from any patient provides an unprecedented platform for gaining in-depth understanding of various diseases, conducting in vitro drug screening, and exploring gene repair strategies alongside

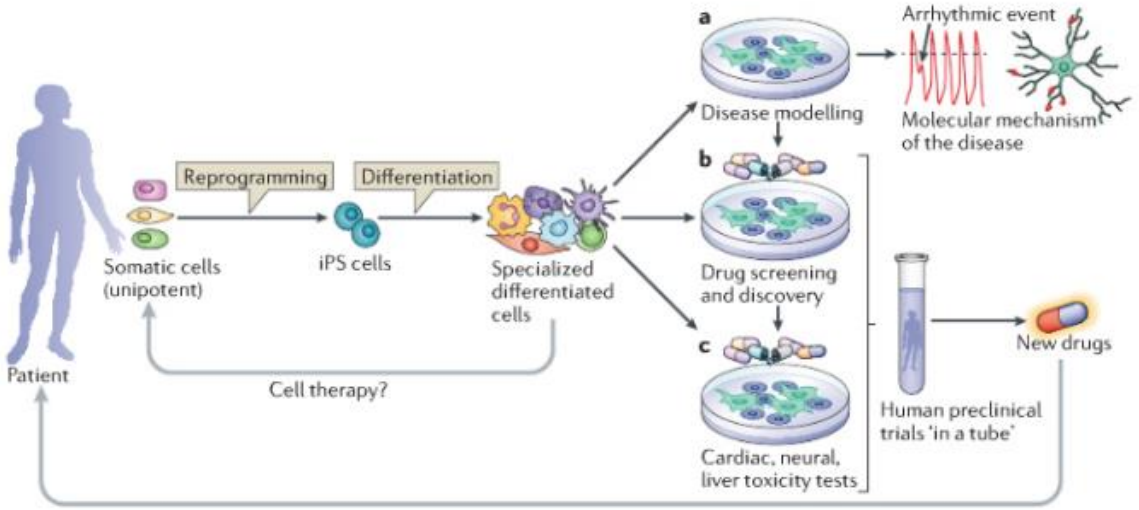

Figure 3: hiPSCs applications. hiPS cells can be derived from somatic cells of the tissue repertoire of a patient. After specific treatments in vitro, hiPSCs can be induced to form cells form specialized cells that have several applications, such as disease modeling, drug screening and testing of cellular toxicity response. Adapted from Bellin et al., 2012[44].

cell-replacement therapies[42–44]. This breakthrough offers limitless opportunities for advancing our knowledge of diseases, developing new therapies, and improving patient-specific treatments (Figure 3).

In the realm of disease modeling, hiPSCs technology has emerged as a powerful tool, particularly for understanding genetically inherited diseases affecting inaccessible tissues like neuropathologies. Numerous studies have showcased the effectiveness of human iPSCs in replicating genetic diseases within laboratory settings. Differentiated cells derived from patients' hiPSCs often mirror the disease traits observed in vivo, providing valuable insights into disease mechanisms.

For instance, hiPSCs derived from spinal muscular atrophy (SMA) patients, a condition characterized by the loss of motor neurons, displayed a progressive loss of motor neurons during in vitro differentiation, resembling the developmental motor neuron loss seen in SMA[45]. Similarly, cardiomyocytes derived from iPSCs of patients with LEOPARD syndrome, a disease associated with hypertrophic cardiomyopathy, exhibited enlargement, reflecting the hypertrophy observed in affected individuals[46].

In the case of diseases like Long QT syndrome and Timothy syndrome, conditions characterized by prolonged QT intervals on electrocardiography, hiPSCs derived from patients were differentiated into cardiomyocytes. These cardiomyocytes displayed prolonged action potentials in single-cell electrophysiological assays, replicating a key feature of these diseases[47,48].

Catecholaminergic Polymorphic Ventricular Tachycardia (CPVT) is a life-threatening condition in young patients characterized by an increased susceptibility to arrhythmias under catecholaminergic stress, despite having a structurally normal heart. In recent studies, researchers utilized cardiomyocytes derived from patient-specific iPSCs to investigate CPVT. These iPSC-derived cardiomyocytes from patients with both dominant and recessive forms of CPVT exhibited clear signs of the disease when compared to healthy controls[49–51]. Additionally, iPSC-derived cardiomyocytes from patients with mutations in SCN5A, a gene associated with cardiac disorders, displayed features of a cardiac 'overlap syndrome.' This syndrome involves the coexistence of Long QT syndrome (LQTS) and Brugada syndrome, highlighting the complexity of cardiac disorders that can be studied using hiPSC technology. Such faithful replication of disease traits in vitro enhances our understanding of these conditions and opens avenues for drug screening and therapeutic development.

This technology also opened new avenues in the search for treatments of complex non-monogenic disorders like schizophrenia and Alzheimer's disease. Studies have indicated that iPSCs can provide valuable insights into potential treatments, including novel antipsychotic drugs for schizophrenia[52] and β-secretase inhibitors for both familial and sporadic Alzheimer’s disease[53]. These findings demonstrate the feasibility of using iPSC-derived cells for predictive drug screening. Additionally, this technology offers a starting point for identifying effective drug dosages while minimizing side effects. Researchers can explore modifications to molecules to reduce toxicity while retaining their therapeutic properties, marking a significant advancement in drug discovery and development.

hiPSCs are an invaluable tool for modeling diseases *in vitro*; however, one of the aspects leading to development of patient-specific stem cells has also been the prospect of generating a ready supply of immune-compatible cells and tissues for autologous transplantation. Although the clinical translation of hiPSCs-based cell therapies seems more futuristic than the *in vitro* use of hiPS cells for research and drug development, a proof of principle study has attempted to use homologous recombination to repair the genetic defect in hiPS cells derived from a humanized mouse model of sickle-cell anemia[54]. Directed differentiation of the repaired hiPS cells into hematopoietic progenitors followed by transplantation of these cells into the affected mice led to the partial rescue of the disease phenotype. The gene

corrected hiPS-cell-derived hematopoietic progenitors showed engraftment in the injected mice and an at least temporal correction of the disease phenotype. Importantly however, a *bona fide* hematopoietic stem cell with the capacity for long-term multilineage reconstitution has yet to be generated from pluripotent cells.

In another landmark study, Wernig and colleagues derived dopaminergic neurons from iPS cells that, when implanted into the brain became functionally integrated and improved the condition of a rat model of Parkinson's disease[55]. The successful implantation and functional recovery in this model is evidence of the therapeutic value of pluripotent stem cells for cell-replacement therapy in the brain, one of the most promising areas for future hiPS cell applications.

1.1.4.2 The molecular landscape of cell reprogramming

The acquisition of pluripotency has been highly dissected at cellular and molecular levels. At the' onset of reprogramming, the presence of MYC promotes rapid cell expansion and resistance to apoptosis; other than that, its role is considered marginal and sometimes even dispensable[56]. In contrast, OCT4, SOX2, and KLF4 allow the suppression of somatic mesenchymal genes and the acquisition of an epithelial phenotype, from which pluripotent clones subsequently emerge[32,33,57]. Many theories have followed on the pathways followed for acquiring pluripotency, postulating that this was a more or less stochastic[58–60] or deterministic[61,62] process. This is because until a few years ago, the focus had not been on the concept that even if reprogramming is not a natural process, it still needs to exploit the principles of development and organogenesis. One possibility, then, is that pluripotency factors gradually restore niches in embryonic development that allow the subsequent transition to a stage closer to pluripotency. The first evidence for this came from the identification in the terminal parts of reprogramming of a transient acquisition by cells transiting to pluripotency of a molecular signature comparable to that of primitive endoderm with the FOXH1 gene as the hallmark of this state[63]. In parallel, it was observed that in the middle stages of reprogramming, as soon as an epithelial phenotype is acquired, the reprogramming cells express genes and signatures of the broad mesoderm, including LEFTY2 and the HOX genes[64].

These transcriptional signatures are obviously accompanied by epigenetic modifications that in-primis activate paused genes that, from a developmental point of view, are closer to the somatic cell of origin and then activate repressed genes and DNA methylated regions[64].

If, therefore, one wants to identify the molecular hallmarks of reprogramming and its dynamics, one can identify:

1. Transient increase in chromatin accessibility and paused/bivalent genes at the expense of exclusively active or repressed regions.
2. Removal of somatic DNA methylation to reach the end of reprogramming in a hypomethylated state.
3. Acquisition of gene regulatory networks of pluripotency and expression of hallmark genes.

If, therefore, differentiation can be defined as the set of cellular choices that lead to the definition of a phenotype and the establishment of transcriptional and epigenetic barriers that irreversibly lead the cell to ultimate somatic development, then reprogramming can be defined as that process whereby forcibly these barriers are removed in the order opposite to which they were placed. Obviously, in implementing this, as cellular reprogramming is not an evolutionarily conserved process, it is possible to accumulate the expansion of reprogramming intermediates or alternative non-productive fates.

In the latter view, it is essential to distinguish what represents a reprogramming intermediate from what instead heads down an unproductive pathway but is sometimes beneficial to the process.

Whilst there is a body of literature describing reprogramming trajectory in mouse[65–67], the fine dynamics of human reprogramming intermediates, which constitute the bottleneck of the process, remain largely unexplored due to the complexity of recognizing and selecting rare phenotypes that will evolve into a hiPSC fate.

It has been recently suggested that reprogramming of murine cells may also depend on population dynamics through cell-non-autonomous mechanisms in a context-dependent manner, i.e. mediated by cell-secreted factors[66,68,69].

1.2 Next Generation Sequencing

Next-generation sequencing (NGS), also known as high-throughput sequencing, encompasses a variety of cutting-edge sequencing techniques that have transformed the fields of genomics and molecular biology in recent years. One of the primary advantages of NGS technologies is their ability to sequence genetic material (i.e., DNA and RNA) in a rapid and cost-effective manner, surpassing the capabilities of earlier sequencing methods. Furthermore, the data generated by NGS exhibit unprecedented quality, robustness, and minimal noise. It's worth noting that the success of an NGS project relies on expertise both in the wet lab for sample preparation and in bioinformatics for data analysis to ensure high-quality data and accurate interpretation. Currently, NGS technologies have become a staple for researchers tackling a wide array of biological questions. The unprecedented scale and efficiency of sequencing achievable today have propelled advancements across diverse areas, from genome analysis to the study of how proteins interact with nucleic acids.

1.2.1 History of DNA sequencing

The fundamental information about the hereditary and biological characteristics of living organisms resides within the arrangement of nucleic acids forming polynucleotide chains. These nucleic acids can be either ribose-based (RNA) or deoxyribose-based (DNA). Consequently, the ability to deduce the order of these nucleotides, known as the sequence, plays a pivotal role in addressing numerous biological inquiries[70]. In 1953, Watson and Crick successfully unveiled the three-dimensional structure and composition of DNA[71]. However, at that time, there were no techniques available to sequence or "decode" DNA. While methods for determining protein sequences, such as Edman degradation, were already in existence, DNA molecules presented distinct challenges due to their considerable length and the relative similarity of their constituent units[72]. From that moment, extensive efforts were undertaken to develop methodologies for DNA sequencing, marking the emergence of the genomic era.

1.2.1.1 First generation sequencing

The effort to sequence DNA was started by Wu and Kaiser in 1965. Their approach involved supplying DNA polymerase and one radiolabeled nucleotide at a time to a phage genome with 5' overhanging "cohesive" ends[73]. This concept was later generalized to the use of specific oligonucleotides to serve as primers for DNA polymerase. A significant breakthrough occurred when the conventional two-dimensional fractionation method, which often included both electrophoresis and chromatography, was replaced by a more powerful single separation technique utilizing electrophoresis through polyacrylamide gels. This

innovative technique was employed in two major protocols: Maxam and Gilbert's "fragmentation technique" in 1975[74] and Sanger's "dideoxy technique" in 1977[75]. The former gained widespread adoption and is often referred to as the initiation of "first-generation DNA sequencing." In contrast, Sanger's methodology was considered a groundbreaking advancement that permanently reshaped the course of sequencing. Briefly, it exploited the concepts of DNA priming, the blockage of DNA polymerase by dideoxy nucleotides and radiolabel to retrieve DNA sequences. Subsequent refinements were made to this approach, including the transition from radiolabeled to fluorophore-ligated nucleotides and the adoption of capillary electrophoresis. These developments paved the way for the creation of automated DNA sequencing machines[76–79]. These advancements marked the onset of a new era in biology, dedicated to unraveling genome composition and the functions of genes. Consequently, in 1990, the Human Genome Project was initiated at the National Institutes of Health (NIH) in the USA. This monumental scientific research project, recognized as one of the largest of its kind, was an international collaborative effort with the primary objective of determining the DNA sequence and mapping the genes on the human genome, both in terms of their physical locations and functional roles.

1.2.1.2 Second generation sequencing

In 1993, a groundbreaking technique known as "pyrosequencing" was introduced by Nyrén, Pettersson, and Uhlen. This innovation laid the foundation for a new era of DNA sequencing. The key advancement was the usage of a dual-step luminescent reaction that employed a pyrophosphate molecule, released during nucleotide incorporation, as its substrate[80]. While both Sanger's and Nyrén's methods are considered "sequence-by-synthesis" (SBS) techniques, pyrosequencing offered several advantages, including the use of natural nucleotides, real-time observation, and the incorporation of paramagnetic beads. These enhancements significantly increased the number of sequenced molecules in a single run. Automated sequencing machines based on this methodology were subsequently developed by 454 Life Sciences (later acquired by Roche).

Following the success of the 454 system, numerous parallel sequencing techniques emerged. Among them, the most notable ones are the Solexa[81] method, later acquired by Illumina, and the SOLiD[82] sequencing technique from Applied Biosystems (later known as Life Technology).

This ushered in the era of "second-generation DNA sequencing," characterized by the ability to perform an extensive number of parallel sequencing reactions on a minuscule scale. Among the available technologies, Illumina platforms have gained widespread adoption and are the most commonly used sequencing platforms. They played a crucial role in producing

the data analyzed in this study and will be further characterized in chapter 1.2.2. These innovations were instrumental in the early completion of the Human Genome Project in 2003, a milestone achieved two years ahead of the expected timeline.

Albeit very recent (2015), DNBSEQ™ is a technology by MGI that falls into this category, being an SBS approach. The innovation stands in the circularization of a single stranded DNA molecule, followed by rolling circle amplification (RCA). This method ensures that the same DNA molecule is used as template, avoiding the propagation of sequencing errors that the DNA polymerase might introduce.

1.2.1.3 Third generation sequencing

Third-generation sequencing technologies are those that can sequence individual molecules without the need for DNA amplification steps. The feature usually is coupled with the capability to sequence very long reads. Among these methods, the single molecule real-time (SMRT) platform by Pacific Biosciences is the most widely utilized. This sequencing process takes place on a chip composed of arrays of microfabricated holes with a diameter of less than 100 nanometers. These nanostructures have a single DNA polymerase molecule at their base, responsible for the sequencing process. Because the wavelength of the excitation light is greater than the diameter of the hole, it allows the light to be focused exclusively at the hole's bottom, a phenomenon known as the evanescent wave. This is where the insertion of fluorescent nucleotides occurs. This approach effectively eliminates background excitation light from non-inserted nucleotides. One of the significant advantages of this technology is its ability to generate reads of up to 10,000 base pairs (10 kb) in length[83].

A second approach has been commercialized by Oxford Nanopore Technologies[84]. It provides a similar output to the SMRT method, but it is based on a different platform. Nanopores are arrays of tiny holes that are embedded in an electro-resistant membrane. Each pore is connected to a channel and a sensor chip. The technology leverages the conformation changes that occur when a single nucleotide passes through the channel, that results in shifts of the electric current. Since every different nucleotide result in a specific current change, the sensor is able to register and assign to each passage through the channel the corresponding nucleotide. One of the advantages relies on the potential to study not only the 5 canonical nucleotides that build up a DNA or RNA molecule, but also any modification attached to them (e.g., methylation, oxidation).

1.2.2 NGS workflow and state of the art

NGS is a versatile technology with a wide range of applications, and the specific use of NGS can vary depending on the biological questions researchers are addressing. However, the

primary distinctions among different applications typically revolve around the type of input material used, its preparation, and the subsequent analysis methods employed to obtain the desired results. Regardless of the sources, the common NGS workflow typically starts with DNA (or cDNA in case of RNA input) molecules and concludes with the bioinformatic processing of raw sequencing data.

1.2.2.1 Experimental workflow

As previously mentioned, the technology applied in this study is based on Illumina platforms. The experimental workflow comprehends three main steps: library preparation, cluster amplification and sequencing.

Library preparation. A library is defined as a collection of DNA fragments that are prepared for sequencing. These fragments are generated randomly from DNA using methods such as sonication or enzymatic digestion. In Illumina-based protocols, these fragments typically fall within the range of 300-400 base pairs in length. Once the fragments are obtained, a ligation step is employed to attach specific oligonucleotide sequences, known as adapters and indexes, to both the 5' and 3' ends of each fragment, thereby creating a library for each sample. Adapters are sequences that are complementary to the oligos fixed to the glass surface of the flow cell. They enable the hybridization of fragments to the flow cell during the sequencing step. Conversely, indexes are unique sequences assigned to fragments originating from the same sample. This enables the sequencing of multiple samples within the same flow cell or lane, thus reducing costs, time, and potential biases.

Cluster amplification. The DNA library is subsequently attached to a glass slide flow cell that is coated with oligonucleotides complementary to the adapters. These hybrids serve as primers for the "bridge amplification" process, which involves repetitive denaturation and polymerase-driven extension steps. This amplification step is crucial as it enables the replication of DNA fragments, resulting in the formation of clonal clusters comprising approximately 1000 identical double-stranded DNA molecules. This clustering is essential to ensure that the sequencing signal is robust enough to be distinctly detected for each base of every fragment.

Sequencing. Illumina sequencing protocols enable the sequencing of fragment ends with read lengths of up to 150 base pairs and employ a "sequencing-by-synthesis" methodology. In this approach, fluorophore-labeled deoxynucleotide triphosphates (dNTPs) and DNA polymerase are introduced into the flow cell channels simultaneously. These dNTPs possess a reversible 3'-OH terminator that makes their incorporation a unique event. During each incorporation, laser excitation is used to illuminate the dNTPs, allowing for the optical identification of individual bases. Following this, the terminator blocking the 3' hydroxyl group is removed to enable the next base incorporation. This iterative process continues, one base at a time, ultimately revealing the sequence of the fragment ends. Modern sequencers, such as the NovaSeq6000, have reduced the number of laser excitation rounds to two (instead of four), which accelerates the fluorescence acquisition step. The four different nucleotides are distinguished as follows: one base does not emit light upon excitation, two bases emit light with either one wavelength or another, and the last base emits light when excited by both wavelengths[85](Figure 4).

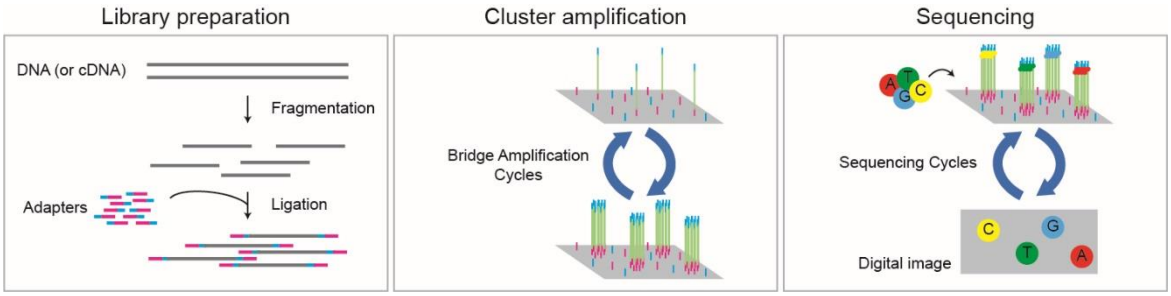

Figure 4: Illumina NGS chemistry overview. Experimental workflow of Illumina technology[85].

A fundamental requirement for the success of an NGS study is the capacity of the generated data to potentially address the specific biological question of interest. This accomplishment hinges not only on the proper execution of the sequencing experiment but also on the establishment of a robust experimental design. Such a design encompasses numerous variables, including the sequencing protocol, sequencing depth, and the number of replicates[86].

During the sequencing step, short DNA sequences known as reads are produced, and these reads typically cover only a portion of the original DNA fragment. As a result, sequencing can be conducted in two modes: single-end (SE) or paired-end (PE). In the SE mode, reads are generated from one end of the fragment, while in the PE mode, reads are derived from both ends of the fragment. The paired-end approach is particularly advantageous for NGS applications that involve the analysis of splice junctions, chimeric entities, and similar

complex structures. In addition to SE or PE sequencing, the length of the generated reads is another crucial feature to consider. Longer reads enhance the mappability of the data, making it easier to align them to a reference genome or transcriptome[87]. Furthermore, NGS libraries can be prepared in either unstranded or strand-preserving (forward or reverse) fashion. Strand-specific libraries help with the analysis and quantification of antisense or overlapping transcripts, particularly in transcriptomic studies.

Another important aspect of the sequencing protocol is the choice of the sequencing depth, defined as the number of sequenced reads for a given sample. This number is largely dependent on the aim of the experiment; deeper sequencing levels corresponds to the identification or quantification of rare and lowly abundant species. For instance, a sequencing depth of 5 million mapped reads could be sufficient to detect medium and highly expressed transcripts, whilst sequencing up to 100 million reads could be necessary to precisely quantify low expressed RNAs (such as lincRNAs).

The inclusion of an appropriate number of biological replicates in NGS studies can be pivotal in obtaining more reliable and robust results. Several factors influence the determination of this number:

- Variability in Sequencing Procedure: Accounting for the inherent variability in the sequencing process itself is important. Replicates can help assess and mitigate technical variability, ensuring that observed differences are more likely due to biological factors of interest rather than technical noise.
- Biological Variability: The biological system under study may exhibit inherent variability among individual samples, such as genetic diversity or biological fluctuations. Replicates enable researchers to capture and account for this natural variation.
- Statistical Power: The desired statistical power of the study plays a role in determining the number of replicates. Statistical power is the ability to detect statistically significant differences among experimental groups. Increasing the number of biological replicates can enhance the statistical power of the study, making it more likely to detect meaningful differences when they exist.

In summary, the decision on the number of biological replicates in an NGS study should be a carefully considered balance between minimizing technical variability, accounting for biological variability, and achieving the desired statistical power to draw meaningful conclusions and detect significant differences in the data.

1.2.2.2 Computational workflow

The computational workflow of sequencing data is organized in three levels, namely primary, secondary, and tertiary analysis. Since the third level comprehends a plethora of applications depending on the NGS technology, I will focus on the first two.

Primary analysis: When sequencing data is generated, the resulting reads are organized based on the sample from which they originated. This process is referred to as *demultiplexing*, which allows for the accurate separation of data from different samples within the same sequencing run (Figure 5). This data and the corresponding quality score are stored in a text-based format file, named FASTQ[88]. The second step recommended in sequencing analysis is the *quality control* of reads. FastQC[89] is an excellent tool used to check the quality of data coming from high throughput sequencing pipelines. The quality control of raw data consists in the analysis of sequence quality, GC content, duplicated sequences that enable the detection of biases, e.g., sequencing errors, PCR artifacts or contaminations. There are guidelines and software programs that provide indications on acceptable scores for these analyses; however, model- and experiment-specific biases may influence these scores, so they must be considered. Finally, the *trimming* step allows to remove reads, entirely or parts of, based on their quality or to remove adapters that have been retained in the read[90,91].

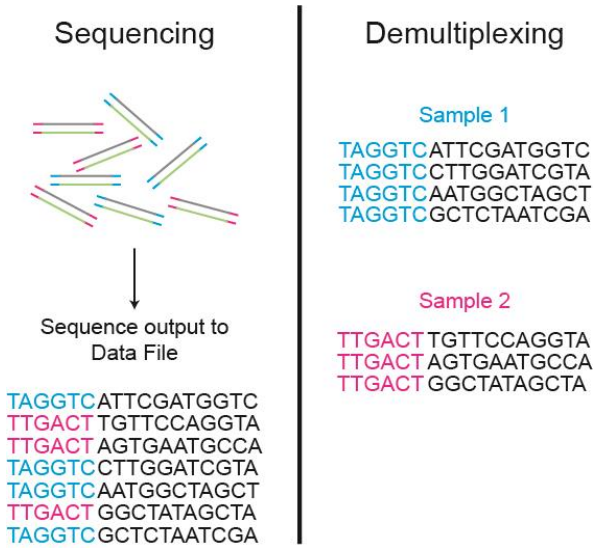

Figure 5: **Library demultiplexing overview**. DNA (or cDNA) fragments are sequenced in pool (left) and then assigned to the correct sample using their specific index (right). Adapted from Illumina[85].

Secondary analysis. It consists of one unique step, the alignment. The alignment step involves the mapping of reads to a reference genome or transcriptome. The quality of this step can be evaluated through the percentage of uniquely mapped reads that should range

between 70% and 90%, when mapping against the human genome[86]. In case reads are mapped against the transcriptome, this value could be lower due to the loss of reads coming from unannotated transcripts. Many software[92–94] have been developed to perform reads alignment, and their different features are best suited for different technologies (Figure 6).

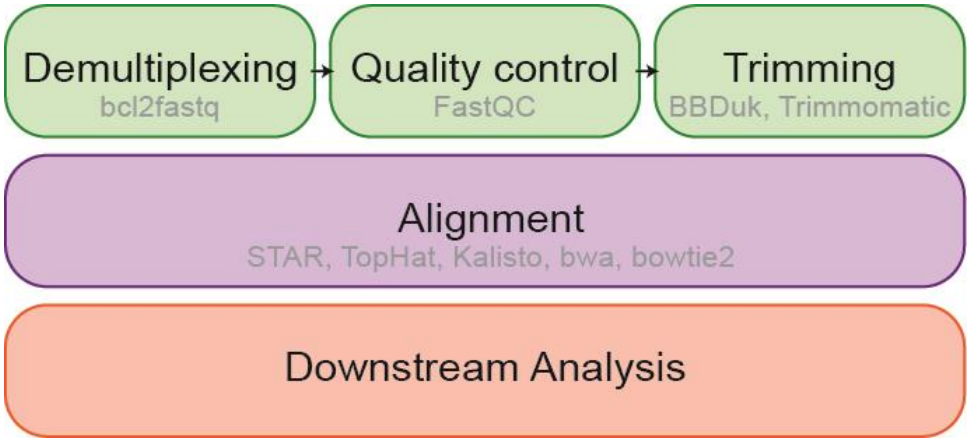

Figure 6: NGS computational workflow. The steps that characterize a typical NGS computational workflow are divided in primary (green), secondary (purple) and tertiary (red) analyses. Common software are reported in grey.

1.2.2.3 NGS State of the art

NGS empowers the simultaneous investigation of hundreds to thousands of genes across multiple samples and facilitates the discovery and analysis of various genomic features in a single sequencing operation. This encompasses the detection of diverse genetic alterations, ranging from single nucleotide variants (SNVs) to copy number variations and structural variants, as well as the identification of RNA fusion events. NGS offers an optimal balance of throughput, making it possible to process a substantial amount of data in a single run, while also ensuring swift and cost-effective studies. Furthermore, NGS boasts several additional advantages, including reduced sample input requirements, heightened accuracy, and the ability to identify variants at lower allele frequencies compared to traditional Sanger sequencing methods.

The speed, throughput, and precision of NGS have initiated a revolution in genetic analysis, unlocking new applications in genomic and clinical research. It has also found application in fields such as reproductive health, environmental studies, agriculture, and forensic science, broadening its impact across diverse domains of science and research (Figure 7).

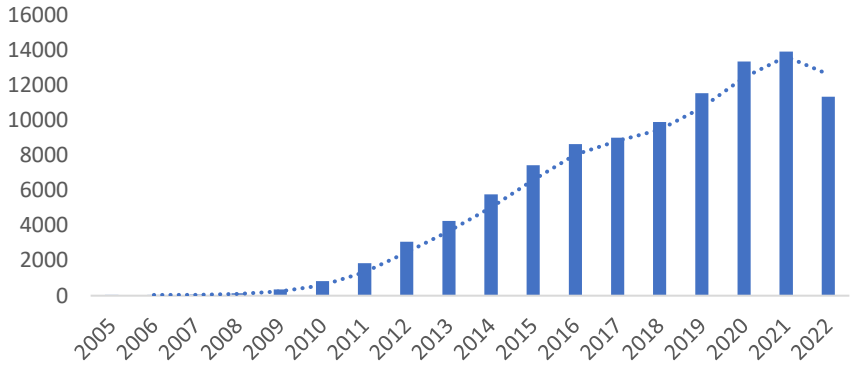

Figure 7: NGS state of the art. Number of publication on PubMed with the pattern "next generation sequencing" per year.

1.2.3 Single-cell RNA sequencing

Over the past decade, population-based RNA sequencing methods, often referred to as bulk RNA-seq, have played a crucial role in unraveling genome-wide variations in gene expression across a wide spectrum of disciplines, including cancer biology, developmental biology, and cellular homeostasis[64,95,96]. However, as bulk RNA-seq data represents an average of gene expression across individual cells, it can obscure the nuanced transcriptional patterns within distinct subpopulations, especially those of the least prevalent cell types or states[97].

The advent of single-cell RNA sequencing (scRNA-seq) has effectively addressed this challenge, providing unprecedented opportunities to explore gene expression profiles at the resolution of individual cells. Since its initial introduction in 2009, scRNA-seq has opened new avenues for uncovering the inherent cellular heterogeneity within complex systems[98,99]. Today, thanks to the emergence of efficient and cost-effective technologies, it is possible to construct sequencing libraries encompassing thousands of individual cells, thus encouraging the adoption of single-cell technology as a routine procedure.

These technological advancements have facilitated the discovery of novel cell types[100,101] and the investigation of dynamic cellular processes at previously unattainable spatial and temporal resolutions[102–105]. Moreover, scRNA-seq has become a fundamental component of the rapidly evolving field of precision medicine[106,107]. The wealth of new information acquired through scRNA-seq has the potential to reshape our understanding of developmental biology, gene regulation, and the diversity of cells in both health and disease.

1.2.3.1 Sample preparation

A key aspect of sample preparation for scRNA-seq analysis revolves around the barcoding of the entire transcriptome of individual cells. Once a viable single-cell suspension is prepared, cell viability is assessed, and lysed cells are removed[108], various technologies have prioritized the refinement of methods for isolating single cells and subsequently applying barcodes to them. In recent years, microfluidic-based technologies have gained significant popularity owing to their cost-effectiveness, high efficiency, and their ability to generate data with moderate file sizes, which helps maintain data integrity and coherence[109,110]. These microfluidic technologies, exemplified by platforms like Chromium[111](Figure 8), inDrop[112], and Drop-seq[113], operate by allowing the passive co-flow of cells, microparticles (often referred to as beads), and a lysis buffer. This combination results in the formation of water-in-oil droplets, each encapsulating precisely one cell and one bead. Within these droplets, the transcriptional content of each cell is captured and subsequently amplified using unique primers that are attached to the surface of individual microparticles. These primers share a common three-part structure:

- Cellular Barcode: A short sequence shared by all primers on a single microparticle, serving the purpose of identifying all transcripts originating from the same cell.
- Unique Molecular Identifier (UMI): A molecular tag specific to each transcript, which ensures the integrity of the sequenced reads by flagging PCR duplicates[114].
- Poly-T Tail: Facilitates the capture and amplification of the 3' end of each transcript.

This approach allows for the precise and efficient capture, barcoding, and amplification of the transcriptome of individual cells within the microfluidic droplets.

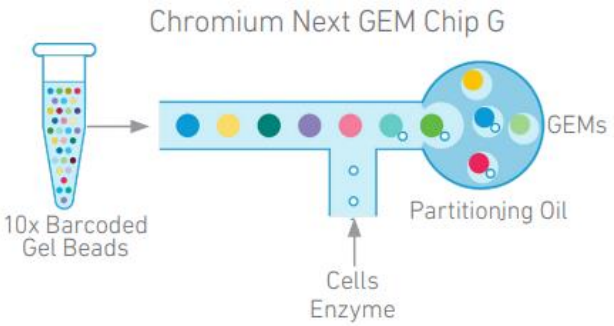

Figure 8: GEMs generation. The Chromium platform uses oil to create micro-chambers where reactions can occur. In each chamber (GEM), a bead covered with primers, cells and enzymes are stored. Adapted from 10X Genomics[186].

1.2.3.2 Applications

Tertiary analysis of scRNA-seq data must face several obstacles, including large-scale data and high levels of noise interference due to dropout events[115–117]. Some standardized pipelines have been proposed to facilitate downstream analysis[118], however standalone software might be helpful to obtain the biological insights of interest. *Reads quantification* of single transcripts is a feature common to every scRNA-seq pipeline. Differently from bulk approaches, the aforementioned UMIs avoid misinterpretation of gene expression levels due to technical artifacts[114]. Once read counts are retrieved, data can be used to:

- **Cluster cells.** As transcriptionally unique populations of cells often correspond to distinct cell types, a primary objective of scRNA-seq is to identify cell subpopulations based on their transcriptional similarities[119]. Therefore, the process of grouping cells into clusters enables the de novo discovery of cell types or the identification of various subpopulations within a single-cell state. This clustering approach is fundamental for unraveling the heterogeneity of cellular populations and gaining insights into the diversity of cell types and states within a given sample.
- **Identify cell markers.** The characterization and annotation of the groups of cells identified by a clustering algorithm can be achieved by identifying marker genes, often referred to as the cluster gene signature, through a process known as differential expression analysis. In this analysis, marker genes are singled out by comparing the gene expression profiles of cells within each individual cluster to those of all other cells in the dataset. This method allows researchers to pinpoint genes that are uniquely and significantly associated with each cluster, providing valuable insights into the distinct functional or phenotypic characteristics of the cell populations within the data.
- **Trajectory inference.** Since many biological mechanisms are inherently dynamic processes, they cannot always be effectively characterized using a discrete approach like clustering. Trajectory inference pipelines are a class of computational methods emerged to model continuous biological systems, including developmental processes. Monocle[120] first introduced the concept of *pseudotime*, a robust methodology to describe developmental systems. Lately, the optimal transport problem has also been implemented to model cellular process over time, hence inferring trajectories[66].

1.2.3.3 Multiome

Complex biological systems might not be easy to disentangle by solely addressing gene expression differences at the single-cell level. Transcriptional regulation upon histone

modification rearrangements, DNA methylation status, and chromatin accessibility plays an important role in many processes. While single-cell epigenetics has seen successful implementation in recent years[121,122], one ongoing challenge is the integration and correlation of epigenetic information with scRNA-seq data. Such integration can be complicated because even replicates, which involve multiple measurements taken from the same experimental condition, are still composed of different individual cells. Each cell, even within the same condition or replicate, can exhibit inherent biological variability due to factors like genetic diversity, stochastic gene expression, and microenvironmental influences.

Multiome analysis allows the user to experimentally overcome such hindrance by combining scRNA-seq and single-cell Assay for Transposable-Accessible Chromatin sequencing (scATAC-seq) data from the same exact cell. 10X Genomics provided a methodology to isolate nuclei and flag accessible DNA and retro-transcribed RNA, by using different barcodes[123](Figure 9).

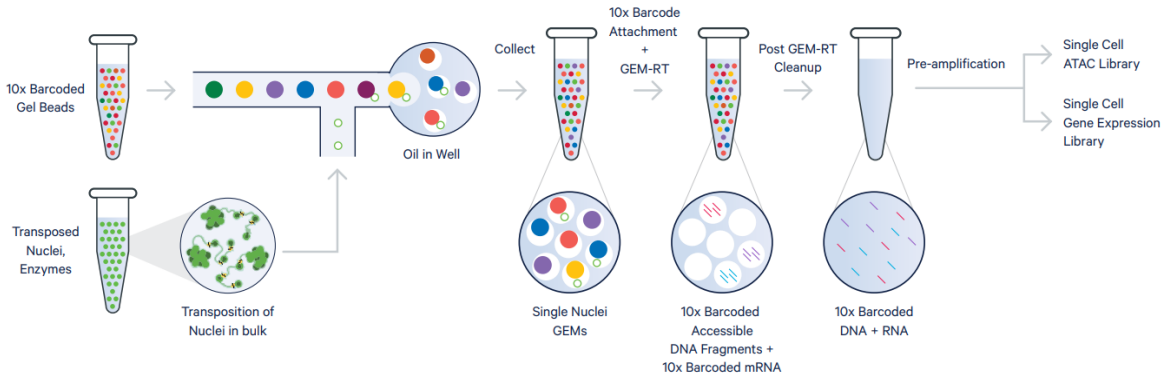

Figure 9: Multiome workflow. Transposed nuclei and enzyme enter the GEM with beads covered in 10X primers. Afterwards, different adapters are ligated to Accessible DNA fragments and retro-transcribed RNA. Adapted from 10X Genomics[123].

2 Materials and Methods

2.1 Microfluidic device

In this work, we used a microfluidic device, fabricated by soft lithography technique and replica molding, previously published by our collaborators' group[124]. Polydimethylsiloxane (PDMS) with a 10:1 base/curing agent ratio (Dow Corning) was coupled to a borosilicate glass slide (Menzel–Gläser) through plasma treatment of surfaces.

Briefly, the microfluidic platform consists of 5 independent culture chambers, with the following dimensions: 18.8 mm of length, 1.5 mm of width, and 0.2 mm height with a 5.6 μL volume. The device is sterilized by autoclaving before use. During experiments the microfluidic chips are placed in a dish, surrounded by a water bath to reduce medium evaporation.

2.2 Cell culture

BJ cells (Miltenyi Biotec, 130-096-726), human newborn skin fibroblasts, were cultured with complete Dulbecco's modified Eagle's medium (DMEM, Thermo Fisher, 41965 or 11965), supplemented with 10% fetal bovine serum (FBS, Thermo Fisher, 10270106 or 10099-141). Cells were maintained at 37 °C in the presence of 5% CO2 and periodically tested for mycoplasma contamination.

2.3 Reprogramming in microfluidics

Microfluidic cell cultures were performed as follows. On day 0 human fibroblasts were seeded in the microfluidic chambers, at a density of 60 cell/mm2, after a coating with 25 μg/mL of cold fibronectin (Sigma Aldrich). Before placing chips in the incubator, 1 ml of PBS 1× was added to the bottom of the dish, in order to maintain proper humidity. From day 1 to day 8, in the morning medium was replaced using Reprogramming Medium, whereas in the night mmRNAs transfection was performed, as reported in Gagliano et al., 2019[124]. From day 9 to day 15, medium change was performed every 12 h using Pluripotency Medium.

2.4 Reprogramming of human fibroblasts to hiPSC colonies

We generated hiPSCs from human foreskin BJ fibroblasts using microfluidic technology as previously described[124]. For proteomic analysis, a total of 10 mRNA transfections were performed using StemRNA-NM reprogramming kit (Stemgent, 00-0076) and StemMACS

mRNA transfection kit (Miltenyi, 130-104-463), in E7 medium, made from E6 medium (Thermo Fisher, A1516401) supplemented with 100 ng/mL FGF2 (Peprotech, 100-18B-1000), switched to E8 medium (Stem Cell Technologies, 05990) from day 11. Whereas, for single-cell RNA-seq, 8 mRNA transfections were performed in supplemented Pluriton medium (Stemgent, 00-0070), switched to StemMACS iPSBrew XF medium (Miltenyi Biotec, 130-104-368) from day 9. Validation experiments were performed either in microfluidics according to single-cell RNA-seq protocol or in standard 24-well plates according to manufacturer's instructions; they were performed under suboptimal conditions to enhance reprogramming efficiency differences, and medium was supplemented with HGF 100 ng/mL (Peprotech, 100-39), IL-6 50 ng/mL (Peprotech, 200-06), IL-6r 10 ng/mL (Peprotech, 200-06 R), NRG1 100 ng/mL (R&D, 396-HB), during the whole process duration, according to the specified perturbation conditions using both Pluriton medium and Nutristem hPSC XF Medium (Biological Industries, 06-5100-01-1 A) supplemented with 20 ng/mL FGF2. The loss of function experiments were performed in microfluidics supplementing the medium with Jak Inhibitor I 1uL (Millipore, 420097) and c-METi 600 uM (Selleck, PF-02341066) from day 1 to day 6. In STAT3 knock-out experiments, siRNA STAT3 10 uM (Qiagen, 1027416) or MOCK siRNA 10 uM (Qiagen, 1027284) was added in the transfection mix from day 1 to day 6. In all cases, the whole process was performed in a hypoxia incubator (5% O2, 5% CO2) at 37 °C.

2.5 Sample preparation for LC-MS/MS

During reprogramming, at every medium change or reprogramming transfection, medium was collected in three replicates, pooling together the conditioned medium from the same 40 channels for each replicate. The media were stored at −80 °C until prepared for proteomic analysis. After thawing, media from four collections (two consecutive days) were pooled together. For example, sample D1-D2 was conditioned by the cells within the microfluidic chamber from day 1 to day 3 mornings. 3 kDa cut-off centrifugation membranes (Amicon Ultra 0.5 mL, Ultracel 3 K, Merck) were used for filter-aided sample preparation (FASP)56. Proteins were concentrated by centrifugation for 20 min at 4 °C and 14,000 g, then washed twice with a 50 mM triethylammonium bicarbonate (TEAB, Thermo Scientific) buffer containing 8 M urea (Sigma-Aldrich). Protein content was quantified by Pierce BCA Protein Assay Kit (Thermo Scientific). Each sample proteins were reduced for 60 min at 56 °C with 100 mM DTT (Sigma-Aldrich) and alkylated for 30 min at room temperature in the dark with 55 mM iodoacetamide (Sigma-Aldrich). Samples were washed with 50 mM TEAB for three times. An equal amount of protein for each sample was digested by trypsin (Promega) at 37 °C for 16 h. Digested peptides were desalted by C-18 spin column (Pierce) and vacuum

dried. Then, labeling by 6-plex Tandem Mass Tag (TMT6, ThermoScientific)[125] was performed according to manufacturer's instructions using 50 μg of peptides from each sample. The six-time point samples of each of the three replicates were pooled, then desalted and vacuum dried.

2.6 Mass spectrometry analysis

25 pre-fractions were collected on UPLC (Agilent 1290) with high pH C18 column (2.1 mm × 30 mm). Before MS analysis, peptides were resuspended in 10 μL of 0.1% formic acid. Thermo Fusion Mass Spectrometer coupled with Thermo EasynLC1000 Liquid Chromatography was used to get peptides profiles. 90 min of LC-MS gradients were generated by mixing buffer A (0.1% formic acid in water) with buffer B (0.1% formic acid in 80% ACN in water) by different proportions. Using NSI as the ion source and Orbitrap as the detector, the mass scan Range was at 300-1800 m/z, and the resolution was set to 120 K. The MS/MS was isolated by Quadrupole and detected by Ion trap, whose resolution was set to 60 K. The activation type was HCD.

2.7 Proteomic bioinformatic analysis

Peak list files were searched against UniProt human reference proteome (UP000005640) by MaxQuant (v. 1.6.3.4)[126]. TMT6 modification and carbamidomethyl on cysteine were set as fixed modifications. The oxidation of methionine, acetylation of protein N-terminus, and phosphorylation (STY) were set as variable modifications. Peptide-spectrum matches (PSMs) were adjusted to 1% and then assembled further to a final protein-level false discovery rate (FDR) of 1%. Proteins not identified in at least 2 replicates in at least one time point were excluded from further analysis. Common contaminants (keratins and Bos taurus proteins) were also filtered out, for a final number of 4542 proteins identified. Missing values were imputed by the mean value of the other two replicates. TMT intensities were normalized according to BCA quantification to obtain a relative quantification proportional to protein concentration in culture. The distributions of the three replicates of TMT intensities were scaled by their respective medians. A principal component analysis (PCA) was performed in MATLAB R2017a (The Mathworks) using mean-centered TMT intensities. A list of secreted proteins was manually annotated by integrating the following resources: secreted proteins predicted by MDSEC as reported in Protein Atlas database[127] (http://www.proteinatlas.org), secreted proteins in Gonzalez et al., 2010[128]; a list of ligands from Gene Ontology-Molecular Function categories "cytokine activity", "growth factor activity", and "hormone activity", and senescence-associated secreted proteins (SASP) annotated from literature[129–132]. Of the proteins identified in this study, only those secreted

according to the criteria above were further studied, in order to avoid the proteins possibly derived from cell death. Differentially secreted proteins between time pairs were assessed with student t-test, using a threshold of 5%. Proteins whose concentration was maximal only at the first time point (D1-D2 sample) were excluded from further analysis, as potential residual proteins from FBS used during fibroblast expansion. Functional enrichment analysis of Reactome pathways was performed using ReactomePA (v1.36.0)[133] Bioconductor package. Reactome hierarchy was visualized using ClueGO (v2.5.6)[134] within Cytoscape (v3.8.0)[135]. Genes specific to different human embryonic stages were derived from a published single-cell RNA-seq study[136], of these core ECM genes were selected based on the annotations in Naba et al., 2012[137]. Proteins playing a role as ligands were taken from Ramilowski et al., 2015[138]. Hierarchical clustering with heat map data visualization was performed in MATLAB R2017a, using Euclidean distance and complete linkage.

2.8 Sample preparation for single-cell RNA-seq

For each time-point, cells were detached using TrypLE-express (ThermoFisher, Gibco 12604). Harvested cells were then centrifuged at 300 g and resuspended at the final cell density of 100 cells/mL using a solution containing 40% KnockOut Serum Replacement (KSR, ThermoFisher, Gibco 10828) in DMEM. For each timepoint, two replicates were produced, each containing cells from 4 independent chips that were pooled together then divided in aliquots containing 5000-80,000 cells. Samples were cryopreserved in DMEM supplemented with 40% KSR and 15% DMSO and stored in liquid nitrogen.

scRNA-seq libraries were generated using one or two samples for each replicate. Briefly, each cryopreserved aliquot was thawed at 37 °C until a tiny ice crystal remained in solution. Then each sample was diluted under gentle shaking by dropwise adding 10 volumes of DMEM supplemented with 40% KSR. Cells were washed twice using a washing buffer containing 8% MACS Running Buffer (Miltenyi, 130-091-221) in PBS. Cells were then resuspended in the washing buffer and filtered through a 40 µm cell strainer (Biosigma, 010198Z). Cell viability and concentration were checked by visual inspection using Trypan Blue (Logos Biosystems, L12002).

Single-cell RNA seq libraries were produced according to 10X Single Cell 3' v2.0 standard protocol and sequenced on Novaseq 6000 (Illumina).

2.9 Single-cell RNA-seq data pre-processing

scRNA-seq data pre-processing was performed using the cellranger software (v 2.2). Fastq files were generated using the cellranger pipeline *mkfastq* using 10X standard Chromium

barcode sequences. Alignment, filtering, barcode and UMI counting were performed using the cellranger *count* pipeline. Human pre-built genome index has been applied (hg38 genome reference and GRCh38 annotation, including protein coding, linc and antisense RNAs). Each feature-barcode matrix from each independent sample was merged to build up the final dataset, containing 33,694 genes and 44,197 cells, then subjected to cells and genes filtering. Cells having less than 1000 detected genes and with the mitochondrial associated reads percentage greater than 10% were filtered out. Furthermore, in order to have a homogenous sampling for each reprogramming day, the cell dataset was randomly subsampled to 2500 cells per time point. The final dataset retained only those genes expressed in at least 5% of all the cells, leading to 12,932 total genes. Gene expression values were normalized to CPM (counts per million) and transformed to the $\log_2$ scale using a pseudocount of 1. Finally, cell-cycle scores and, consequently, phases were assigned to each cell by Seurat's (v.3.1.5) *CellCycleScoring* function.

2.10 Single-cell RNA-seq data visualization and clustering

To better visualize and characterize single cell data, high dimensionality was reduced. First, we computed the neighborhood graph using the function *compute_neighborhood_graph* from the Python (v 3.9.5) package wot (v 1.0.5)[66], using 50 neighbors and choosing the first 100 PCA components and the first 20 diffusion map components. The resulting 120 components were used as input to initialize the Force-Directed Layout Embedding (FLE) algorithm, using forceatlas2 (v 1.0.3) with 1000 iterations and reducing the space to 2 dimensions (FLE1 - FLE2). The same components were also applied to perform an unsupervised graph-based algorithm (louvain) using the *FindNeighbours* and *FindClusters* (resolution = 0.6) functions in the Seurat (v.3.1.5)[139] package. This step resulted in the identification of 12 clusters, annotated based on the enrichment of somatic and developmental signatures[64] at the single-cell level (SR = somatic related; DR = developmental related; NA = not assigned) and ordered by their composition in terms of time-points.

2.11 Single-cell RNA-seq differential gene expression and gene sets enrichment

Differentially expressed genes among clusters were identified using the *FindAllMarkers* function from Seurat (v.3.1.5), taking just LFC ($\log_2$ fold change) more than 0.25. For each gene, significance was assessed with the Wilcoxon rank-sum test P values, adjusted for multiple testing using the Benjamini–Hochberg correction to retrieve the false discovery rate (FDR). Only genes with FDR < 0.01 were considered. As expected, many gene markers were

shared by clusters from the same group (SR or DR) because of the continuous nature of data. We therefore decided to select unique markers and to take duplicated markers once, preferring the cluster where the LFC was the highest.

To perform enrichment of gene signatures in clusters, we used pre-ranked Gene Set Enrichment Analysis (GSEA) from fgsea (v 1.14.0)[140] R package. Pre-ranked lists for each cluster were generated by assigning to each gene its LFC relative to the average expression across all the other clusters. Common pathways were defined as belonging to several databases, i.e. Hallmark[141], KEGG, Biocarta, Reactome and Gene Ontology Biological Process.

Enrichment scores (ES) of gene signatures at the single cell level were obtained by computing the z-score for each gene across the data sheet. After truncating these scores at 5 or −5, the enrichment score was defined by the average z-score over all genes in the gene set.

2.12 Single-cell RNA-seq trajectory inference

To infer the reprogramming trajectory, two different approaches were used: wot (v 1.0.5)[66] and Monocle3 (v 0.2.3.0)[142]. The former applies the Mass Optimal Transport theory to the gene expression space to infer, for each cell in a given sample, the most probable ascending and descending cells in the previous and following timepoints. First, birth-death rates were computed for each cell by applying a logistic function to the enrichment scores for Cell-cycle[143] and Apoptosis (R-HSA-109581, hsa04210, HALLMARK_APOPTOSIS in Liberzon et al., 2015[141]). ß and δ logistic functions were optimized (center = −0.1 and center = 0.15, respectively). Second, transport maps were generated in batch for each pair of subsequent time-points using the functions *wot.ot.OTModel* (epsilon = 0.2) and *compute_all_transport_maps*. Finally, trajectories were inferred using *population_from_cell_sets* and trajectories functions starting from D15 cells that showed high enrichment (> 2) for the signatures Matrisome[137] and Late pluripotency[64]. For each timepoint, cells having a trajectory probability greater than the mean were considered to belong to the trajectory.

Monocle 3, on the other hand, learns a trajectory graph looking at the gene expression changes required for each cell to move from one state to another during a dynamic biological process. In particular, UMAP coordinates in Monocle 3 were replaced with the FLE ones, in order to obtain an FLE-based Monocle trajectory. Furthermore, *cluster_cells* and *learn_graph* were performed by tuning the parameters k (30) and ncenter (96), respectively.

2.13 Single-cell RNA-seq interaction analyses

Interaction analyses have been performed on a set of 82 ligand-receptor pairs obtained as follows.

A putative list of 3333 couples has been generated from the ligands identified in the secretome analysis with every possible receptor. Afterwards, receptors have been filtered out in case they were not defined as receptor on BioGrid or they did not belong to any of these GO terms: GO-CC:0009897, GO-CC:0098802 and GO:0004714. The resulting list of 1082 pairs was then filtered based on the expression of both ligand and receptor in at least one cell (491). Finally, we selected only those pairs that were experimentally validated[138].

Interaction scores between trajectories throughout the time-course were evaluated as shown in Schiebinger et al., 2019[66]. Top interactors were selected by ordering the results by standardized interaction score (sIS). Then, the highest ligand-receptor pair for each day was assessed. All the unique couples with a sIS comprised between the first and the last day-specific occurrence was taken.

HGF/MET cluster-to-cluster interaction scores were computed as the product between the average gene expression value of MET in one cluster and the value of HGF in another. Significance was assessed with empirical p-value, generating a null distribution of 1000 permutations on the association between cells and clusters.

2.14 STAT3 targets expression

STAT3 targets were identified using a ChIP-seq dataset on HUS64 human embryonic stem cells[144]. In particular, STAT3 target genes were defined as genes with STAT3 significant peaks at ±3000 bp from the transcription start site. For each cell, the STAT3 pathway enrichment was computed from the scaled gene expression matrix as the average value for all the STAT3 targets. For each enrichment value, the corresponding p-value was calculated by performing a hypergeometric test and using a random gene list to obtain the null distribution.

2.15 Bulk RNA-seq analysis of reprogramming data

To analyze the relationship between mouse feeders and human reprogramming cells at day 8, we re-analyzed bulk RNA-seq data from Cacchiarelli et al., 2015[64]. Fastqs have been trimmed using Trim Galore (https://github.com/FelixKrueger/TrimGalore) for quality and adapters removal. Then, reads have been mapped with TopHat (v. 2.1.0)[145] and Bowtie2 (v. 2.3.2)[94] with default parameters against a hybrid build of the human (hg38) and mouse

(mm10) genomes. Reads aligned to the mouse reference were few (alignment rate <20%), but it was consistent with the purified nature of the samples, where mouse cells should just represent contamination. Finally, read quantification was performed with HTSeq (v. 0.9.1)[146] on GENCODE human (GRCh38) and mouse (mm10) genome annotations, including protein coding, linc and antisense RNAs. The final count matrix was created by merging mouse and human genes by orthology and differential expression analysis was performed between human and mouse (feeders) samples using DESeq2[147].

2.16 Immunofluorescence staining

For immunofluorescence staining, cells were fixed in 4% paraformaldehyde for 10 min at room temperature, then permeabilized with 0.1% Triton X-100 for 10 min, blocked in blocking solution (DPBS with 10% horse serum and 0.1% Triton X-100 for intracellular targets) for 45 min, followed by overnight incubation with primary antibodies. The following antibodies were used for immunofluorescence: rabbit anti-NANOG (Cell Signaling, 4903)(1:200), mouse anti TRA1-60 (Millipore, MAB4360)(1:100), mouse anti-STAT3 (Cell Signaling, 9139)(1:300), goat anti- HGFR/c-MET (R&D, AF276)(1:200). Alexa488 or Alexa594 conjugated rabbit, mouse or goat secondary antibodies (1:200) were used (Life Technologies, A21202; A21207; A11058). The nuclei were stained with Hoechst 33342 (Life Technologies).

Images were acquired on a confocal TCS SP5 microscope (Leica) at 40x magnification and on a fluorescence microscope DM6B (Leica) at 5 and 10x magnification.

2.17 Assessment of reprogramming efficiency

Reprogramming efficiency was quantified after immunostaining with TRA1-60 and NANOG markers. When the efficiency of reprogramming was too high to allow counting single colonies, it was quantified as relative TRA1-60+ and NANOG+ cell area divided by the total area occupied by the cells. Since TRA1-60 is a membrane/extracellular marker and NANOG is a nuclear marker, we considered TRA1-60 area positive only where it overlapped with NANOG positive nuclear area, for having the double positive cells as result.

2.18 Secondary reprogramming experiments

Secondary reprogramming experiments were performed as previously reported in Cacchiarelli et al., 2015[64]. Briefly, 105 TERT-immortalized secondary fibroblasts (hiF-T) harbouring a doxycycline-inducible OSKM cassette were seeded with or without irradiated mouse embryonic fibroblast (MEF) in a 3:1 ratio. The day after seeding, cells were treated with doxycycline (Sigma Aldrich, D9891-1G) (2 μg/mL) to start the OSKM expression. In

addition, LSD1 inhibitor RN-1 (MERK, 489479) was added at the final concentration of 10 nM to further increase the reprogramming efficiency. Both treatments were prolonged for 21 days. Colony counting and visualization in bright-field were performed by using a TRA-1-60 chromogenic staining[148].

2.19 Sample preparation for Multiome analysis

For Multiome analysis, each cryopreserved aliquot was thawed at 37°C as mentioned in chapter 2.8. The cells from each time point were combined to generate a final sample containing an equal representation of each reprogramming day. Subsequently, 15,000 viable cells were filtered through a 40µm cell strainer (Biosigma, 010198Z) and used as input for the standard Single Cell Multiome ATAC + Gene Expression assay (10X Genomics, 1000285). Briefly, the cells were permeabilized using digitonin and incubated at 37°C for 30 minutes with the Tn5 transposase. After the incubation, the transposed nuclei were loaded into the Chromium controller (10X Genomics, 1000204) for DNA and mRNA capture and single-nuclei level barcoding. The sample obtained was divided into two parts to generate scRNA-seq and scATAC-seq libraries, respectively. Both were sequenced independently on an Illumina Novaseq 6000, following the 10X Genomics specifications.

2.20 Multiome data pre-processing

Multiome data pre-processing was performed using the cellranger-arc software (v 2.0). Fastq files were generated using the cellranger-arc pipeline *mkfastq* using 10X standard Chromium barcode sequences. Alignment, filtering, barcode and UMI counting, as well as peak calling (for scATAC-seq profiles) were performed using the cellranger-arc *count* pipeline. Human pre-built genome index has been applied (hg38 genome reference and GRCh38 annotation, including protein coding, linc and antisense RNAs). To be able to merge this dataset with the scRNA-seq data from this work, scRNA-seq data was re-processed using a newer version of cellranger (v 7.0.0). Gene-barcode matrix from Multiome was subjected to cells filtering. Cells having less than 1000 detected genes and with the mitochondrial associated reads percentage greater than 7% were filtered out, 0. The final dataset retained only those genes expressed in at least 5% of all the cells, leading to 36601 total genes. The scRNA-seq dataset from this work and the Multiome were merged using reciprocal PCA (RPCA) function of Seurat (v 4.3.0.1)[139] package to find anchors. With this approach, we project each dataset into the others' PCA space and constrain the anchor by the same mutual neighborhood requirement. The merged gene expression values were normalized to CPM (counts per million) and transformed to the log2 scale using a pseudocount of 1.

2.21 Merged scRNA-seq data visualization and clustering

Principal Component Analysis (PCA) dimensionality reduction and uniform manifold approximation and projection (UMAP) were performed using the functions *RunPCA* and *RunUMAP* from Seurat. The same components were also applied to perform an unsupervised graph-based algorithm (louvain) using the *FindNeighbours* and *FindClusters* (resolution = 0.5) functions from Seurat.

2.22 Multiome data processing

Multiome data was processed using Signac (v 1.10.0)[149]. Plots for the concomitant visualization of gene expression and chromatin accessibility were performed with the functions *LinkPeaks* and *CoveragePlot*.

2.23 Statistics and reproducibility

Sequencing data were analyzed and plots were produced in R (v 4.2.0). Data variability is presented as boxplots, where bars indicate the median, boxes indicate the 25th and 75th percentiles, whiskers represent median + /- the interquartile (25-75%) range multiplied by 1.5. The number of replicates and the tests used to assess statistical differences are reported within each figure caption. Experiments shown in Figure 4 have been repeated 10 times independently with similar results.

3 Results

3.1 Human cell reprogramming in microfluidic system

Low efficiency has long hindered the capability of dissecting the molecular regulatory logic behind human somatic reprogramming. However, it has been recently reported that the generation of hiPSCs can be drastically improved in a microfluidic confined environment[124,150,151], due to the accumulation of secreted factors[152–157] that sustains the acquisition of both primed[124,151] and naive human pluripotency[158]. Therefore, we took advantage of reprogramming in microfluidics (uF) to test the hypothesis that the communication between distinct intermediate sub-populations and their shared extracellular environment lying in-between contributes to shaping the route to pluripotency. With respect to non-uF methods, this system generates a more efficient reprogramming, with a considerably and significantly higher number of pluripotent colonies retrieved at the end of the process (Figure 10).

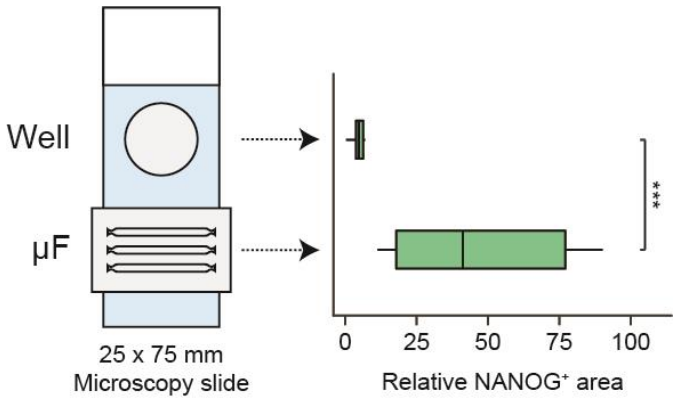

Figure 10: Comparison of efficiency between reprogramming systems. Left: Schematics of the in-scale conventional (Well) and microfluidic (µF) setup. Right: Comparison of reprogramming efficiency between the two systems. Two-sided Wilcoxon's test was used to assess differences among the conditions. n = 8 for Well and n = 15 for µF (***P = 0.0001593).

3.1.1 Experimental design

The main objective was to develop an integrated temporal multi-omic approach, combining high-efficiency reprogramming with high-throughput single-cell RNA sequencing (scRNA-seq) and tandem mass spectrometry (LC-MS/MS) on conditioned media to decipher finely regulated dynamics of secreted proteins accumulating in the extracellular space and corroborate it with cellular heterogeneity arising during intermediate stages of reprogramming.

Reprogramming of human fibroblasts was achieved with daily transfections of non-modified messenger RNAs (mRNAs) encoding for POU5F1 (OCT4), SOX2, KLF4, MYC, LIN28, and NANOG (Methods). The reprogramming protocols used to generate the single-cell transcriptome and secretome data were almost identical: however, some adjustments were made to maximize the effectiveness of identifying the endogenous secreted proteins (Figure 11).

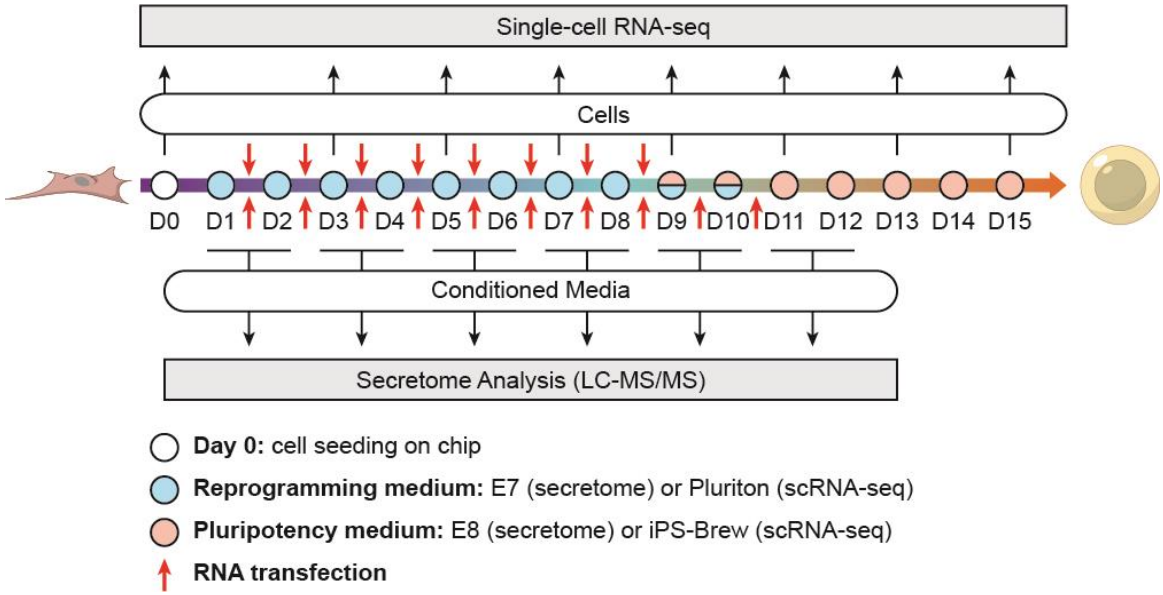

Figure 11: Experimental design. scRNA-seq data were collected by stopping parallel experiments at day 0, 3 and every 48 h. Proteomic data were obtained by tandem mass spectrometry analysis of conditioned media along the same reprogramming experiments. The differences in medium usage and reprogramming factors suppliance (RNA transfections) are reported.

For instance, we used a chemically defined medium based on E6 medium with the addition of FGF2 which shows to preserve the high efficiency of microfluidic reprogramming while enabling high-resolution and accurate detection of cell-secreted proteins (Figure 12).

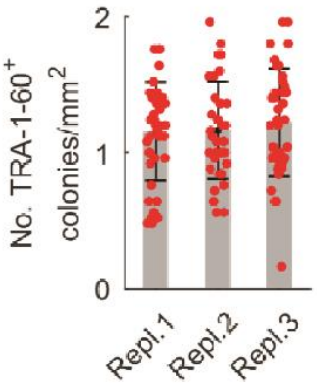

Figure 12: Reprogramming efficiency of the protocol used for secretome data. Efficiency evaluated within the same microfluidic channels used for proteomic analyses (n=40 for Repl.1, n=35 for Repl.2 and n=39 for Repl.3). Data are presented as mean values +/- SD.

3.1.2 Quality assessment

It is crucial to address the quality of reprogramming before processing the samples for any downstream analysis of interest. To achieve this goal, it is worth looking at the morphological changes associated with conventional human cell reprogramming, since it has been already reported that they are recapitulated in our microfluidic system[124]. These critical steps can be considered hallmarks and they consist of quick mesenchymal to epithelial transition (MET - before day 5), epithelial cells clustering (from day 5 to day 8) and hiPSCs colony formation as soon as day 9 (Figure 13). To further confirm the presence of pluripotent colonies at the end of the process, we also performed immunostaining of multiple channels at day 14 using antibodies against NANOG and TRA-1-60, known pluripotency markers (Figure 13). The presence of both signals further demonstrates that reprogramming in microfluidics yields a great proportion of pluripotent cells with respect to the seeded ones.

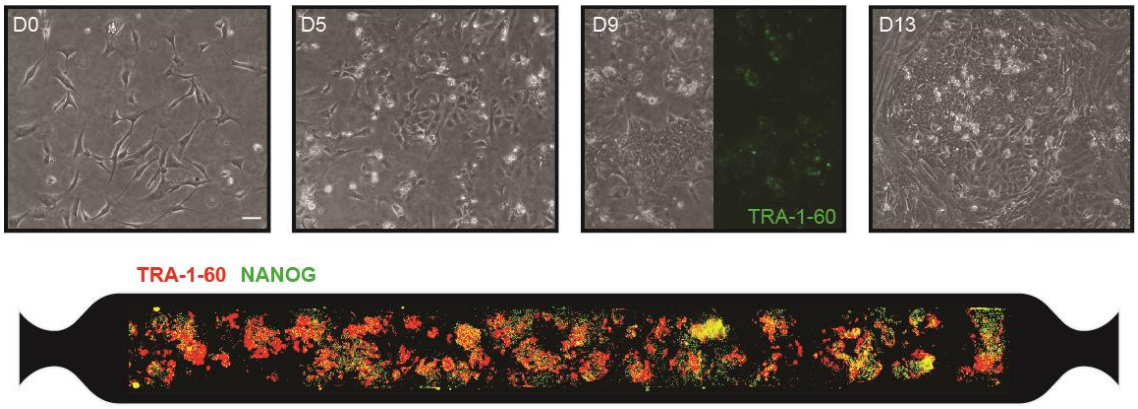

Figure 13: Morphological evaluation. Top: Morphological changes occurring during reprogramming, sampled at day 0 (D0), day 5 (D5), day 9 (D9) and day 13 (D13). Bottom: Immunostaining of a single microfluidic channel at day 14 for pluripotency markers (NANOG and TRA-1-60). Scale bar: 100nm.

3.2 Development of a temporal multi-omic approach

3.2.1 Quality controls

The analysis of the secretome data was performed on three independent replicates of conditioned media, pooled from microfluidic channels every 2 days (Figure 11). By filtering out proteins solely quantified in one replicate, 4542 proteins were kept, the majority identified in 3 replicates (81%) and the others identified in only 2 replicates. First, consistency of data was addressed by correlating protein concentration of one replicate against the other two. The scatterplots show that replicates from the same sampling day are indeed highly correlated. Then, secretome data dimensionality was reduced via a principal

component analysis (PCA). It is a linear statistical procedure that allows for large datasets to be visualized. It results in a graph where all the features (e.g., secretome profile) of a sample are represented as a dot in a bidimensional space. The distance between dots (i.e., samples) correspond to how similar they are. In this peculiar case, on one hand we could observe that replicates grouped together according to their day of origin, and, on the other hand, that samples followed a reprogramming temporal trajectory (Figure 14). Altogether, these results confirmed that data was highly reproducible.

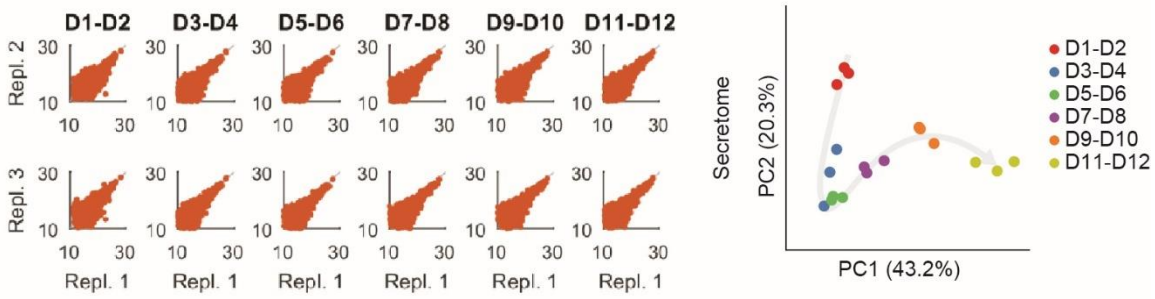

Figure 14: Secretome data consistency. Left: Visualization of proteomic data correlation between replicates. Each dot represents an identified protein. Log_2 relative quantification is shown on the axes (a.u.). Right: PCA plot proteomic data, shown as the distribution of the proteomic pattern for sampled conditioned medium over a 48-hour period.

For the sequencing data, cells were collected before the first transfection (D0), 3 days after transfection (D3) and then every 2 days (D5-D15). We generated sequencing libraries from independent captures for at least two replicates per time-point, collecting altogether more than 40,000 single-cell transcriptomes. Sequencing quality was evaluated by looking at the number of detected genes, number of reads per cells and percentage of reads associated with mitochondrial genes (Figure 15).

Indeed, if a cell shows a low number of genes and their associated counts, as well as high levels of mitochondrial genes expression, it might suggest that some ambient RNA has been mis-interpreted as a cell by the software. For this reason, we kept cells with more than 1,000 detected genes and less than 10% of mitochondrial gene counts (Figure 15). To avoid any bias due to the inconsistent number of cells per each sampling day, we also randomly subsampled the dataset to consist of 20,000 high-quality single-cell transcriptomes for a total of 12,932 features detected.

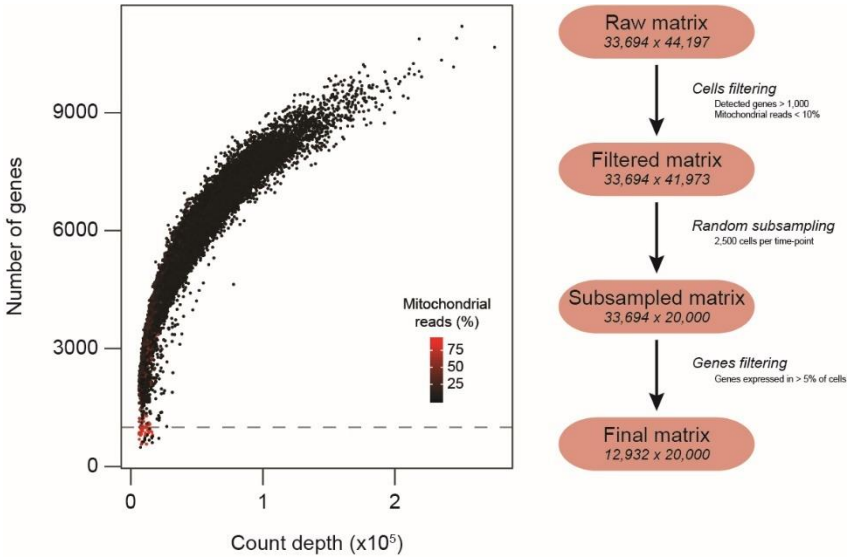

Figure 15: Sequencing data quality control and filtering. Left: Scatter plot representing the number of reads (x axis) over the number of detected genes (y axis) for each cell. Color gradient shows the percentage of reads associated with mitochondrial genes. The dotted line has been put at 1,000 detected genes, used for filtering. Right: Schematic representation of cells/genes filtering from raw data to the final dataset.

In a manner akin to our approach with secretome data, we also ensured a high degree of correlation among the replicates in the scRNA-seq dataset. However, since single-cell transcriptional data is sparse (i.e., contains a considerable number of missing values), dimensionality reduction cannot be achieved via linear models (e.g., PCA). Therefore, it was accomplished through a non-linear algorithm, specifically the force layout embedding (FLE – Figure 16).

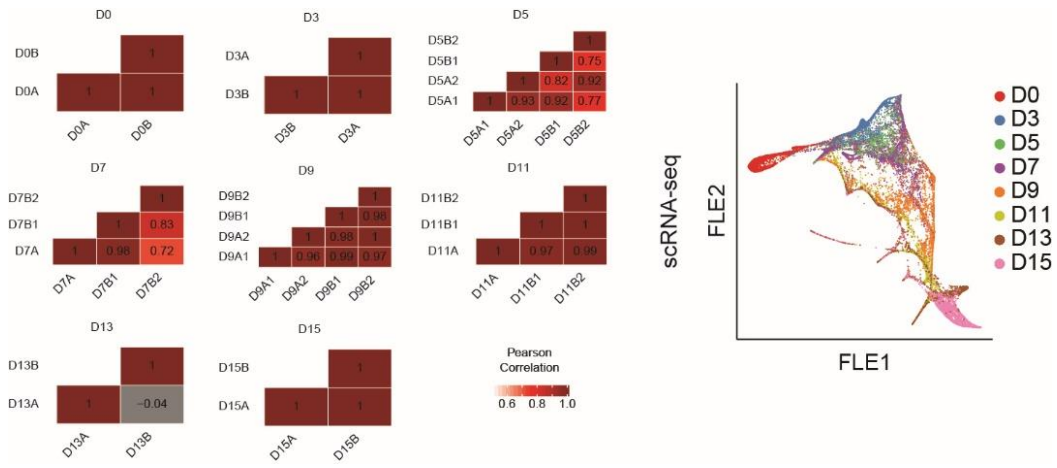

Figure 16: scRNA-seq data consistency. Left: Heatmaps of Pearson correlation coefficient for each replicate, divided by each time-point. Correlation has been evaluated by comparing the distribution of each replicate in the clusters identified in Figure 21. Right: FLE plot for scRNA-seq data. Sequencing data is shown as the distribution of transcriptional patterns for single cells across sampled time-points.

The resultant diagram visually represents the gene expression profile of each cell as a point within a Euclidean space, with cells being clustered based on their transcriptional similarities. Intriguingly, the graphical representation demonstrates a significant degree of uniformity within the fibroblast population on day 0 (D0), but greater diversity in cell placement as subsequent sampling days progress, reflecting their transcriptional variation over time.

3.2.2 Transcriptional waves contribute to the secretion of specific proteins

To test the hypothesis that both cellular and extracellular dynamics are interconnected, we compared the differential features of each dataset along the time. As expected, during the transition from D0 to D3, gene expression was the most influenced by the transfection of reprogramming factors, as evidenced by the high number of differentially expressed genes. From D3 on, transcriptional changes start to decrease until D7, where they reach the minimum magnitude. Finally, we observed at D7-D9 and D11-D13, two more transcriptional waves in line with the onset of developmental transitions and final acquisition of pluripotency. Notably, in between the two first transcriptional waves (from D5 to D7), we observed a great increase in the number of secreted proteins (Figure 17).

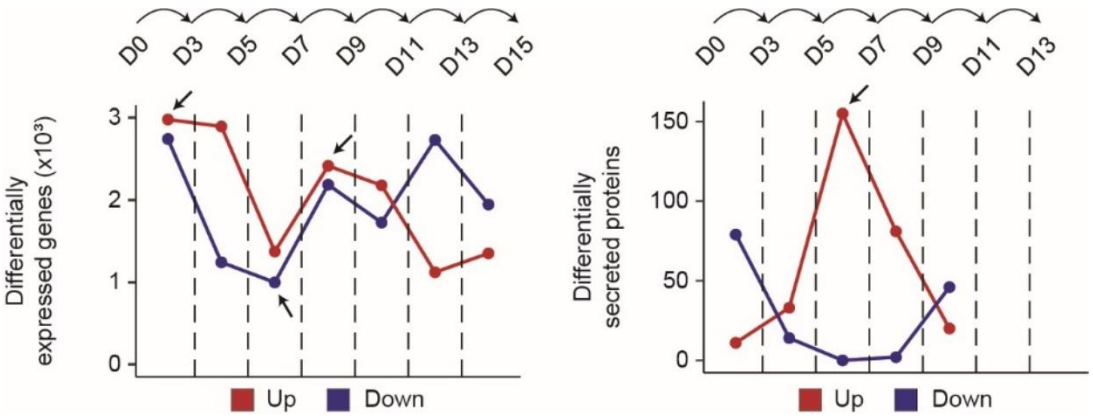

Figure 17: Comparative dynamics of features. Absolute number of differential features for each -omics data, both up- (Up - red) and down-regulated (Down - blue). Each value refers to the differential analysis between subsequent time-points. Peaks of deregulation are highlighted (arrows).

We have examined the temporal dynamics pertaining to both transcription and secretion of this set of proteins that peaked between D5 and D7 (Figure 18). The transient up-regulation of the genes encoding for these proteins peaked at D7, followed by their maximum level of secretion at D9. We hence reason that the initial massive changes in gene expression might induce the specification of a set of secreted molecules that becomes manifest in the medium at D7.

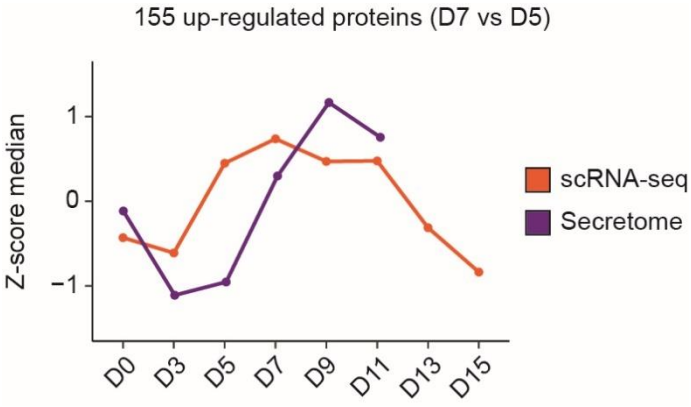

Figure 18: Definition of a relationship between transcription and secreted proteins. Median z-score of the 155 proteins found up-regulated from day 5 (D5) to day 7 (D7) in proteomic data. Trends have been evaluated along the time-course for both -omics data.

3.3 A rich extracellular signaling environment is shaped during human cell reprogramming

The massive number of secreted proteins at D5-D7 pointed us to investigate the quality of secreted proteins and cell population dynamics occurring in such a peculiar window of time.

3.3.1 Accumulation of embryonic extracellular matrix

To characterize the secreted proteins, we specifically selected 555 proteins known to be secreted. Besides proteins that were only up-regulated at the initial stages of reprogramming (D1-D2), we also got rid of intracellular proteins potentially released by dead cells. The identified categories were classified into two broad groups, one of them being extracellular matrix (ECM)-related functional annotation. Many ECM-related categories were highly significant, including ECM deposition, degradation, and remodeling, and both integrin- and non-integrin-mediated cell-ECM interactions (Figure 19). A previous RNAi screen also identified the critical role of cell adhesion in human reprogramming, highlighting the role of intercellular factors needed for filament assembly, branching, and disassembly[159].

In our data, we found an overall increasing trend of ECM-related protein accumulation, with different ECM components exhibiting distinct dynamics (Figure 19). These dynamic changes started already at days 3-4 (SPP1, COL4A1/2, SPARC), in some cases at days 5-6 (LAMC1), or even later (COL18A1). We wondered whether the observed global changes somehow resembled embryo development stages. We selected the ECM proteins in our data that were previously reported to be expressed at mRNA level at different stages of human embryo development[136]. The concentration dynamics of these proteins in this system showed the progressive establishment of an ECM that recapitulates the one deposited at the stage of

the late inner cell mass. In conclusion, our data support the idea that during reprogramming, not only fibroblasts are converted to a primed pluripotent phenotype, but also the extracellular context is shaped accordingly.

3.3.2 Dynamics of extrinsic regulatory signals

The other class of categories that we identified was related to soluble and regulatory signals. we narrowed the results to a selection of signaling pathways enriched within the Reactome database. Among receptor tyrosine kinase pathways, PDGF and WNT have already been shown to be implicated in embryo development and reprogramming[160,161]. we also identified the MET pathway as a link between cell-cell communication via soluble environment, and cell-ECM interaction via PTK2 (also known as FAK) adhesion. Moreover, the regulation of insulin-like growth factor (IGF) pathway through IGF binding proteins (IGFBPs) was significantly enriched, in line with previous studies[162] (Figure 20).

Looking at the temporal profiles of enriched signaling pathway proteins and ligands, we found a progressive accumulation of proteins that were previously shown to play a role in mouse cell-non-autonomous reprogramming regulation: some senescence-associated secreted proteins (SASP), such as CXCL1 (also known as Gro-α), CXCL8, CCL2, IL6[163]; YAP-target CCN1, also known as CYR61[164]; inflammatory cytokines, such as IL6/11/19, CSF1/2/3, LIF[69] (Figure 20). We conclude that secreted proteins follow precise dynamics during reprogramming and encompass several potential regulators of autocrine/paracrine signaling, including those involved in ECM-mediated and soluble communication.

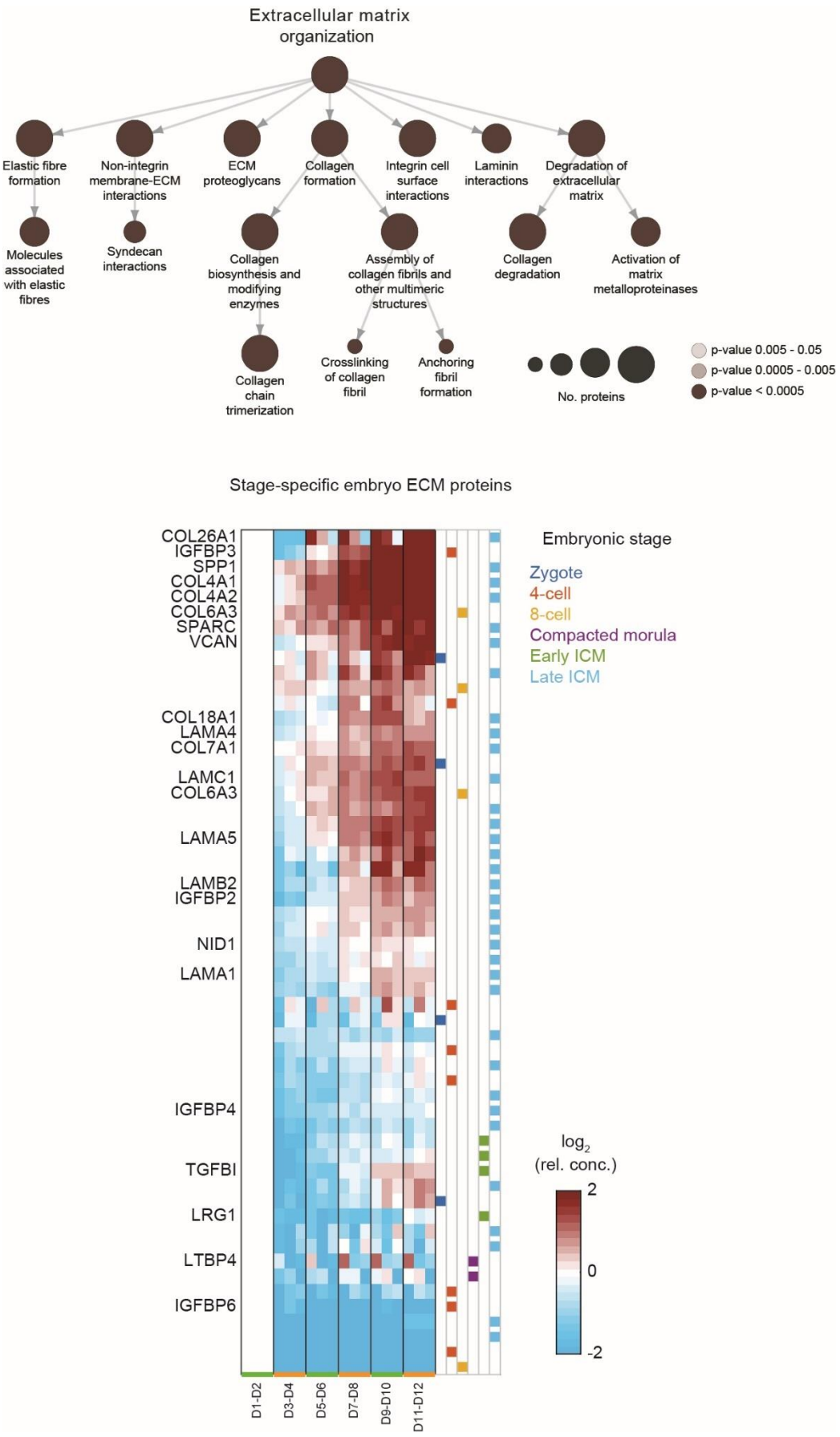

Figure 19: Accumulation of embryonic extracellular matrix. Top: ECM-related results from the enrichment analysis within the Reactome database of the 555 proteins identified as secreted. Edges connecting different categories reproduce Reactome hierarchy relationships. Bottom: Hierarchical clustering of proteins identified in this study and belonging to the core ECM components[137] at specific stages of embryo development[136].

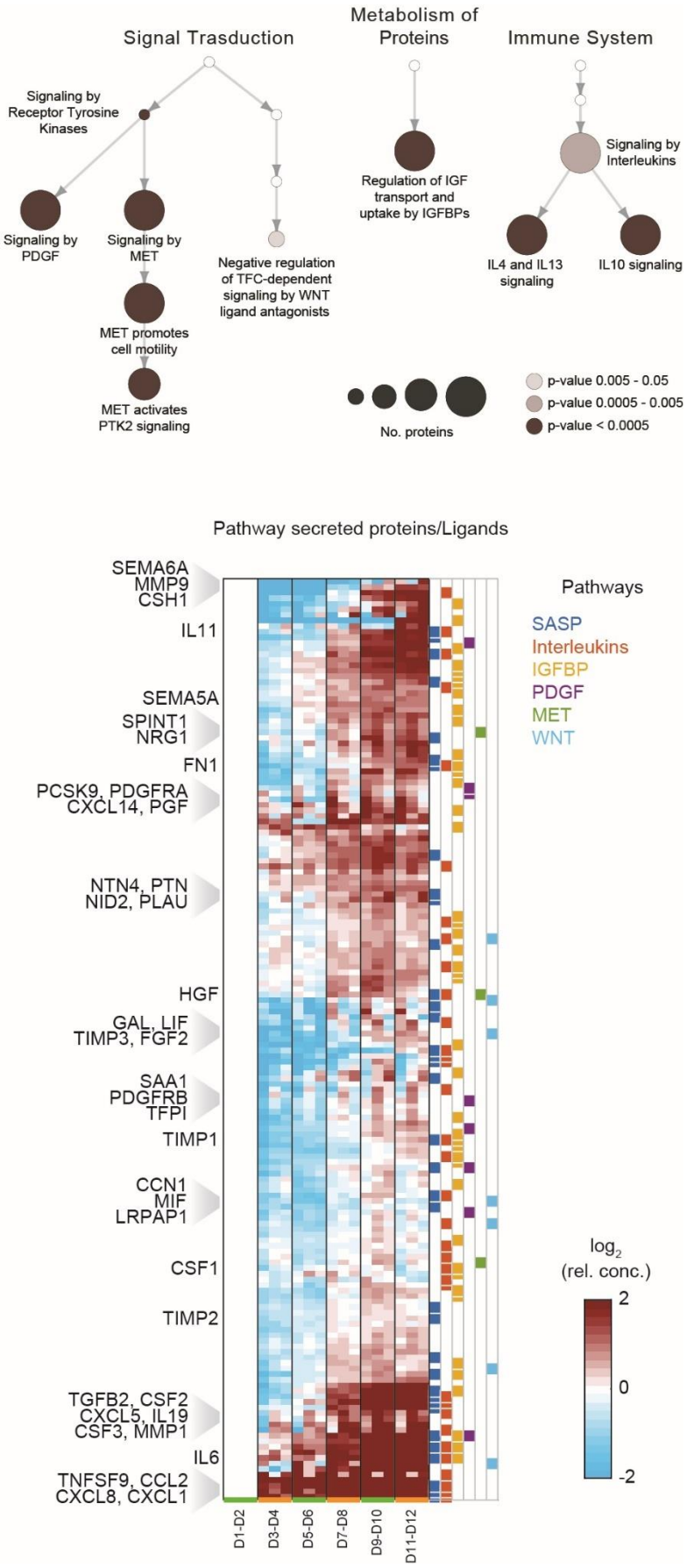

Figure 20: Dynamics of extrinsic regulatory signals. Top: Signaling-related results from the enrichment analysis within the Reactome database of the 555 proteins identified as secreted. Edges connecting different categories reproduce Reactome hierarchy relationships. Bottom: Hierarchical clustering of proteins identified in this study and belonging to signaling pathways.

3.4 Extracellular environment and cell heterogeneity are interconnected

In Figure 18, we have already speculated the potential existence of a connection between the transcriptome of cells and the way the environment undergoes remodeling. However, which cells are responsible for this contribution was still unclear. We hence leveraged the power of scRNA-seq in evaluating heterogeneity to tackle this issue.

3.4.1 Resolving cell population heterogeneity

We applied an unsupervised community detection algorithm[165] to our scRNA-seq data and we identified 12 cell clusters. We then took advantage of formerly defined reprogramming-associated gene signatures from my lab[64] to annotate them (Figure 21). 7 clusters showed high expression of somatic genes ("Somatic-Related" clusters, SR), whereas 4 clusters were highly enriched by the developmental signature ("Developmental-Related" clusters, DR). Finally, a residual cluster was not enriched by either of those signatures and it was characterized by a lower number of detected genes and total UMI counts, thus we named it "NA" and excluded it from further analyses.

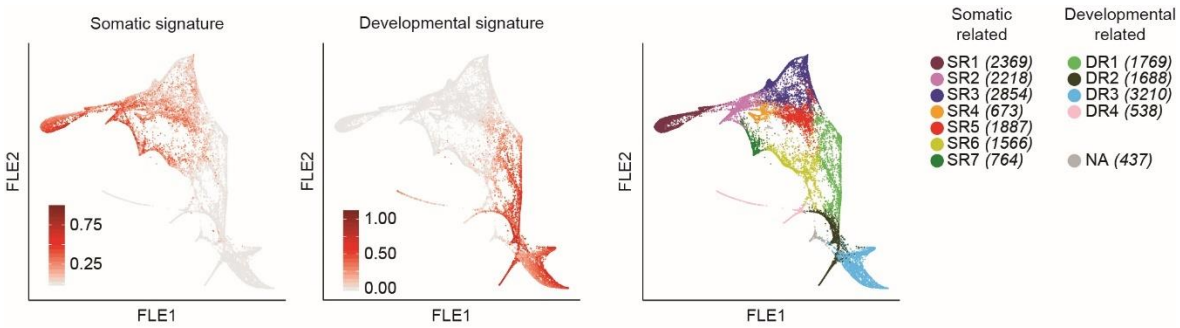

Figure 21: Cell clustering and annotation. Left and middle: Somatic and Developmental signatures enrichment scores shown along the FLE map. Right: FLE map showing the distribution of cells across identified clusters.

As expected, SR clusters included non-transfected fibroblasts (SR1) and cells captured at earlier days (SR2-5), while DR clusters were enriched by cells collected at later time points (from D9 to D15) and highly cycling (Figure 22). However, also SR6 and SR7 displayed more than 97% of cells from day 11 and were characterized by low but detectable expression of embryonic genes (e.g., POU5F1, LEFTY2): nevertheless, they were negative for NANOG, indicating reshaping of fibroblast identity but at the same time inefficient acquisition of pluripotency. Furthermore, these cells were in the G0/G1 phase of the cell cycle, thus confirming their somatic nature and suggesting peculiar identity in the reprogramming timeline (Figure 22). Despite their developmental features, DR4 cells also

did not express NANOG, while showing high and very specific transcriptional levels of mesendoderm genes (e.g., CER1, EOMES), suggesting a possible similarity with a differentiating stage. DR clusters exhibited higher expression of pluripotent and embryonic-related signatures[136], thus they appear to contain the productively reprogramming cells.

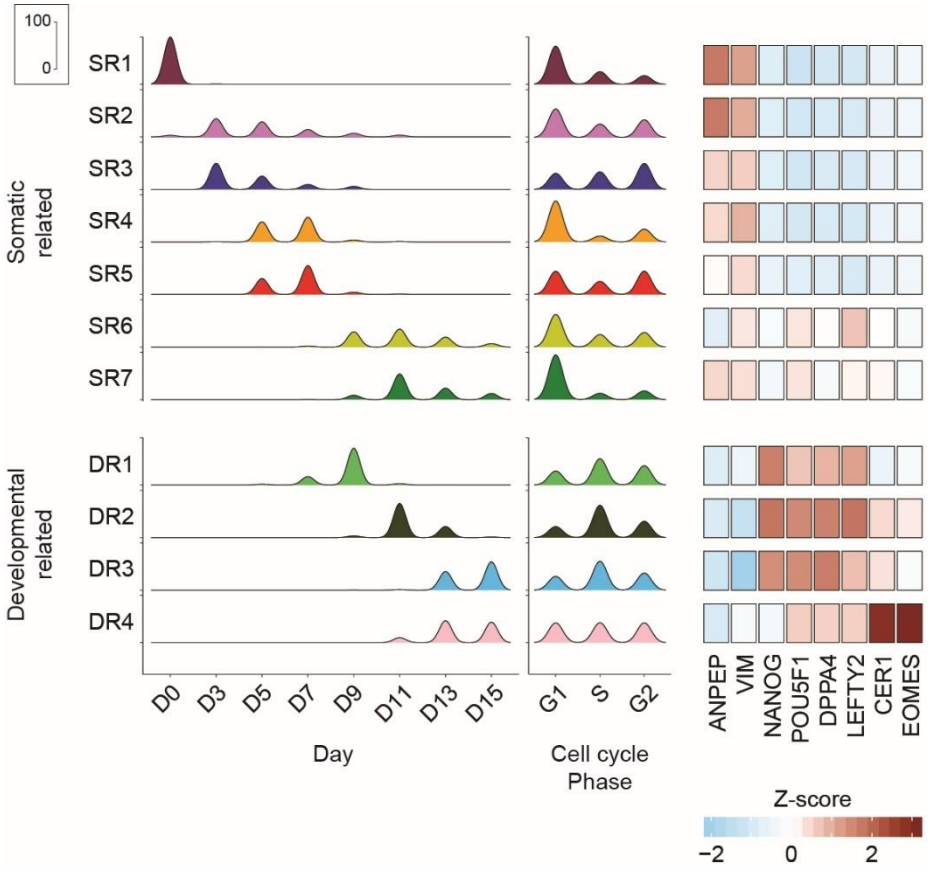

Figure 22: Characterization of cell heterogeneity. Left: Time-points and cell-cycle phase distribution for each cluster. Right: Heatmap of Z-scored normalized counts, averaged by clusters, for key reprogramming related genes (right). NA cluster not shown.

3.4.2 SR clusters contribute to the establishment of a specific environment

Although we disentangled their features, the role of the SR clusters was still less clear. However, when looking at the transcription levels of proteins highly concentrated in the environment (Figures 19 and 20), we observed that most of them were heavily transcribed by these clusters (Figure 23). We confirmed this relationship by performing Gene Set Enrichment Analysis (GSEA) using the secreted proteins previously identified and some gene signatures that were found enriched in the proteomic analysis (Figure 24). As speculated, the secreted proteins detected by mass spectrometry appear to be transcribed by the cells in the SR clusters, except for SR3 that might not be involved in the secretory

phenotype. These results highlight the presence of an unproductive somatic fate, whose role is to express and secrete those factors that we found to be shaping the extracellular environment during reprogramming and that have been found to characterize later stages of embryonic development.

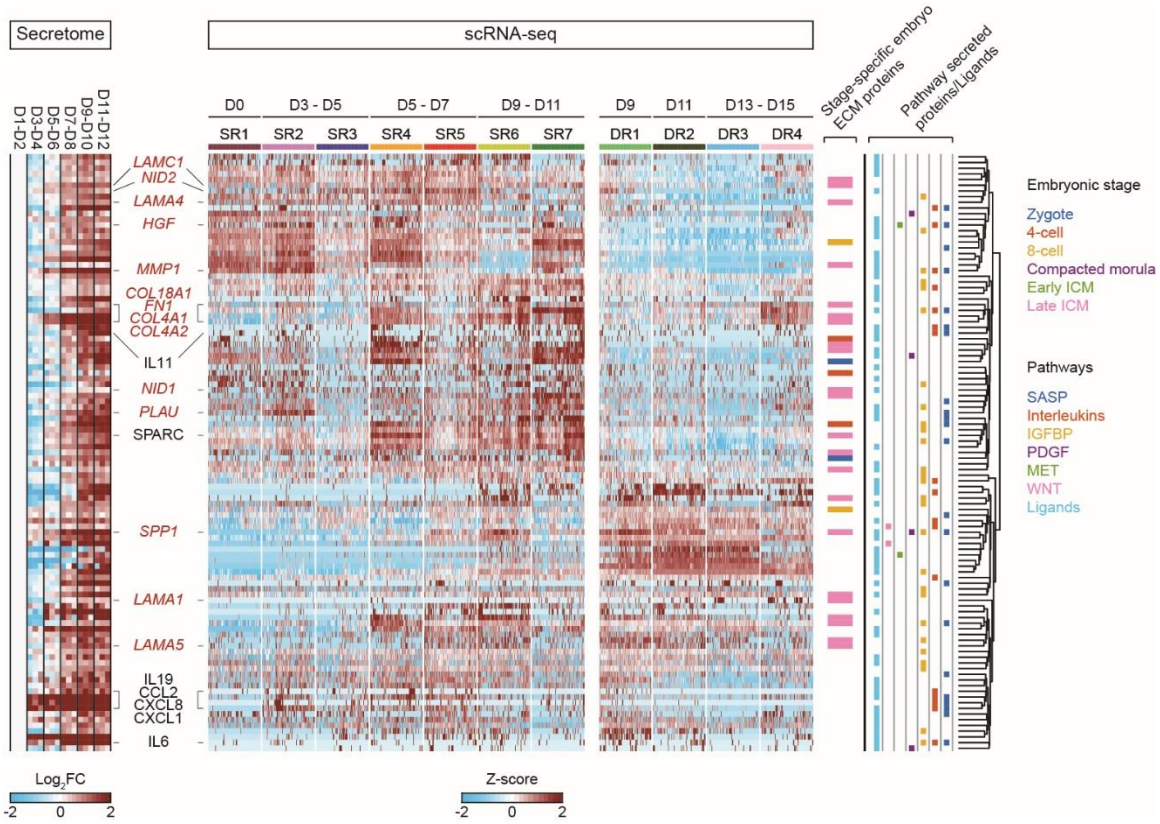

Figure 23: Contribution of SR clusters to the secretome. Heatmaps of highly dynamic proteins from the secretome analysis. The colors display log2 fold change protein concentration with respect to D1-D2 (Secretome - left) and Z-scored log2 counts per million (scRNA-seq - right). Hierarchical clustering was performed on scRNA-seq data according to each separate cluster of cells. Proteins involved in primitive node formation are highlighted (red names).

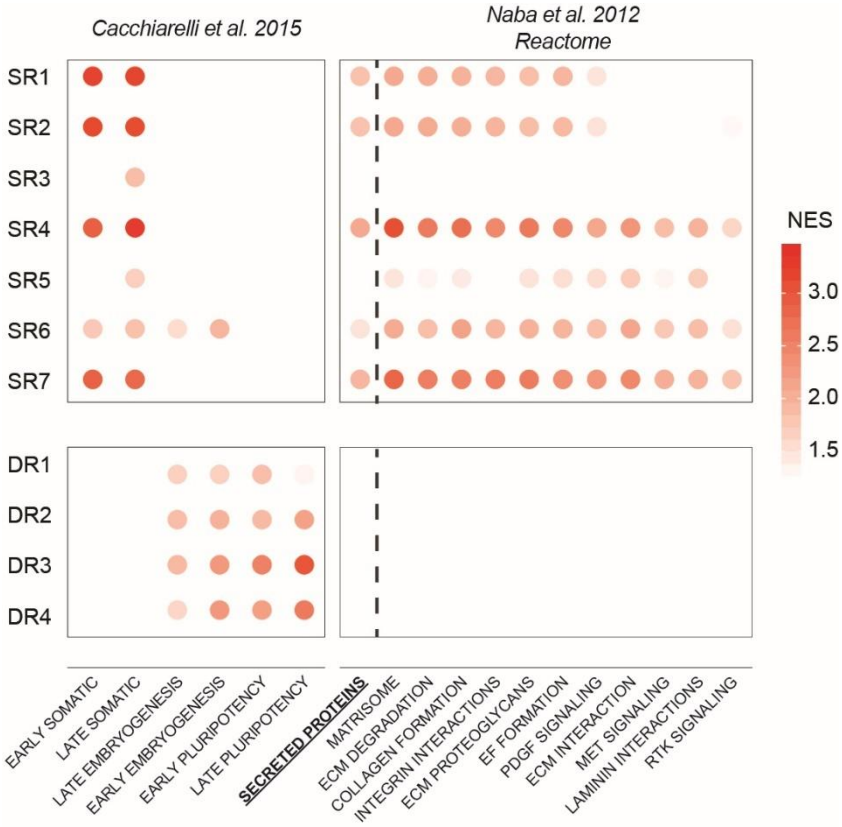

Figure 24: SR clusters profile is enriched by secreted pathways. GSEA results for each cluster. Only significant results are shown. The gene set made of the secreted proteins found in this work is written in bold. NES, Normalized enrichment score.

3.5 Trajectory inference reveals different fates

Among all the gene sets analyzed, Matrisome[137] and Late pluripotency[64] associated genes were found to best describe the phenotype of D13-15 endpoints. Therefore, we decided to computationally investigate the routes linking such states to the somatic start-point by applying Waddington Optimal Transport (WOT)[66]. The algorithm applies the Mass Optimal Transport theory to the gene expression space to infer, for each cell in a given sample, the most probable ascending and descending cells in the previous and following timepoints. By applying this tool to selected cells at the furthest time-point (D15), we were able to compute the route from D15 to D0. Results showed a common path until day 5 (D5), after which cells started to exhibit different trajectories. We validated these findings through an unsupervised pseudotime-based approach using Monocle3[142,166], which not only confirmed the bifurcation at day 7 (D7) leading to endpoints inside SR7 matrisomal and DR3 pluripotent clusters, but also introduced two additional outcomes inside DR4 and SR2, respectively (Figure 25).

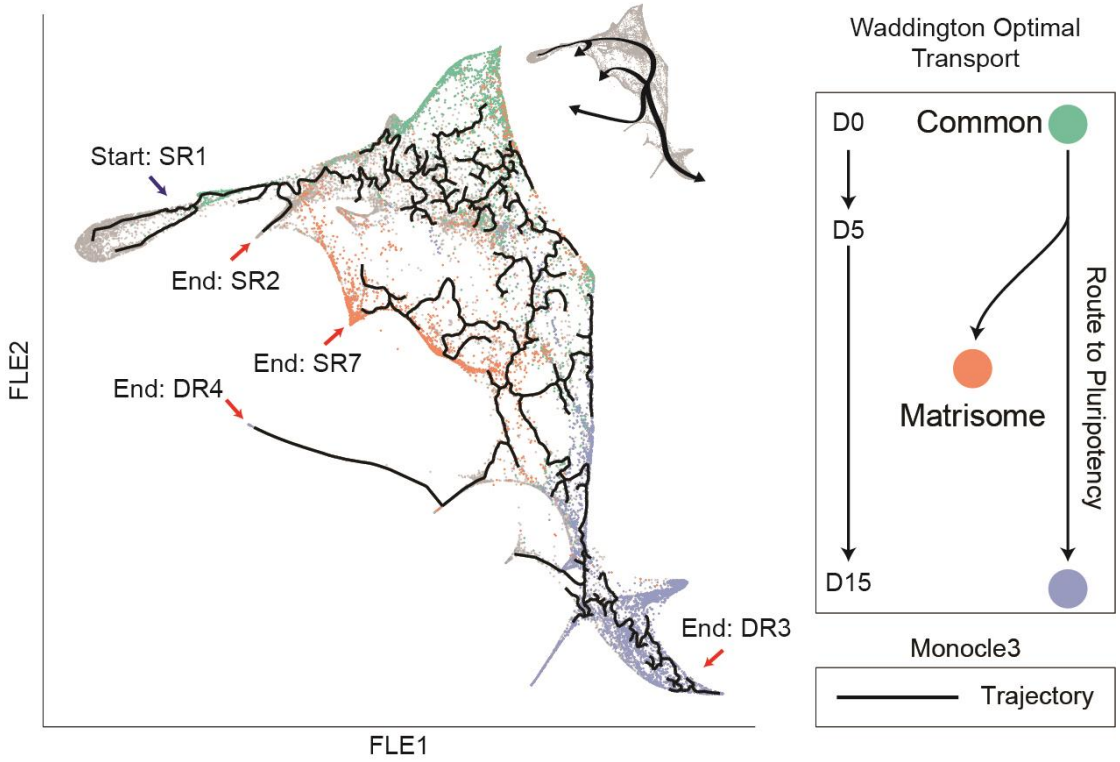

Figure 25: Trajectory inference analysis. Monocle3 (black line) and WOT (colored dots) trajectory inferences are displayed on the FLE graph. Arrows point to the starting point (blue) and 4 end points (red) of the inferred trajectories. A representative scheme of the trajectories is shown on the top-right.

3.5.1 An early reprogramming fate is characterized by a secretory phenotype

While the mesendodermal nature of DR4 was previously assessed (Paragraph 3.4.1), we focused on the characterization of SR2. GSEA using common pathways (Methods) revealed the enrichment for terms related to signaling molecules, therefore, we hypothesized that this cluster might be implicated in the secretion of the ligands detected in the medium. Indeed, most of them were significantly enriched, with SASP having the highest enrichment score (Figure 26). We found SASP genes are highly expressed and specific of this cluster, such as cytokines (CXCL1, IL1B, CXCL8), metalloproteases (MMP1, MMP3), HGF and its activators, PLAU and PLAUR. Notably, almost all of them were detected by LC-MS/MS with some (CXCL1, CXCL8, CCL2, SPP1, PLAU) being the first to be accumulated in the medium (Figure 27).

In conclusion, we were able to define human somatic reprogramming as a process consisting of two major outcomes, matrisomal and pluripotent, deriving from the same starting cells which bifurcate around day 7 (D7). Moreover, among matrisomal somatic cells, we identified and characterized an early sub-population of cells which contributes to the expression and secretion of SASP-related signaling molecules.

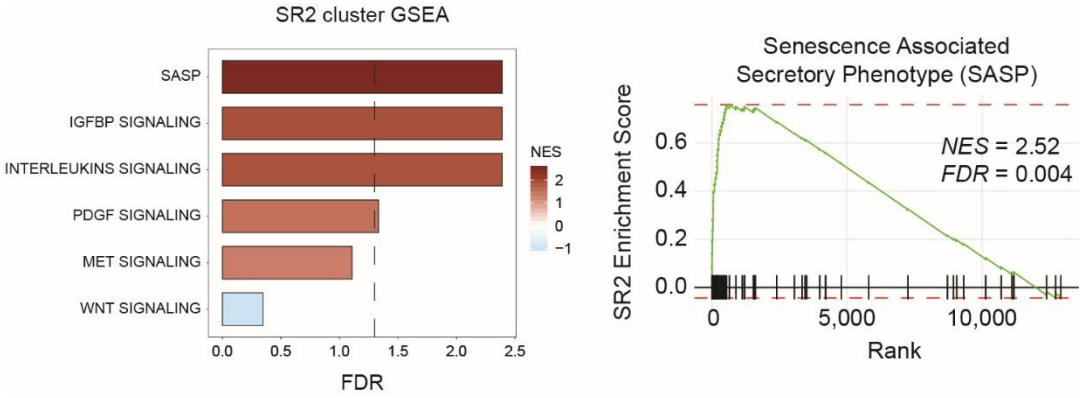

Figure 26: SR2 cluster shows a secretory phenotype. Left: GSEA has been performed on SR2 cluster using signaling-related genesets used in Fig. 17. The results are shown as a barplot, displaying FDR (x axis) and NES (colors). Right: Enrichment Score graph relative to the GSEA of SR2 cluster for senescence-associated secreted proteins geneset (SASP). Black lines on the x axis represent a match between the ranked list and the geneset analyzed. NES, Normalized enrichment score. FDR, False Discovery Rate.

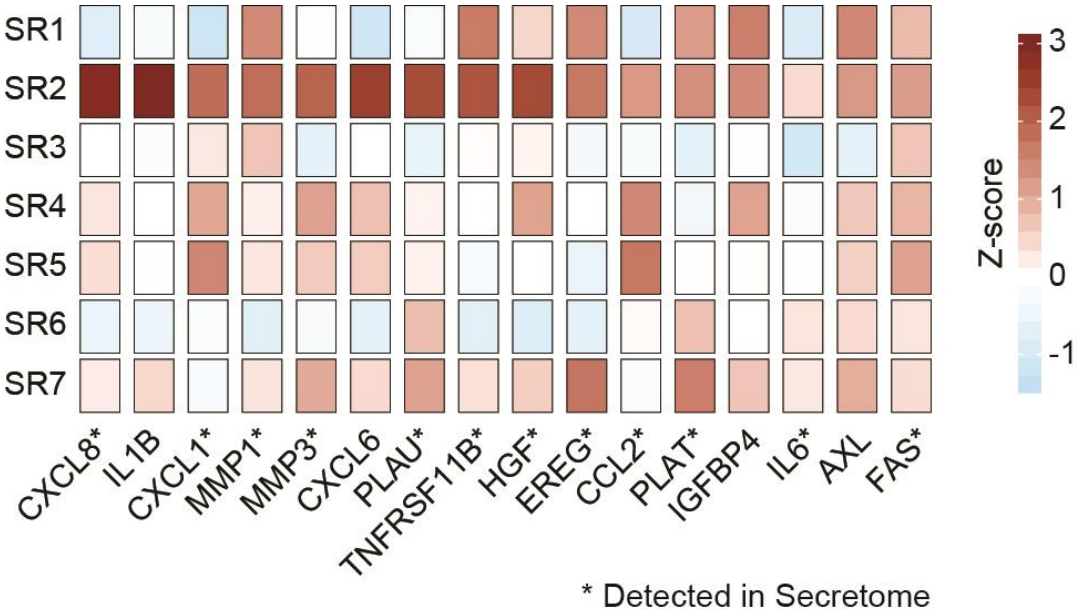

Figure 27: SASP genes are specific to SR2 cluster. SR2 cluster marker genes relative expression, shown in a heatmap of Z-scored normalized counts, averaged by clusters. Genes with (*) have been detected in secretome analysis.

3.6 Reprogramming fates interact through different ligand-receptor pairs

To rationally understand whether somatic subpopulations arising during reprogramming are actively involved in the population cross-talk with productive reprogramming intermediates, we developed a ligand-receptor interaction analysis from the cells laying on the somatic trajectory towards the reprogramming ones (Figure 28). Using the previously identified

secreted proteins (Fig. 1) that fall in the list of experimentally validated ligand-receptor couples[138], we restricted the number of putative interactors involved in subpopulation crosstalk to a set of 82 pairs (Figure 28). We were able to identify a standardized interaction score (sIS) by leveraging the gene expression trends of ligands along the matrisome route and of receptors along the path to pluripotency (Methods).

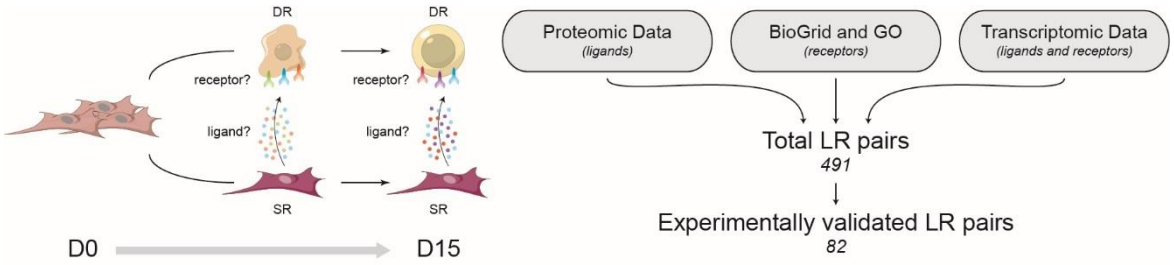

Figure 28: Interaction Score analysis. Left: Schematic representation of ligand-receptor interactions hypothesized during reprogramming. Fibroblasts (D0, left) develop two fates: a somatic secretory phenotype (bottom) and induced pluripotency (top). Black arrows show the directionality of the examined interaction. Right: Schematic representation of ligand-receptor pairs selection for interaction score analyses, as described in Methods.

The results showed that almost every ligand-receptor pair had a significant sIS in at least one time-point (Figure 29). Moreover, when looking at the couples with the greatest scores, we observed many ligands involved in signaling cascades which are already known to be associated with pluripotency maintenance, such as Wnt, Tgfβ and Inhb signalling[160,167,168]. Among these interactors, 8 ligands were related to SASP, of these 4 were soluble and highly dynamic in both transcriptomic and proteomic data: SPP1, INHBA, NRG1 and HGF. As INHBA is a known pluripotency regulator[168], and SPP1 is the major HGF-regulated gene[169], we focused our analyses on HGF and NRG1.

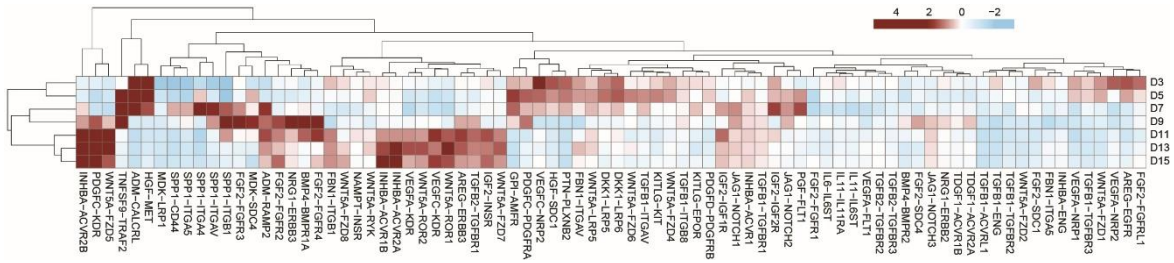

Figure 29: Interaction Score results. Heatmap of z-scored standardized interaction scores for all the ligand-receptor pairs analyzed.

3.6.1 HGF-MET

The HGF-MET interaction occurred at early time-points of the reprogramming (Figure 30) with HGF expressed by cluster SR2 and SR5 and its receptor MET expressed by cluster DR1. Both HGF and MET were highly expressed in the early intermediate stages and decreased in the later time points, suggesting a role in the reprogramming intermediates.

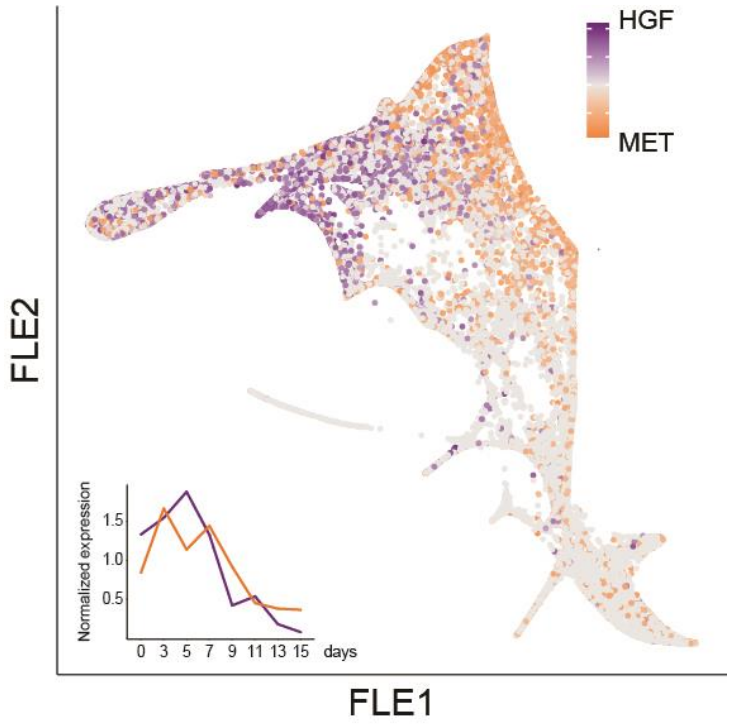

Figure 30: HGF-MET dynamics in microfluidic system. HGF and MET gene expression profiles (log_2 CPM) displayed on the FLE map as fold change relative to HGF and averaged across the time course (bottom-left).

Thus, we explored whether the same HGF-MET dynamics was present in a conventional (i.e., Petri dish) human reprogramming approach[170] and not strictly related to the microfluidic environment. scRNA-seq data exploration, using authors-defined clusters, showed that the cluster noRepro1, enriched for SR signatures, expressed high levels of HGF, whereas MET expression was observed in the mixed intermediate cluster, overlooked by the authors (Figure 31). Remarkably, the analysis of RNAseq data from reprogramming of secondary human fibroblasts cultured on mouse embryonic fibroblast feeder (MEF)[64], showed the expression of HGF only from MEFs while MET was upregulated in human cells undergoing reprogramming at day 8 (OSKM - Figure 31). Therefore, we performed reprogramming experiments with depletion or addition of MEFs and observed a drastic reduction in the ability of generating pluripotent colonies when cultured in absence of feeder cells (Figure 31), suggesting a pivotal role of HGF-MET interaction in sustaining pluripotency. These results showed a common behavior of HGF vs MET expression in the

early phase of the reprogramming, being expressed by matrisome producing/supporting cells and reprogramming intermediates respectively, regardless of reprogramming approach and culture system.

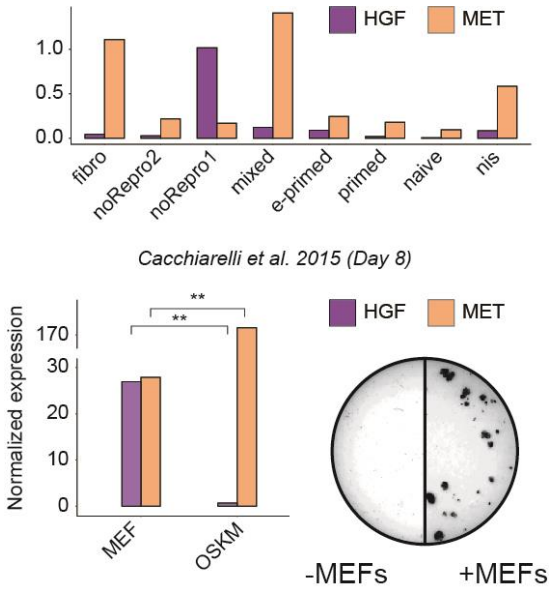

Figure 31: HGF-MET dynamics in other systems. HGF and MET gene expression profiles shown in Liu et al., 2020[170], as averaged across their identified clusters (top) and in Cacchiarelli et al., 2015[64], shown as mouse and human mean normalized expression at sampling day 8 (bottom-left; ** BH-adjusted p-value <0.01). Bottom-right: Representative pictures of HiF-T DOX secondary reprogramming performed with or without depletion of MEFs in standard 12-well plates, assessed by immunostaining of TRA-1-60.

3.6.2 NRG1-ERBB3

On the other hand, the NRG1-ERBB3 interaction showed higher sIS between clusters along the same developmental trajectory in a sequential fashion: NRG1 is expressed by DR clusters at earlier stages (until D9), while its receptor, Erb-B2 Receptor Tyrosine Kinase 3 (ERBB3), is expressed by late DR clusters (starting from D7) (Figure 32). The same information can be retrieved from Liu et al., 2020[170] and Cacchiarelli et al., 2015[64], observing the sequential expression of NRG1 then ERBB3 only along the reprogramming intermediates, with NRG1 decreasing halfway during reprogramming route, ERBB3 increasing from halfway, and a central timeframe of co-presence (Figure 33). Therefore, as NRG1-ERBB3 expression occurs only along the reprogramming trajectory, we did not get significant results when comparing MEFs versus human reprogramming intermediates from our human secondary system[64].

Altogether, these findings suggest a crosstalk between cell subpopulations, with an active role of non-pluripotent cells in supporting the route of other cells to pluripotency. We

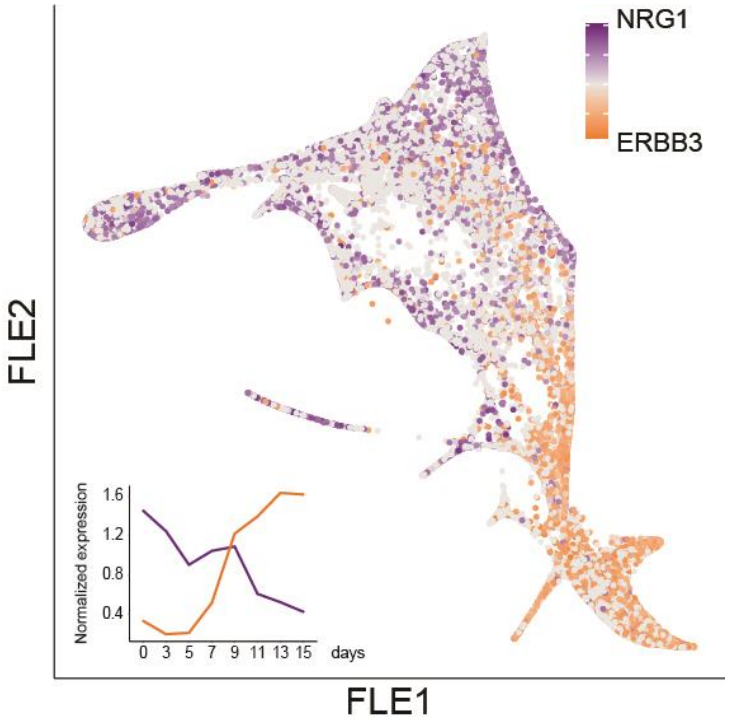

Figure 32: NGR1-ERBB3 dynamics in microfluidic system. NGR1 and ERBB3 gene expression profiles (log_2 CPM) displayed on the FLE map as fold change relative to NRG1 and averaged across the time course (bottom-left).

demonstrated that such non-pluripotent cells can be part of the same (i.e. NRG1 and ERBB3 both expressed during DR trajectory to pluripotency) or different trajectories (i.e. HGF ligand expressed by SR trajectory towards matrisome vs MET receptor expressed by DR trajectory towards pluripotency).

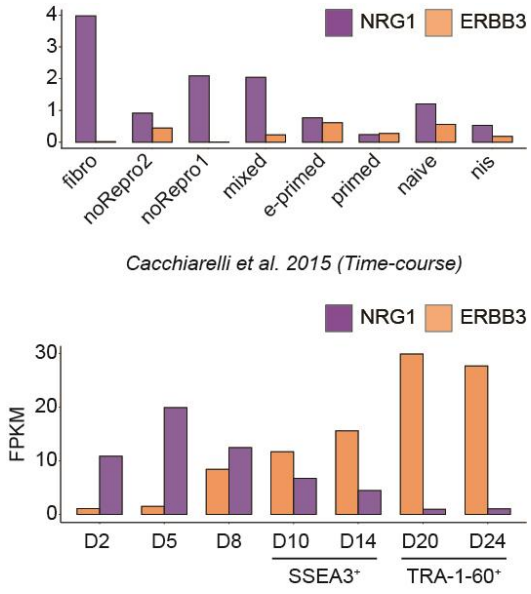

Figure 33: NRG1-ERBB3 dynamics in other systems. NRG1 and ERBB3 gene expression profiles shown in Liu et al., 2020[170], as averaged across their identified clusters (top) and in Cacchiarelli et al., 2015[64], shown as mean FPKM across the time-course (bottom).

3.7 HGF-MET crosstalk functionally sustains the acquisition of pluripotency through STAT3

Considering the results from the ligand-receptor analyses, we wondered whether the HGF-MET interaction has a functional role in the progression of intermediate states towards pluripotency. HGF is a growth factor involved in many cell functions and it is mostly secreted by mesenchymal cells, while acting on epithelial ones[171]. In our reprogramming, it is biologically active as its activator complex PLAU/PLAUR was also found in the secreted medium (Figure 23). On the other hand, MET is a tyrosine kinase receptor activated by its ligand HGF. This binding induces MET catalytic activity and results in downstream initiation of multiple pathways, including STAT3 direct phosphorylation or via Janus kinase 1 (JAK1 – Figure 34). This activation axis is shared with other two ligands (i.e., LIF and IL6), known to be involved in murine pluripotency[68,172]. However, their gene expression pattern cannot justify their action in this context, with IL6R and LIF not being expressed in scRNA-seq.

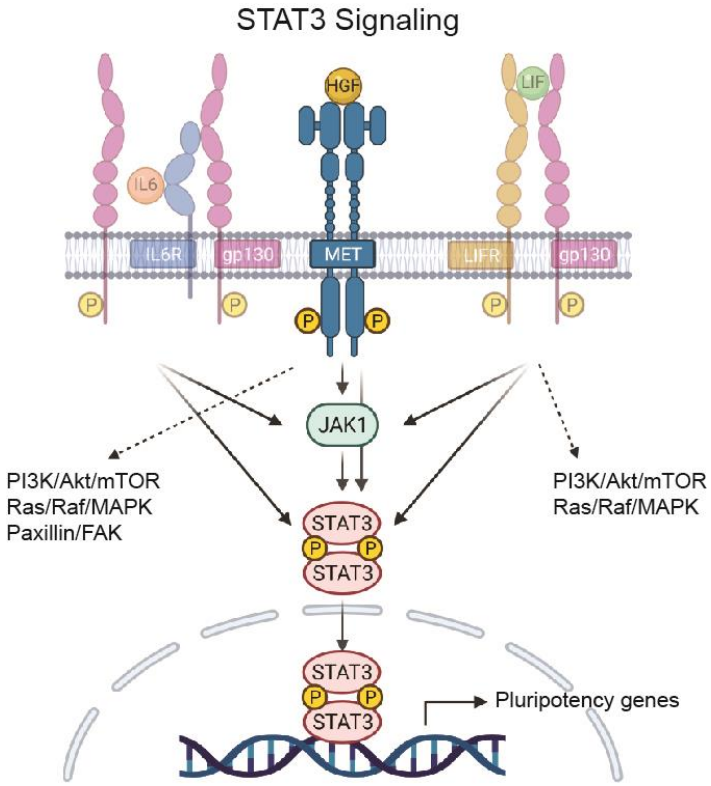

Figure 34: HGF/c-MET/STAT3 axis. A schematic representation of HGF/c-MET/STAT3 signaling pathway (Created with BioRender.com).

3.7.1 STAT3 pathway is active in reprogramming cells

To test the STAT3 pathway involvement in our reprogramming setup, we investigated its activation throughout the reprogramming process in microfluidics. First, HGF and MET were differentially expressed by SR (higher HGF) and DR (higher MET) clusters and came up as early interactors in a cluster-based interaction analysis. Furthermore, STAT3 nuclear target transcriptional enrichment[144] revealed their activation from day 5, along the reprogramming route (Figure 35), in agreement with MET signalling activity. Finally, at the protein level, we observed STAT3 nuclear localization (indicative of STAT3 activation) during intermediate days (D4, D7) and at the end of the process (D12 – Figure 35).

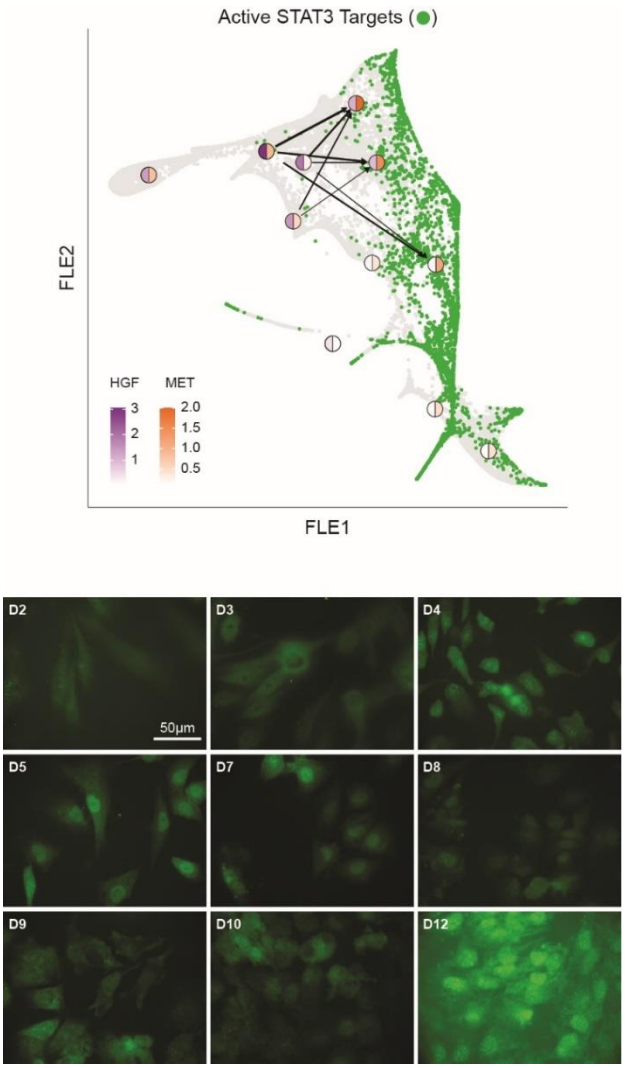

Figure 35: STAT3 nuclear activation. Top: In the FLE graph, green dots represent cells with positive enrichment scores for STAT3 target genes (Methods). Bigger circles summarize averaged HGF (left) and MET (right) gene expression in identified clusters. Significant inter-cluster HGF-MET interactions are displayed (arrows). Arrow thickness relates to the strength of the interaction. Bottom: Time course of STAT3 activation during the microfluidic reprogramming process. Representative images showing a first wave around day 4 to day 7, and then a second wave at the end of the process, when it is active just in the hiPSC colonies.

To give further evidence, we then investigated the localization of MET and STAT3 at day 6. We found that cells of smaller size (undergoing the mesenchymal-to-epithelial transition) show the highest intensity of both c-MET and nuclear STAT3 (Figure 36).

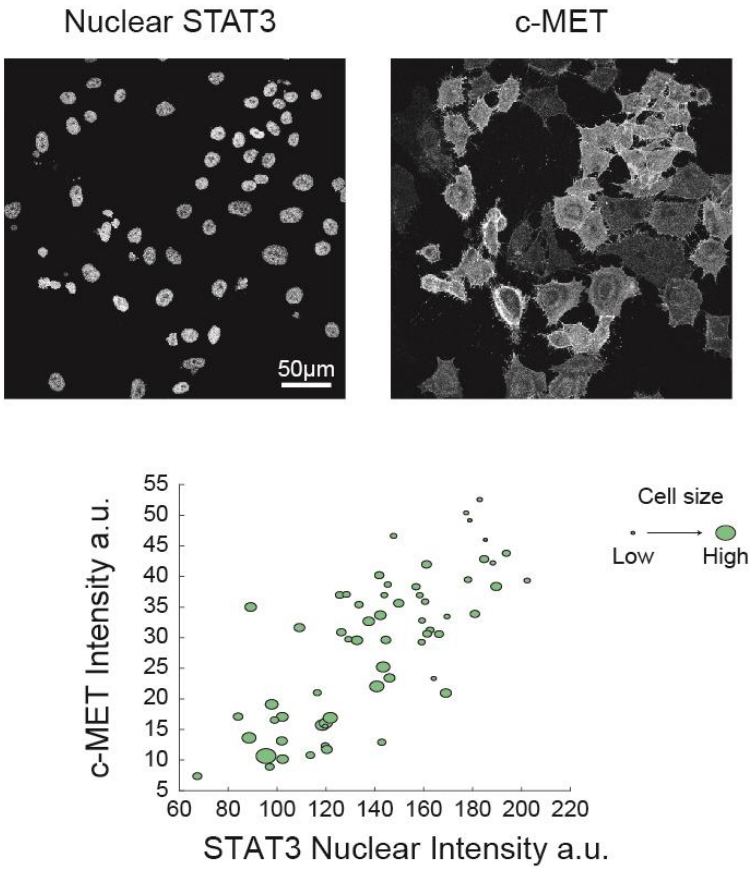

Figure 36: MET co-localizes with nuclear STAT3. Top: Representative images of expression of nuclear STAT3 and c-MET during reprogramming performed in microfluidics at day 6. Bottom: Correlation between the expression intensity of nuclear STAT3, c-MET, and cell size obtained from experimental data shown on top. Data from n = 61 cells (n = 3 independent experiments).

3.7.2 Perturbation of STAT3 pathway components affect the efficiency of reprogramming

Once the involvement of STAT3 was established, we separately inhibited two kinases along the STAT3 axis, MET and JAK1, using small molecules and assessed reprogramming efficiency by immunostaining analysis of NANOG at day 12. Consistent with our hypothesis, we observed a significant loss of reprogramming efficiency upon inhibition of STAT3. These data were strengthened by a direct knock-down of STAT3 mRNA using specific siRNA, that efficiently reduced reprogramming efficiency at day 12 (Figure 37).

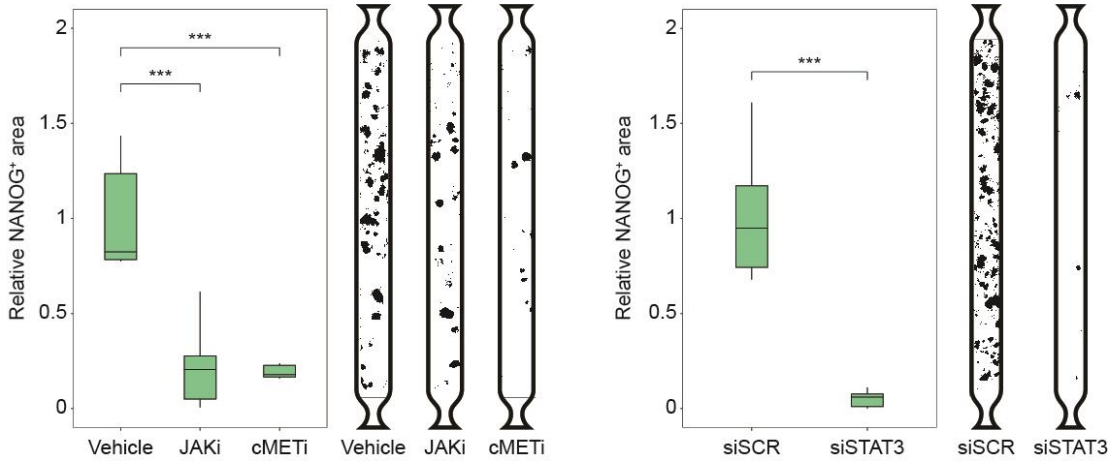

Figure 37: STAT3 inhibition impairs reprogramming. Left: Reprogramming efficiency in microfluidics measured as the relative area occupied by NANOG+ colonies in cells upon inhibition of c-Met and JAK1 kinases using small molecules at day 12, compared to the ones treated with the vehicle (n = 6 for vehicle, n = 12 for JAKi and n = 7 for c-METi); ANOVA followed by two-sided Dunnett's multiple comparisons test was used to assess differences among the conditions (JAKi – *** FDR = 0.0001; cMETi – *** FDR = 0.0001). Representative quantification pictures in microfluidic channels assessed by immunostaining of NANOG are shown. Right: Reprogramming efficiency in microfluidics upon knock-down of STAT3 using siRNAs at day 12 (n = 8 for scramble siRNA, n = 11 for siSTAT3); two-sided unpaired t-test was used to assess differences among the conditions (***P < 0.0001). Representative quantification pictures in microfluidic channels assessed by immunostaining of NANOG are shown.

Lastly, we tested whether the addition of signaling molecules was capable of further improving reprogramming yield in conventional culture systems that are otherwise far less efficient than microfluidic systems. For this purpose, we selected molecules that were found dynamic in the secretome analysis or involved in cell-cell interactions (e.g. HGF, IL6 and NRG1). In conventional culture (i.e., Petri dishes), we saw a significant increase of about 2-fold in reprogramming efficiency in terms of relative TRA-1-60+/NANOG+ area when medium was supplemented with either HGF, IL6 and its soluble receptor (sIL6R) to activate STAT3 signaling, and NRG1 throughout the reprogramming process (Figure 38). Consistent with the idea that multiple signals are involved in the first phase of reprogramming and the second phase of hiPSCs stabilization, secretome and single-cell RNA sequencing data showed more accumulation of HGF and IL6 in early phases of the reprogramming process, while NRG1 came out at later stages. To mimic this timing, we added HGF alone or with IL6/sIL6R in the first half, and NRG1 in the second half. This resulted in a further increase in the reprogramming efficiency up to three folds (Figure 38). However, when supplementing the medium with HGF, IL-6, sIL-6R and NRG1 together, we were able to reach the highest efficiency (i.e. 5-fold over controls), thus suggesting that the combination of specific signalling pathways further boosts hiPSCs formation (Figure 38).

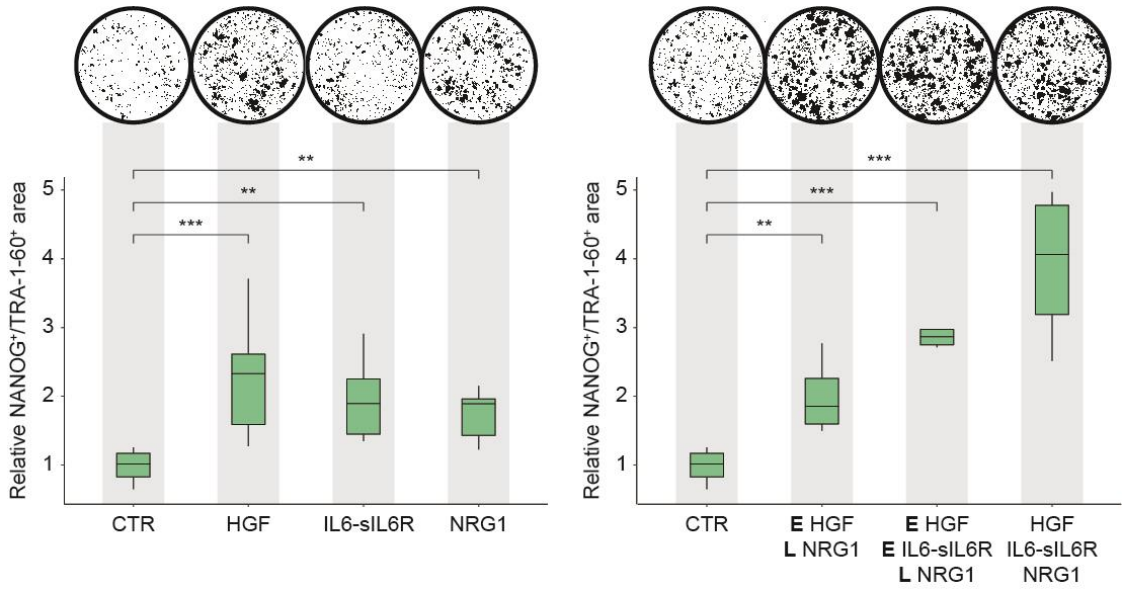

Figure 38: Exogenous signals improve reprogramming. Left: Reprogramming efficiency in standard 24-well plates upon addition of HGF, IL-6 and soluble IL6 receptor (sIL6R), or NRG1 at day 9 (n = 14 for control, n = 19 for HGF, n = 5 for IL6 + sIL6R, n = 16 for NRG1); ANOVA followed by two-sided Dunnett's multiple comparisons test was used to assess differences among the conditions (HGF - *** FDR = 0.0001; IL6 + sILR – **FDR = 0.0083; NRG1 – ** FDR = 0.0021). Representative quantification pictures in standard 24-well plates assessed by immunostaining of NANOG and TRA-1-60 are shown. Right: Reprogramming efficiency in standard 24-well plates upon temporally modulate addition of HGF, IL6 and soluble IL6 receptor (sIL6R), and NRG1 at day 9 (n = 14 for control, n = 6 for HGF in the early phase and NRG1 in the late phase, n = 4 for HGF + IL6 + sIL6R in the early phase and NRG1 in the late phase, n = 4 for HGF + IL6 and sIL6R + NRG1 for the entire process); ANOVA followed by two-sided Dunnett's multiple comparisons test was used to assess differences among the conditions (E HGF + L NRG1 - ** FDR = 0.0038; E HGF/IL6 + sILR + L NRG1 – *** FDR = 0.0001; ALL – *** FDR = 0.0001). Representative quantification pictures in standard 24-well plates assessed by immunostaining of NANOG and TRA-1-60 are shown.

3.8 Future perspectives

3.8.1 Multiome approach to dissect the regulatory logic behind different fates

One of the questions arising from what has been done so far concerns the origin of the unproductive fate sustaining the reprogramming process. We hypothesized that there might be some epigenetic differences between cells that can efficiently respond to stimuli that induce pluripotency, and those that cannot. Whilst transcriptional profiling enabled us to characterize these populations, it fails in depicting the causes that generate this phenomenon. Therefore, we have profiled a pool of cells from all the time-points from this work using a Multiomic approach (Methods). It consists of profiling the transcriptome and the chromatin accessibility coming from the same exact cell. The workflow of this analysis is as follows:

1. Combine the gene expression profile from Multiome with the previously generated scRNA-seq data from this work.
2. Use the combined data to transcriptionally characterize the cells from the Multiome (Figure 39).
3. Correlate the transcriptional profile of selected genes with the accessibility of their genomic loci.
4. Create Gene Regulatory Networks (GRNs) able to describe what differs one state from the other.

Preliminary results encompass the first three steps and demonstrate the validity of the approach. Further analysis will cover the last step and hopefully reveal new interesting results.

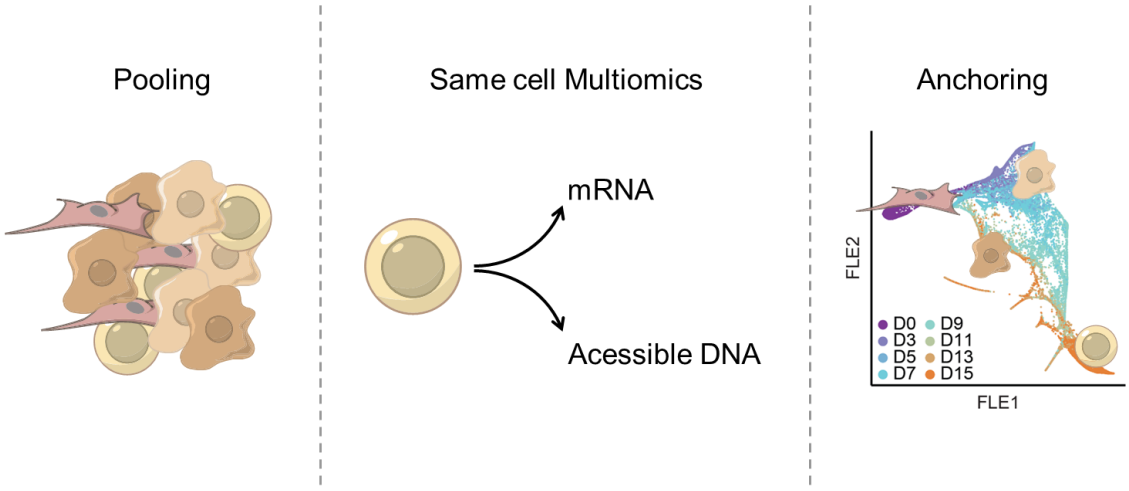

Figure 39: Multiome approach. Cells from the whole time-course are pooled together (Pooling) and sequenced for both mRNA and accessible DNA (Same cell Multiomics). The transcriptional profile of scRNA-seq data from this work is used to characterize the gene expression phenotype of pooled cells (Anchoring).

3.8.1.1 Anchoring

First, the gene expression profile from the Multiome data must be characterized. Addressing the phenotype of the pool of cells is needed to be able to assign specific epigenetic rearrangement to the correct cells. We re-analyzed scRNA-seq data to render them compatible with the newly generated data (Methods). Reciprocal PCA (RPCA) is an algorithm designed to identify anchors when cell types are conserved, but there are very substantial differences in gene expression across experiments. Thus, it is recommended during integrative analysis. We visualized scRNA-seq data on a new UMAP and obtained a very similar pattern to the FLE from Figure 16 (Figure 40). When highlighting the

transcriptional profiles from Multiome onto the same UMAP, an almost complete coverage of all the described phenotypic states could be reached, confirming the suitability of this approach to our purposes (Figure 40).

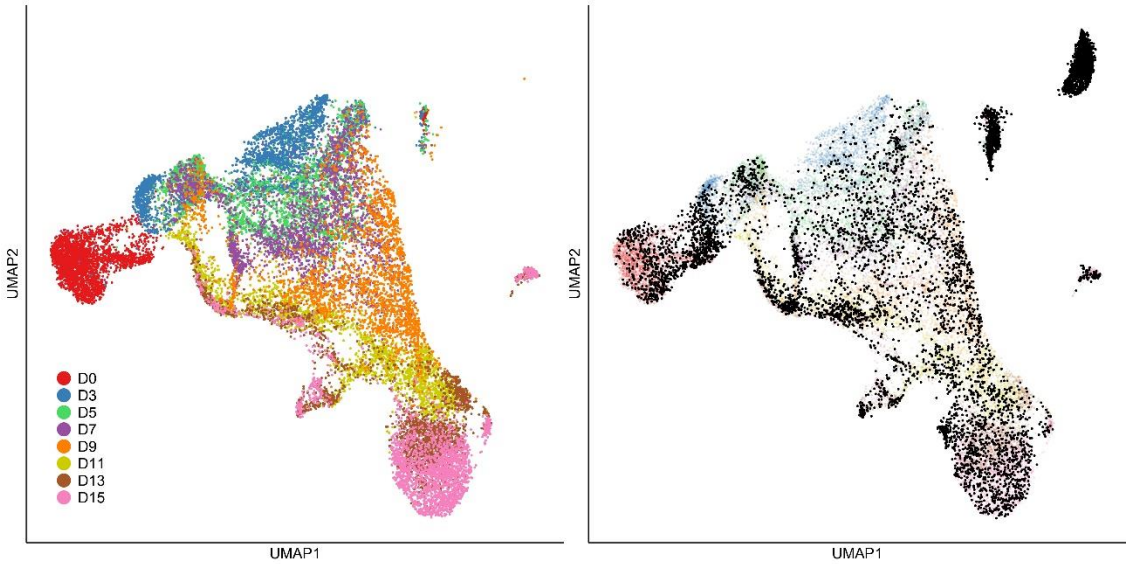

Figure 40: RPCA enables correct anchoring. UMAP visualization of scRNA-seq data from this work after re-analysis with (right) or without (left) pooled cells from Multiome approach (black dots).

3.8.1.2 Chromatin accessibility and gene expression correlation

The anchoring step was followed by the identification of new clusters (Methods). The results showed 14 different clusters that were used to evaluate the agreement with the old ones. This led to the annotation of the new clusters with the old nomenclature. To evaluate the robustness of the phenotypic characterization of the Multiome cells, we looked at the expression of marker genes of late time-points for both fates, i.e. developmental and somatic. Since the Multiome approach gives the opportunity to link gene expression with chromatin accessibility, the markers were evaluated from both points of view (Figure 41). SOX2, a pluripotency marker, shows increasing expression levels from DR1 to DR3 (clusters that define the pluripotency route). The same pattern can be observed at the chromatin levels, where the locus associated with SOX2 promoter is progressively open from DR1 to DR3, resulting in closed in the other clusters. On the other hand, ELN characterizes the late stages of the somatic commitment. Albeit its expression peak occurred in SR7, as expected, chromatin accessibility revealed that ELN promoter does not appear to be differentially open between clusters. In contrast, there is a distal region downstream of ELN locus that is specifically open in SR clusters.

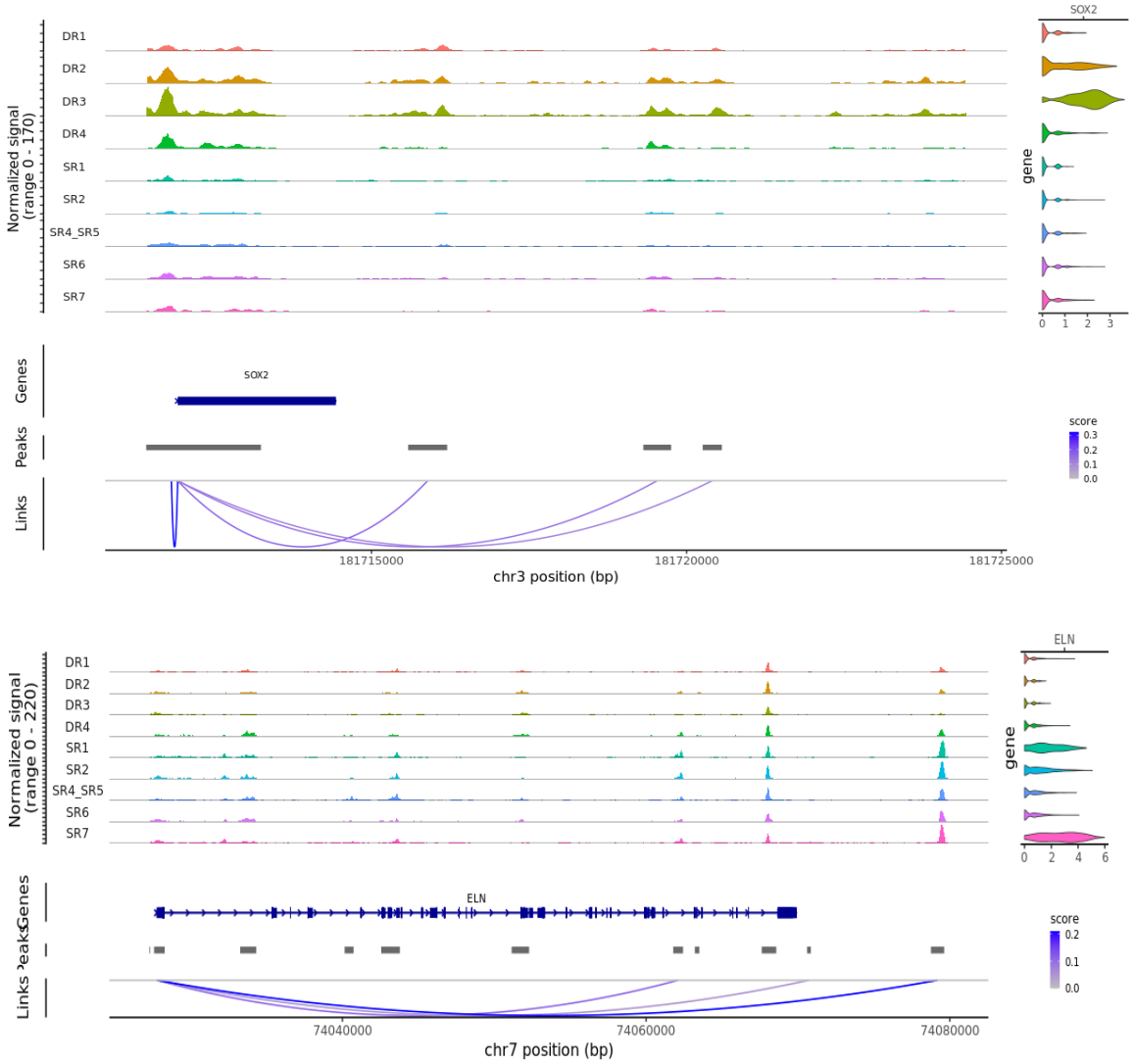

Figure 41: Correlation of gene expression and chromatin accessibility at marker genes loci. Reads distribution from accessible DNA at SOX2 (top) and ELN (bottom) loci, divided by clusters. Violin plots of gene expression (log_2 CPM) are reported on the right. Links are generated by correlating gene expression with chromatin accessibility at called peaks.

4 Discussion

4.1 Identification and characterization of an unproductive cell fate during reprogramming

Our integrative approach of secretome and single-cell transcriptomic analyses revealed a previously unappreciated crosstalk between subpopulations during the intermediate stages of human reprogramming. Whilst population heterogeneity was also described in recent papers, both in mouse[65–67] and human[170,173,174], these works reported the formation of distinctive cell clusters and diversification of pluripotent trajectories, viewing the unproductive/refractory subpopulations as a "problem" or limitation in the process. Instead, here we highlight the crucial role of reprogramming intermediates and the positive contribution of non-pluripotent clusters as actively supporting and shaping the route of the reprogramming cells towards a hiPSC identity.

The efficiency of human somatic cell reprogramming heavily relies on the successful transient accessibility and overcoming of specific intermediate stages but given the generally low reprogramming efficiency, these stages have been hard to identify. Few strategies were previously adopted to capture human intermediate reprogramming-committed subpopulations such as cell sorting[32,173] and secondary reprogramming systems[64].

Supported by the microfluidic culture system, we took a step further through the unbiased identification of the reprogramming subpopulation trajectories and interactions based on an integrative secreted proteome and scRNA-seq analysis. The former identified a number of secreted cytokines, growth factors and ECM-related proteins actually present in the extracellular space during reprogramming and contributing to establish an environmental signaling resembling the early embryo basal lamina. scRNA-seq identified two main trajectories during reprogramming, with one almost exclusively responsible for secretory activity and one committed to reprogram. It was probably the reduced secretory activity of nascent hiPSCs or their low abundance that led previous works to overlook the role of the extracellular environment, failing to recognize nascent hiPSCs as a secretome target[173]. Recently, a few works suggested the potential for cross-population signalling in mouse reprogramming[65–67] including the role of SASP and senescence[163], but until now the molecular mechanisms and rationale behind human non-cell autonomous signaling remained unclear.

4.2 Role of the somatic fate in subpopulation crosstalk

In this study, scRNA-seq could identify the putative subpopulation interaction dynamics during microfluidic reprogramming. In particular, the identification of the two distinctive trajectories, somatic secretory and reprogramming, was instrumental for scoring the putative ligand-receptor association responsible for the unidirectional support of the developmental trajectory towards pluripotency. Secretome analysis, performed here for the first time, could further reduce the dimensionality of the interactions, restricting them to those whose soluble ligand was actually detected as secreted at protein level. Only four ligands passed these restrictive selection criteria: INHBA, SPP1, NRG1 and HGF. INHBA was previously described[168], SPP1 is downstream of the HGF pathway[169], thus we focused on NRG1 and HGF, not previously implicated as reprogramming regulators. Interestingly, NRG1 signalling occurred within the reprogramming trajectory, while HGF involved population cross-talk from the secretory somatic to the reprogramming trajectory.

HGF is part of SASP, however it was not measured in Mosteiro et al., 2016[163] who instead identified IL6 in mouse cell reprogramming. Both HGF and IL6 signaling have STAT3 as a common effector, although via different receptors[175], and other works reported a positive correlation between STAT3 activity and in vivo reprogramming efficiency[68,176]. In our human reprogramming systems[64,124], IL6 was present both at transcriptional and proteomic level, however we could not detect its receptor, IL6R, in any subpopulation at any stage. Indeed, we were able to enhance reprogramming efficiency with IL6 only upon providing a soluble form of IL6R. The axis HGF/MET/STAT3 was first reported in cancer stemness and promotes the expression of pluripotent genes[175]. HGF-MET was demonstrated to take part in a mesenchymal-epithelial cross-talk[177].

We performed extensive experimental validation both in microfluidics and in conventional culture systems. Our loss of function data clearly show that MET activation and STAT3 signaling play an important role in preserving the efficiency of reprogramming, supporting the idea that HGF/MET/STAT3 may have a crucial role in the phenotypic conversion of developmental subpopulation towards pluripotency. Our gain of function experiments within the conventional culture system (i.e., Petri dish) support our hypothesis of the role of miniaturization in concentrating endogenous HGF and show the possibility of scaling up our findings for wider applicability. Whilst a positive role of STAT3 signaling has been extensively characterized during maintenance and induction of mouse naive pluripotency[178], STAT3 signaling pathway is not active in primed human hiPSCs. It is therefore particularly striking that we find transient STAT3 activity to be of benefit during human reprogramming to primed hiPSC identity, and highlights that we must consider the environmental niche

requirements of the intermediate states, which may differ from those of the endpoint target identity.

4.3 An early embryonic stage is recapitulated during reprogramming

HGF/MET physiological expression during development starts in the primitive streak where they take part into the so-called branching morphogenesis[179,180]. Therefore, it is intriguing to observe in our data the recapitulation of ECM organization resembling this state[181,182], with HGF secreted within the somatic trajectory, while its receptor, MET, especially present along the reprogramming one (Figure 42).

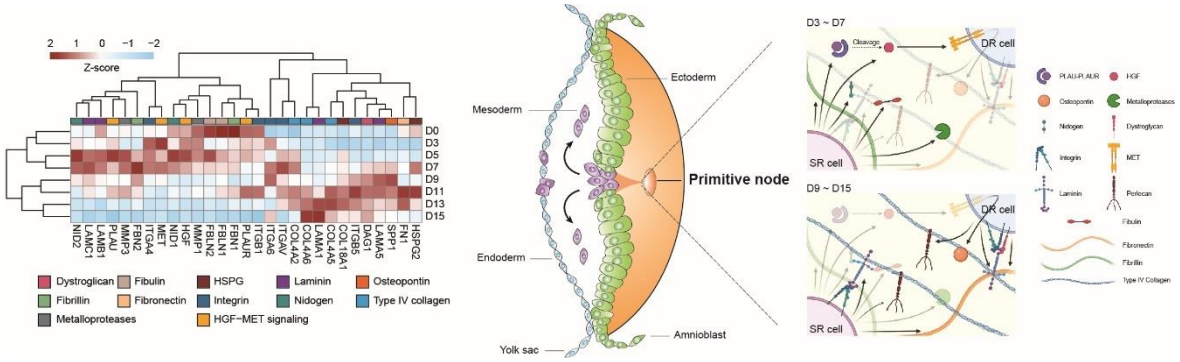

Figure 42: Primitive node formation and reprogramming. Left: Heatmap of Z-scored log2 counts per million, averaged by day, of genes encoding for primitive node components. Right: Schematic representation of primitive node formation (Adapted from Boccaccio and Comoglio, 2006) and primitive node components (Created with BioRender.com). Snapshot of early and late events are reported according to the expression dynamics on the left. Black arrows show the contribution of SR and DR cells based on their average gene expression.

In our work, we followed an unbiased approach that supports the idea that the route to pluripotency can be broadened by cell-non-autonomous mechanisms. Paracrine signaling is established by highly regulated dynamics with multi-factorial contribution. We showed the use of HGF for gain of function during reprogramming in a conventional culture system, but this efficiency was amenable to further enhancement when multifactorial contributions were used. In particular, we used IL6 and soluble IL6R for a more effective downstream activation of STAT3. Moreover, we found that NRG1 contribute to enhance efficiency of hiPSC formation consistently with previous works, which upon binding ERBB2/ERBB3 receptors activates MAPK/ERK pathway and showed improved maintenance and passage of hiPSCs[183,184].

5 Conclusion

In conclusion, this work reports an overview of the environment-mediated subpopulation cross-talk during reprogramming and identifies some specific critical players. Important implications of our work are related to in vivo reprogramming, where environmental factors cannot be controlled but may affect potential applications. Moreover, strategies to reprogram in vitro fibroblasts from any donor with high efficiency are down the road and unlock the possibilities of using hiPSC as modeling systems for a large number of patients, including their use as diagnostic tools in predicting patient-specific genotype-phenotype associations in disease.

6 References

1. Kelly, S. J. Studies of the developmental potential of 4- and 8-cell stage mouse blastomeres. *J Exp Zool* **200**, 365–376 (1977).

2. Sukoyan, M. A. *et al.* Embryonic stem cells derived from morulae, inner cell mass, and blastocysts of mink: comparisons of their pluripotencies. *Mol Reprod Dev* **36**, 148–158 (1993).

3. Evans, M. J. & Kaufman, M. H. Establishment in culture of pluripotential cells from mouse embryos. *Nature* **292**, 154–156 (1981).

4. Martin, G. R. Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells. *Proc Natl Acad Sci U S A* **78**, 7634–7638 (1981).

5. Larijani, B. *et al.* Stem Cell Therapy in Treatment of Different Diseases. *Acta Med Iran* **50**, 79–96 (2012).

6. Gehring, W. Clonal analysis of determination dynamics in cultures of imaginal disks in Drosophila melanogaster. *Dev Biol* **16**, 438–456 (1967).

7. Hadorn, E. [Constancy, variation and type of determination and differentiation in cells from male genitalia rudiments of Drosophila melanogaster in permanent culture in vivo]. *Dev Biol* **13**, 424–509 (1966).

8. Lièvre, C. L. Le & Douarin, N. L. Le. Mesenchymal derivatives of the neural crest: analysis of chimaeric quail and chick embryos. *J Embryol Exp Morphol* (1975).

9. Gurdon, J. B. The egg and the nucleus: a battle for supremacy. *Development* **140**, 2449–2456 (2013).

10. Gurdon, J. B. & Melton, D. A. Nuclear reprogramming in cells. *Science* **322**, 1811–1815 (2008).

11. Gurdon, J. B. & Byrne, J. A. The first half-century of nuclear transplantation. *Biosci Rep* **24**, 545–557 (2004).

12. Wilmut, I., Schnieke, A. E., McWhir, J., Kind, A. J. & Campbell, K. H. S. Viable offspring derived from fetal and adult mammalian cells. *Nature* **385**, 810–813 (1997).

13. Graf, T. Historical origins of transdifferentiation and reprogramming. *Cell Stem Cell* **9**, 504–516 (2011).

14. Graf, T. & Enver, T. Forcing cells to change lineages. *Nature* **462**, 587–594 (2009).

15. Lassar, A. B., Paterson, B. M. & Weintraub, H. Transfection of a DNA locus that mediates the conversion of 10T1/2 fibroblasts to myoblasts. *Cell* **47**, 649–656 (1986).

16. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* **51**, 987–1000 (1987).

17. Kulessa, H., Frampton, J. & Graf, T. GATA-1 reprograms avian myelomonocytic cell lines into eosinophils, thromboblasts, and erythroblasts. *Genes Dev* **9**, 1250–1262 (1995).

18. Nerlov, C. & Graf, T. PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. *Genes Dev* **12**, 2403–2412 (1998).

19. Laiosa, C. V., Stadtfeld, M., Xie, H., de Andres-Aguayo, L. & Graf, T. Reprogramming of committed T cell progenitors to macrophages and dendritic cells by C/EBP alpha and PU.1 transcription factors. *Immunity* **25**, 731–744 (2006).

20. Xie, H., Ye, M., Feng, R. & Graf, T. Stepwise reprogramming of B cells into macrophages. *Cell* **117**, 663–676 (2004).

21. Zhou, Q., Brown, J., Kanarek, A., Rajagopal, J. & Melton, D. A. In vivo reprogramming of adult pancreatic exocrine cells to β-cells. *Nature 2008 455:7213* **455**, 627–632 (2008).

22. Ieda, M. *et al.* Direct reprogramming of fibroblasts into functional cardiomyocytes by defined factors. *Cell* **142**, 375–386 (2010).

23. Qian, L. *et al.* In vivo reprogramming of murine cardiac fibroblasts into induced cardiomyocytes. *Nature* **485**, 593–598 (2012).

24. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035–1041 (2010).

25. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

26. Maherali, N. *et al.* Directly reprogrammed fibroblasts show global epigenetic remodeling and widespread tissue contribution. *Cell Stem Cell* **1**, 55–70 (2007).

27. Okita, K., Ichisaka, T. & Yamanaka, S. Generation of germline-competent induced pluripotent stem cells. *Nature* **448**, 313–317 (2007).

28. Wernig, M. *et al.* In vitro reprogramming of fibroblasts into a pluripotent ES-cell-like state. *Nature* **448**, 318–324 (2007).

29. Meissner, A., Wernig, M. & Jaenisch, R. Direct reprogramming of genetically unmodified fibroblasts into pluripotent stem cells. *Nat Biotechnol* **25**, 1177–1181 (2007).

30. Kang, L., Wang, J., Zhang, Y., Kou, Z. & Gao, S. iPS cells can support full-term development of tetraploid blastocyst-complemented embryos. *Cell Stem Cell* **5**, 135–138 (2009).

31. Boland, M. J. *et al.* Adult mice generated from induced pluripotent stem cells. *Nature* **461**, 91–94 (2009).

32. Takahashi, K. *et al.* Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. *Cell* **131**, 861–872 (2007).

33. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917–1920 (2007).

34. Park, I. H. *et al.* Disease-specific induced pluripotent stem (iPS) cells. *Cell* **134**, 877 (2008).

35. Li, W. *et al.* Generation of rat and human induced pluripotent stem cells by combining genetic reprogramming and chemical inhibitors. *Cell Stem Cell* **4**, 16–19 (2009).

36. Liu, H. *et al.* Generation of induced pluripotent stem cells from adult rhesus monkey fibroblasts. *Cell Stem Cell* **3**, 587–590 (2008).

37. Friedrich Ben-Nun, I. *et al.* Induced pluripotent stem cells from highly endangered species. *Nat Methods* **8**, 829–831 (2011).

38. Aasen, T. *et al.* Efficient and rapid generation of induced pluripotent stem cells from human keratinocytes. *Nat Biotechnol* **26**, 1276–1284 (2008).

39. Hanna, J. *et al.* Direct reprogramming of terminally differentiated mature B lymphocytes to pluripotency. *Cell* **133**, 250–264 (2008).

40. Di Stefano, B. *et al.* C/EBPα poises B cells for rapid reprogramming into induced pluripotent stem cells. *Nature* **506**, 235–239 (2014).

41. Yamanaka, S. & Blau, H. M. Nuclear reprogramming to a pluripotent state by three approaches. *Nature* **465**, 704–712 (2010).

42. Wu, S. M. & Hochedlinger, K. Harnessing the potential of induced pluripotent stem cells for regenerative medicine. *Nat Cell Biol* **13**, 497–505 (2011).

43. Robinton, D. A. & Daley, G. Q. The promise of induced pluripotent stem cells in research and therapy. *Nature 2012 481:7381* **481**, 295–305 (2012).

44. Bellin, M., Marchetto, M. C., Gage, F. H. & Mummery, C. L. Induced pluripotent stem cells: the new patient? *Nat Rev Mol Cell Biol* **13**, 713–726 (2012).

45. Ebert, A. D. *et al.* Induced pluripotent stem cells from a spinal muscular atrophy patient. *Nature* **457**, 277–280 (2009).

46. Carvajal-Vergara, X. *et al.* Patient-specific induced pluripotent stem-cell-derived models of LEOPARD syndrome. *Nature* **465**, 808–812 (2010).

47. Yazawa, M. *et al.* Using induced pluripotent stem cells to investigate cardiac phenotypes in Timothy syndrome. *Nature* **471**, 230–236 (2011).

48. Moretti, A. *et al.* Patient-specific induced pluripotent stem-cell models for long-QT syndrome. *N Engl J Med* **363**, 1397–1409 (2010).

49. Fatima, A. *et al.* In vitro modeling of ryanodine receptor 2 dysfunction using human induced pluripotent stem cells. *Cell Physiol Biochem* **28**, 579–592 (2011).

50. Jung, C. B. *et al.* Dantrolene rescues arrhythmogenic RYR2 defect in a patient-specific stem cell model of catecholaminergic polymorphic ventricular tachycardia. *EMBO Mol Med* **4**, 180–191 (2012).

51. Novak, A. *et al.* Cardiomyocytes generated from CPVTD307H patients are arrhythmogenic in response to β-adrenergic stimulation. *J Cell Mol Med* **16**, 468–482 (2012).

52. Brennand, K. J. *et al.* Modelling schizophrenia using human induced pluripotent stem cells. *Nature* **473**, 221–225 (2011).

53. Israel, M. A. *et al.* Probing sporadic and familial Alzheimer's disease using induced pluripotent stem cells. *Nature* **482**, 216–220 (2012).

54. Hanna, J. *et al.* Treatment of sickle cell anemia mouse model with iPS cells generated from autologous skin. *Science* **318**, 1920–1923 (2007).

55. Wernig, M. *et al.* Neurons derived from reprogrammed fibroblasts functionally integrate into the fetal brain and improve symptoms of rats with Parkinson's disease. *Proc Natl Acad Sci U S A* **105**, 5856–5861 (2008).

56. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).

57. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell* **122**, 947–956 (2005).

58. Hanna, J. *et al.* Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature* **462**, 595–601 (2009).

59. Buganim, Y. *et al.* Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase. *Cell* **150**, 1209–1222 (2012).

60. Yamanaka, S. Elite and stochastic models for induced pluripotent stem cell generation. *Nature* vol. 460 49–52 Preprint at https://doi.org/10.1038/nature08180 (2009).

61. Chronis, C. *et al.* Cooperative Binding of Transcription Factors Orchestrates Reprogramming. *Cell* **168**, 442-459.e20 (2017).

62. Papp, B. & Plath, K. Epigenetics of reprogramming to induced pluripotency. *Cell* **152**, 1324–1343 (2013).

63. Takahashi, K. *et al.* Induction of pluripotency in human somatic cells via a transient state resembling primitive streak-like mesendoderm. *Nat Commun* **5**, 1–9 (2014).

64. Cacchiarelli, D. *et al.* Integrative Analyses of Human Reprogramming Reveal Dynamic Nature of Induced Pluripotency. *Cell* **162**, 412–424 (2015).

65. Tran, K. A. *et al.* Defining Reprogramming Checkpoints from Single-Cell Analyses of Induced Pluripotency. *Cell Rep* **27**, 1726-1741.e5 (2019).

66. Schiebinger, G. *et al.* Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell* **176**, 928-943.e22 (2019).

67. Zhao, T. *et al.* Single-Cell RNA-Seq Reveals Dynamic Early Embryonic-like Programs during Chemical Reprogramming. *Cell Stem Cell* **23**, 31-45.e7 (2018).

68. Mosteiro, L., Pantoja, C., de Martino, A. & Serrano, M. Senescence promotes in vivo reprogramming through p16 INK4a and IL-6. *Aging Cell* **17**, (2018).

69. Mahmoudi, S. *et al.* Heterogeneity in old fibroblasts is linked to variability in reprogramming and wound healing. *Nature* **574**, 553–558 (2019).

70. Heather, J. M. & Chain, B. The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**, 1 (2016).

71. Watson, J. D. & Crick, F. H. C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature 1953 171:4356* **171**, 737–738 (1953).

72. Hutchison, C. A. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res* **35**, 6227–6237 (2007).

73. Wu, R. & Kaiser, A. D. Structure and base sequence in the cohesive ends of bacteriophage lambda DNA. *J Mol Biol* **35**, 523–537 (1968).

74. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc Natl Acad Sci U S A* **74**, 560–564 (1977).

75. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* **74**, 5463–5467 (1977).

76. Ansorge, W., Sproat, B., Stegemann, J., Schwager, C. & Zenke, M. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic Acids Res* **15**, 4593 (1987).

77. Luckey, J. A. *et al.* High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res* **18**, 4417 (1990).

78. Swerdlow, H. & Gesteland, R. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res* **18**, 1415–1419 (1990).

79. Prober, J. M. *et al.* A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–341 (1987).

80. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242**, 84–89 (1996).

81. Voelkerding, K. V., Dames, S. A. & Durtschi, J. D. Next-generation sequencing: from basic research to diagnostics. *Clin Chem* **55**, 641–658 (2009).

82. McKernan, K. J. *et al.* Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res* **19**, 1527–1541 (2009).

83. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* **13**, 278 (2015).

84. Jain, M., Olsen, H. E., Paten, B. & Akeson, M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* **17**, (2016).

85. Illumina. *An introduction to Next-Generation Sequencing Technology*. www.illumina.com/technology/next-generation-sequencing.html.

86. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology 2016 17:1* **17**, 1–19 (2016).

87. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat Methods* **8**, 469–477 (2011).

88. Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L. & Rice, P. M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* **38**, 1767 (2010).

89. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

90. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114 (2014).

91. BBMap download | SourceForge.net. https://sourceforge.net/projects/bbmap/.

92. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

93. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15 (2013).

94. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

95. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).

96. Mitelman, F., Johansson, B. & Mertens, F. The impact of translocations and gene fusions on cancer causation. *Nat Rev Cancer* **7**, 233–245 (2007).

97. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25**, 1491–1498 (2015).

98. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Exp Mol Med* **50**, (2018).

99. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and biological applications. *Mol Aspects Med* **59**, 36–46 (2018).

100. Plasschaert, L. W. *et al.* A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature 2018 560:7718* **560**, 377–381 (2018).

101. Suo, S. *et al.* Revealing the Critical Regulators of Cell Identity in the Mouse Cell Atlas. *Cell Rep* **25**, 1436-1445.e3 (2018).

102. Fischer, D. S. *et al.* Inferring population dynamics from single-cell RNA-sequencing time series data. *Nature Biotechnology 2019 37:4* **37**, 461–468 (2019).

103. Velasco, S. *et al.* Individual brain organoids reproducibly form cell diversity of the human cerebral cortex. *Nature 2019 570:7762* **570**, 523–527 (2019).

104. Liu, Z. *et al.* Single-cell transcriptomics reconstructs fate conversion from fibroblast to cardiomyocyte. *Nature 2017 551:7678* **551**, 100–104 (2017).

105. Cacchiarelli, D. *et al.* Aligning Single-Cell Developmental and Reprogramming Trajectories Identifies Molecular Determinants of Myogenic Reprogramming Outcome. *Cell Syst* **7**, 258-268.e3 (2018).

106. Moghe, I., Loupy, A. & Solez, K. The Human Cell Atlas Project by the numbers: Relationship to the Banff Classification. *Am J Transplant* **18**, 1830 (2018).

107. Savas, P. *et al.* Single-cell profiling of breast cancer T cells reveals a tissue-resident memory subset associated with improved prognosis. *Nat Med* **24**, 986–993 (2018).

108. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell* **65**, 631-643.e4 (2017).

109. Baran-Gale, J., Chandra, T. & Kirschner, K. Experimental design for single-cell RNA sequencing. *Brief Funct Genomics* **17**, 233–239 (2018).

110. Salomon, R. *et al.* Droplet-based single cell RNAseq tools: a practical guide. *Lab Chip* **19**, 1706–1727 (2019).

111. Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).

112. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature Communications 2017 8:1* **8**, 1–12 (2017).

113. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).

114. Islam, S. *et al.* Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods 2013 11:2* **11**, 163–166 (2013).

115. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol* **21**, (2020).

116. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145 (2015).

117. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* **9**, (2017).

118. Slovin, S. *et al.* Single-Cell RNA Sequencing Analysis: A Step-by-Step Overview. *Methods Mol Biol* **2284**, 343–365 (2021).

119. Andrews, T. S. & Hemberg, M. Identifying cell populations with scRNASeq. *Mol Aspects Med* **59**, 114–122 (2018).

120. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology 2014 32:4* **32**, 381–386 (2014).

121. Rotem, A. *et al.* Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nature Biotechnology 2015 33:11* **33**, 1165–1172 (2015).

122. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925–936 (2019).

123. *Simultaneous profiling of the transcriptome and epigenome from the same cell, Document number LIT000099 Rev C, 10x Genomics (2021)*.

124. Gagliano, O. *et al.* Microfluidic reprogramming to pluripotency of human somatic cells. *Nat Protoc* **14**, 722–737 (2019).

125. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat Methods* **6**, 359–362 (2009).

126. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* **11**, 2301–2319 (2016).

127. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (1979)* **347**, (2015).

128. Gonzalez, R. *et al.* Screening the mammalian extracellular proteome for regulators of embryonic human stem cell pluripotency. *Proc Natl Acad Sci U S A* **107**, 3552–3557 (2010).

129. Coppé, J. P., Desprez, P. Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: The dark side of tumor suppression. *Annual Review of Pathology: Mechanisms of Disease* vol. 5 99–118 Preprint at https://doi.org/10.1146/annurev-pathol-121808-102144 (2010).

130. Coppé, J.-P. *et al.* Senescence-Associated Secretory Phenotypes Reveal Cell-Nonautonomous Functions of Oncogenic RAS and the p53 Tumor Suppressor. *PLoS Biol* **6**, e301 (2008).

131. Lopes-Paciencia, S. *et al.* The senescence-associated secretory phenotype and its regulation. *Cytokine* **117**, 15–22 (2019).

132. Acosta, J. C. *et al.* A complex secretory program orchestrated by the inflammasome controls paracrine senescence. *Nat Cell Biol* **15**, 978–990 (2013).

133. Yu, G. & He, Q. Y. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**, 477–479 (2016).

134. Bindea, G. *et al.* ClueGO: A Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25**, 1091–1093 (2009).

135. Shannon, P. *et al.* Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498–2504 (2003).

136. Boroviak, T. *et al.* Single cell transcriptome analysis of human, marmoset and mouse embryos reveals common and divergent features of preimplantation development. *Development (Cambridge)* **145**, (2018).

137. Naba, A. *et al.* The matrisome: In silico definition and in vivo characterization by proteomics of normal and tumor extracellular matrices. *Molecular and Cellular Proteomics* **11**, (2012).

138. Ramilowski, J. A. *et al.* A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* **6**, 1–12 (2015).

139. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology 2015 33:5* **33**, 495–502 (2015).

140. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *bioRxiv* 060012 (2016) doi:10.1101/060012.

141. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* **1**, 417–425 (2015).

142. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

143. Kowalczyk, M. S. *et al.* Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells. *Genome Res* **25**, 1860–1872 (2015).

144. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell differentiation. *Nature* **518**, 344–349 (2015).

145. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).

146. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).

147. Love, M., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *bioRxiv* 2832 (2014) doi:10.1101/002832.

148. Manos, P. D., Ratanasirintrawoot, S., Loewer, S., Daley, G. Q. & Schlaeger, T. M. Live-cell immunofluorescence staining of human pluripotent stem cells. *Curr Protoc Stem Cell Biol* **Chapter 1**, (2011).

149. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nature Methods 2021 18:11* **18**, 1333–1341 (2021).

150. Luni, C., Gagliano, O. & Elvassore, N. Derivation and Differentiation of Human Pluripotent Stem Cells in Microfluidic Devices. *Annu Rev Biomed Eng* **24**, 231–248 (2022).

151. Luni, C. *et al.* High-efficiency cellular reprogramming with microfluidics. *Nat Methods* **13**, 446–452 (2016).

152. Hu, Q., Luni, C. & Elvassore, N. Microfluidics for secretome analysis under enhanced endogenous signaling. *Biochem Biophys Res Commun* **497**, 480–484 (2018).

153. Rui, X. *et al.* Extracellular phosphoprotein regulation is affected by culture system scale-down. *Biochimica et Biophysica Acta (BBA) - General Subjects* **1866**, 130165 (2022).

154. Michielin, F. *et al.* The Microfluidic Environment Reveals a Hidden Role of Self-Organizing Extracellular Matrix in Hepatic Commitment and Organoid Formation of hiPSCs. *Cell Rep* **33**, (2020).

155. Tolomeo, A. M. *et al.* NGN2 mmRNA-Based Transcriptional Programming in Microfluidic Guides hiPSCs Toward Neural Fate With Multiple Identities. *Front Cell Neurosci* **15**, (2021).

156. Gagliano, O. *et al.* Synchronization between peripheral circadian clock and feeding-fasting cycles in microfluidic device sustains oscillatory pattern of transcriptome. *Nat Commun* **12**, (2021).

157. Cesare, E. *et al.* 3D ECM-Rich Environment Sustains the Identity of Naïve Human iPSCs. *SSRN Electronic Journal* (2021) doi:10.2139/SSRN.3761454.

158. Giulitti, S. *et al.* Direct generation of human naive induced pluripotent stem cells from somatic cells in microfluidics. *Nat Cell Biol* **21**, 275–286 (2019).

159. Qin, J. & Gronenborn, A. M. Weak protein complexes: Challenging to study but essential for life. *FEBS Journal* vol. 281 1948–1949 Preprint at https://doi.org/10.1111/febs.12744 (2014).

160. Marson, A. *et al.* Wnt Signaling Promotes Reprogramming of Somatic Cells to Pluripotency. *Cell Stem Cell* vol. 3 132–135 Preprint at https://doi.org/10.1016/j.stem.2008.06.019 (2008).

161. Andrae, J., Gallini, R. & Betsholtz, C. Role of platelet-derived growth factors in physiology and medicine. *Genes and Development* vol. 22 1276–1312 Preprint at https://doi.org/10.1101/gad.1653708 (2008).

162. Phanstiel, D. H. *et al.* Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nat Methods* **8**, 821–827 (2011).

163. Mosteiro, L. *et al.* Tissue damage and senescence provide critical signals for cellular reprogramming in vivo. *Science (1979)* **354**, (2016).

164. Hartman, A. A. *et al.* YAP Non-cell-autonomously Promotes Pluripotency Induction in Mouse Cells. *Stem Cell Reports* **14**, 730–743 (2020).

165. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, 10008 (2008).

166. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).

167. Tesar, P. J. *et al.* New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature* **448**, 196–199 (2007).

168. James, D., Levine, A. J., Besser, D. & Hemmati-Brivanlou, A. TGFβ/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development* **132**, 1273–1282 (2005).

169. Medico, E. *et al.* Osteopontin is an autocrine mediator of hepatocyte growth factor-induced invasive growth. *Cancer Res* **61**, 5861–5868 (2001).

170. Liu, X. *et al.* Reprogramming roadmap reveals route to human induced trophoblast stem cells. *Nature* **586**, 101–107 (2020).

171. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733–D745 (2016).

172. Graf, U., Casanova, E. A. & Cinelli, P. The role of the leukemia inhibitory factor (LIF) - Pathway in derivation and maintenance of murine pluripotent stem cells. *Genes* vol. 2 280–297 Preprint at https://doi.org/10.3390/genes2010280 (2011).

173. Xing, Q. R. *et al.* Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Sci Adv* **6**, (2020).

174. Guo, L. *et al.* Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq A generic bifurcation model for cell fate decisions. *Mol Cell* **73**, 815–829 (2019).

175. Boccaccio, C. *et al.* Induction of epithelial tubules by growth factor HGF depends on the STAT pathway. *Nature* **391**, 285–288 (1998).

176. Khourieh, J. *et al.* A deep intronic splice mutation of STAT3 underlies hyper IgE syndrome by negative dominance. *Proc Natl Acad Sci U S A* **116**, 16463–16472 (2019).

177. Prat, M. *et al.* The receptor encoded by the human C-MET oncogene is expressed in hepatocytes, epithelial cells and solid tumors. *Int J Cancer* **49**, 323–328 (1991).

178. Chen, H. *et al.* Reinforcement of STAT3 activity reprogrammes human embryonic stem cells to naive-like pluripotency. *Nat Commun* **6**, (2015).

179. Andermarcher, E., Surani, M. A. & Gherardi, E. Co-expression of the HGF/SF and c-met genes during early mouse embryogenesis precedes reciprocal expression in adjacent tissues during organogenesis. *Dev Genet* **18**, 254–266 (1996).

180. Sonnenberg, E., Meyer, D., Weidner, K. M. & Birchmeier, C. Scatter factor/hepatocyte growth factor and its receptor, the c-met tyrosine kinase, can mediate a signal exchange between mesenchyme and epithelia during mouse development. *Journal of Cell Biology* **123**, 223–235 (1993).

181. Nakaya, Y., Sukowati, E. W., Alev, C., Nakazawa, F. & Sheng, G. Involvement of dystroglycan in epithelial-mesenchymal transition during chick gastrulation. *Cells Tissues Organs* **193**, 64–73 (2010).

182. Mogi, K. & Toyoizumi, R. Invasion by matrix metalloproteinase-expressing cells is important for primitive streak formation in early chick blastoderm. *Cells Tissues Organs* **192**, 1–16 (2010).

183. Kuo, H. H. *et al.* Negligible-Cost and Weekend-Free Chemically Defined Human iPSC Culture. *Stem Cell Reports* **14**, 256–270 (2020).

184. Marinho, P. A., Chailangkarn, T. & Muotri, A. R. Systematic optimization of human pluripotent stem cells media using Design of Experiments. *Sci Rep* **5**, 1–13 (2015).

185. Eckfeldt, C. E., Mendenhall, E. M. & Verfaillie, C. M. The molecular repertoire of the 'almighty' stem cell. *Nat Rev Mol Cell Biol* **6**, 726–737 (2005).

186. Chromium Next GEM Single Cell 3′ Reagent Kits v3.1. (2019).