Moritz Schlapansky

Benchmarking a Novel Screening Platform to Decipher Cell State Transitions

Research Report

Marshall Plan Scholarship

University of Vienna
Center for Molecular Biology

University of California, San Francisco
Department of Microbiology and Immunology

Supervision

Professor Michael T. McManus, PhD
Professor Peter Fuchs, PhD

Acknowledgements

I want to thank the Austrian Marshall Plan Foundation for their generous support of this project. I also want to give a warm “Thank you” to the following people, without whom this thesis would not have been possible. I want to thank Professor Michael McManus for being an inspiring mentor, for his guidance and for giving me the opportunity to work on this project. Also, many thanks to Professor Peter Fuchs for support and advice. I also want to thank my colleagues at the McManus lab for creating an amazing, open and warm atmosphere ideal for doing science, for always having an open ear for my many questions and being extremely helpful in every way. Finally, I want to thank my friends and family for always supporting me and helping me through the difficult times.

Table of Contents

Abstract

Cell state transitions play a principal role in many fields of biology and might hold answers to several questions about fundamental biological phenomena. For instance, tumor cells can transition into a non-genetic, therapy-resistant state or into a metastatic state, immune cells can transition between different effector states and any differentiation can be viewed as an extreme alteration of cell state. However, current technology does not offer adequate tools to investigate the mechanisms and factors underlying cell state transitions. I aimed to overcome current limitations by leveraging a novel high-throughput genetic screening paradigm to develop a method to identify factors and mechanisms underlying cell state transitions. This method will allow asking more complex questions than conventional pooled genetic screening methods and will thus support our quest for new, fundamental discoveries.

1 Introduction

Cells were first discovered by Robert Hooke in 1665 as he observed cork and other plant tissues through his microscope[1]. However, it took about 200 years until the botanist Matthias Jakob Schleiden and the zoologist Theodor Schwann amongst others suggested the cell theory: the theory that all living organisms and their tissues are formed of cells or their products and that cells can only be generated from other cells[2]. Hence, the eukaryotic cell can be viewed as the most basic unit of structure and function of multicellular organisms. Acting through macromolecules like DNA, RNA and proteins, phenotypes on an organismal level arise through the cellular level as each cell is a functional unit regulating their own gene expression program. To understand biological processes and their dysregulation in disease, it is vital that we understand them at the cellular level. There are a lot of diseases which are caused by a molecular fault (e.g. a mutation) that becomes apparent in a certain cell type or certain cell types. For instance, a mutation in the gene that encodes Dystrophin (DMD) causes molecular alterations in skeletal muscle cells that lead to muscular dystrophy[3]. Certain mutations in the Cystic fibrobsis membrane conductance regulator (CFTR) gene can cause misfolding of its gene product and thereby impair the regulation of flow of ions in cells in the sweat glands, lungs, pancreas, and all other remaining exocrine glands in the body. This in turn causes cystic fibrosis[4]. Also, certain genetic variants in the IL2RG gene impair the development of T and B cells which leads to a near complete failure of the immune system of those patients[5].

To truly understand those diseases, it is not enough to consider the consequences on the molecular level – like to identify the causal mutations in the genome or to characterize the misfolding of CFTR or the truncated Dystrophin. One also has to find out which cell types are affected and to decipher the specific consequences for the affected cell types. Only then it is possible to leverage this understanding to interfere with the disease.

That is why biologists have been trying to classify cells into certain types since the formation of cell theory in the middle of the 19th century[2]. The properties for classification have increased in complexity over the years as new technologies were developed. Using light microscopy and synthetic dyes, cells were mostly categorized by morphology, location and staining patterns[6,7]. As technologies such as monoclonal antibodies[8], immunohystochemistry[9] and flow cytometry[10] arose, it became clear that even cells with similar shapes can be significantly different in their expression pattern of surface proteins. The development of the Fluorescence *in situ* hybridization (FISH) method allowed distinguishing cells by allowing the sequence-specific detection of specific loci or transcripts[11]. Those molecular differences in RNA and protein signatures were found to often correspond to functional differences[7]. This array of technology allowed the classification of cells of for example the hematopoietic system or of the immune system at unprecedented depth[12,13]. Building on new 'omics' technology like antibody-based imaging, mass spectrometry-based proteomics, transcriptomics and systems biology, the Human Protein Atlas project is in the process of mapping all proteins in cells, tissues and organs in health and cancer[14]. Furthermore, the transformative Human Cell Atlas project holds promise to create a map of all cell types in the human body[15].

However, substantial obstacles persist on the road to a complete understanding of cell types and cell states. While many cells are classified according to their morphology such as hair, rod or cone

cells, others are classified according to their functionally such as retinal ganglion cells by their electrophysiological properties[7]. Some cells types are defined by the presence or absence of certain molecules on their surface like for example CD (cluster of differentiation) antigens[16]. Recent advances in gene expression profiling and single-cell technology have led to many cell types being defined according to their gene expression profile[17,18]. In addition to this lack of a systematic and comprehensive approach on how to define a cell type, there is also the conceptual problem of clearly distinguishing cell type and cell state. Cell type normally refers to a stable, non-changing state of a cell like for example a hepatocyte or a myocyte, while cell state implies a more volatile, unstable or reversible state of a cell like being at a certain stage of the cell cycle, immune cells performing different functions, non-genetic resistance states of cancer cells or a quiescent cells[15]. However, often the distinction between those concepts is unclear due to fluent transition between them. For instance, M1 and M2 polarization of macrophages leads them to have quite different function, morphology and molecular signatures[19], but is it different enough to classify them as separate cell types? Similarly, some cells might share important marker genes like CD4 but might still differ in their function or gene expression profile, therefore suggesting hidden diversity and subpopulations within this cell type[20]. One cell type might perform different functions and have different molecular signatures at a given time, but to what extent does this difference justify a distinction in cell type? Moreover, viewing cell types in the light of development significantly complicates the situation, since all cells arise from the same zygote and at a lot of stages during the process of “fate” decisions during differentiation, the classification into cell type or cell state becomes ambiguous[7].

The static description of cell states and types does not capture the full picture, because it cannot explain the dynamic and intricate changes in gene expression programs that underlie changes in cell state or type and are caused either intrinsically or by intercellular signaling processes or in response to changes in the extracellular environment. Conrad Waddington offered an elegant framework to metaphorically explain how cells “roll” along their path to differentiation on the epigenetic Waddington landscape (Figure 1)[21].

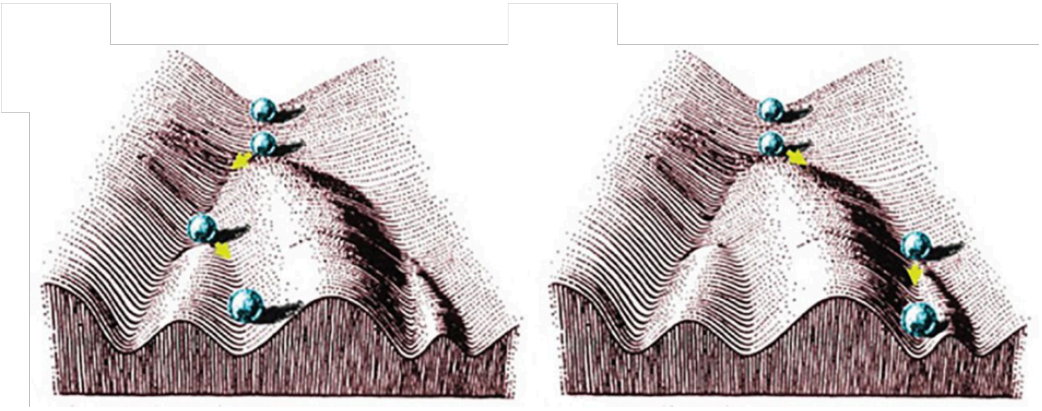


Figure 1 | Waddington landscape from Conrad Waddington's publication "The strategy of the genes. A discussion of some aspects of theoretical biology"[21]

Building on this framework, cell states and types can be viewed as points on a free energy landscape which is defined by gene expression state, analogous to the free energy landscape of protein folding or catalysis[22–24]. One might view the trajectory of a cell along its differentiation

path not as a smooth rolling downhill as Waddington postulated[21], but as a path with some bumps and hills along the way, which have to be overcome (Figure 2).

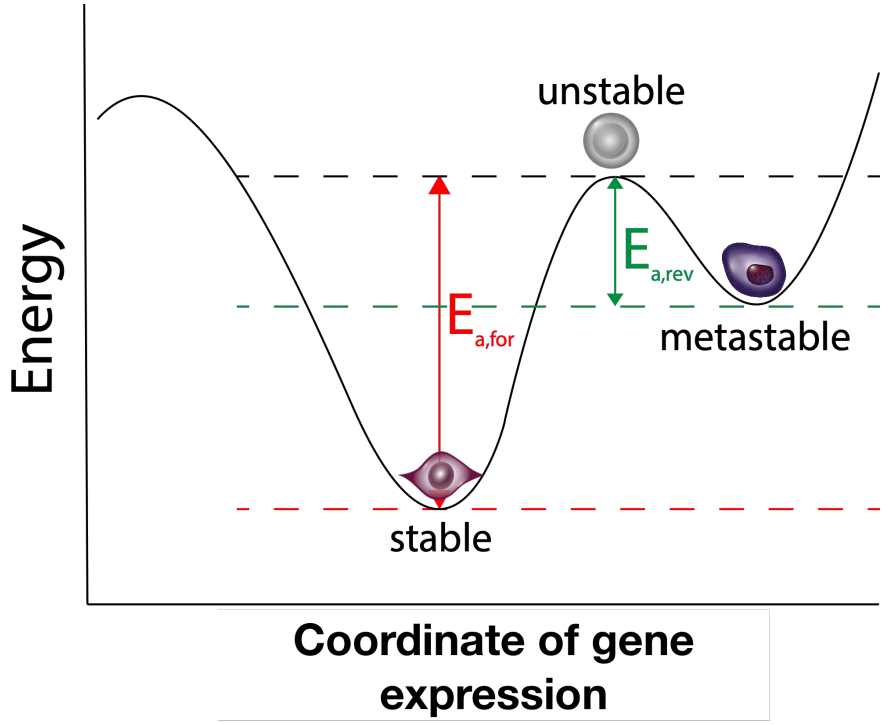


Figure 2 | Analogy of transitions of cell state or type to a free energy landscape

Those hills can be overcome by some sort of activation energy, such as a change in gene expression elicited by an extracellular or intrinsic event – for example a pioneering transcription factor or a master regulator[25,26]. It further suggests that some cell states are more stable than others and that metastable as well as unstable states are possible[27]. The very stable or 'attractor' states could correspond to what we call cell types due to their notion of persistence[28]. But also, unstable and metastable cell states which are lowly populated could be crucial to understand certain physiological processes and their dysregulation in disease. Unstable states are found on hills of the gene expression landscape and could represent intermediates between cell types. Metastable states on the other hand are found in local minima (valleys) and could represent volatile, relatively easily reversible states such as a non-genetic therapy-resistant state. For instance, in a tumor it is important to understand all possible cellular states in a tumor to target them appropriately and ultimately cure the disease[29]. Of particular interest in cancer is the persister cells state, a state in which cells can withstand drug concentrations many orders of magnitudes higher than their "normal" cancer cell counterpart[30]. This makes it highly relevant for therapeutic considerations, since it is hypothesized that those cells can lead to drug-resistant tumors and relapse after treatment[31]. Interestingly, this state is epigenetically determined, reversible and lowly populated in absence of drug, which is why studying it remains a challenge[30].

In many cases, the overexpression of one or more genes is enough to move a cell towards a certain state or even type. Overexpression of the master transcription factor MyoD in murine fibroblasts is sufficient to stably convert them to myoblasts[32]. Also, overexpression of the

developmental transcription factors Gata4, Mef2c, and Tbx5 in murine fibroblasts lead to transdifferentiation into cardiomyocyte-like cells[33]. Overexpressing Ptf1a in mouse embryonic fibroblasts transdifferentiates them into self-renewable induced neural stem cells[34]. Similarly, mouse embryonic fibroblasts (MEFs) can be directly reprogrammed into functional neurons by the ectopic overexpression of Ascl1, Brn2, and Myt1l[35]. It is even possible to transdifferentiate human fibroblasts into hepatocytes which are able fulfill drug-metabolizing functions[36]. The most famous example that ushered in this era of lineage reprogramming is the Nobel-prize winning discovery by Shinya Yamanaka that both mouse and human fibroblasts can be converted to induced pluripotent stem cells (iPSCs) with the introduction of four specific transcription factors: Oct4, Sox2, Klf4 and c-Myc[37,38]. Building on this groundbreaking realization, it has been shown that many different cell types can be reprogrammed to iPSCs[39].

The 'activation energy' to overcome the barrier between two cell types or states can also be overcome by the use of small molecules instead of ectopic overexpression. However, in most cases the small molecules influence cell state or type by modulating signaling pathways or altering DNA methylation or histone modifications[40,41]. For instance, chemically induced neural stem-like cells can be generated from mouse fibroblasts by treatment with 8 small molecules which influence important signaling cascades and the growth factor bFGF[42]. It has also been shown that MEFs can be reprogrammed to chemically induced iPSCs by using a combination of 7 small molecules at a higher efficiency than when using ectopic overexpression of Yamanaka factors[43]. Functional mouse neurons could also be generated from MEFs using only 4 chemical entities[44] and functional human neurons using a cocktail of 7 molecules[45].

These examples give big hope not only to navigate the vast and incredibly complex gene expression space of cell type or cell state transitions and to better understand them, but also for clinical uses to replace damaged tissues. For instance, first clinical applications to treat macular degeneration have shown positive results and improved vision in a patient[46].

However, most of the studies to discover factor or small molecule combinations to transition cells between states or types are quite laborious due to the large solution space of possible combinations. They are also often experience-driven and based on assumptions, which might inhibit unexpected discoveries. The still relatively low efficiency and often stochastic nature of most reprogramming and transdifferentiation approaches[47] could reflect our incomplete understanding of those processes on a cellular level. Moreover, unbiased high-throughput approaches are lacking. Furthermore, more subtle and reversible cell state transitions such as the persister cell state could be of high importance for our understanding of physiological and pathological processes. Those transitions are however much harder to delineate due to the high similarity and flux between states, and appropriate methods to study them are not available to date.

Technologies that allow measurements in single cells such as single-cell RNA-Seq[48], CyTOF[49], single-cell genome[50] and epigenome[51] sequencing potentially offer a way to overcome current limitations and to describe cell types and states. Their main advantage over bulk assays is that they overcome the fundamental constraint of averaging across all cells, thus allowing deeper insights into cell-to-cell variability[52]. However, some problems remain to be addressed. I will focus on the problems of scRNA-Seq, since it is not only the most relevant to this thesis, but

arguably the currently most advanced and widely used single-cell technology, and its problems extend for the most part to the other mentioned technologies. One of the most profound limitations of scRNA-Seq is the cost. While there are already many methods available and the price range is large, the cost of a scRNA-Seq workflow including sequencing costs are still quite high ranging from about $700 to about $11,000 depending on the method[53]. Furthermore, for the droplet-based methods commercial microfluidic equipment is necessary, which inflates the costs of such protocols[53]. It should be noted that sequencing costs are the main factor for the price issue, since cost of library preparations are normally very low ranging from about $0.1 to $30[53]. Related to the cost issue, the throughput of scRNA-Seq workflows is still relatively limited. Despite significant advances in library generation and microfluidics technology over the last years, most datasets encompass thousands to hundred-thousands of cells[54] (moores law figure). Of course, as sequencing costs decrease and with new technological advances, it should be feasible to sequence millions of single cells routinely within the next decade[55]. For instance, a recent publication profiled about 2 million cells in a single experiment[56]. Since in a single-cell RNA-Seq experiment a library corresponds to a single cell, library amplification can vary significantly between cells and lead to large amplification biases[55]. Furthermore, due to the low amount of RNA used for library generation, transcripts might 'drop out' due to capture or amplification errors[57]. Another problem is that since there are no replicates *per se* (each cell is measured only once), there is a high risk of confounding batch effects[55]. Furthermore, scRNA-Seq currently yields only a relatively shallow insight into the transcriptome of a cell – of course depending on the sequencing depth to some degree – since lowly expressed genes often 'drop out' due to poor capture or amplification, or due to temporal fluctuations in gene expression[57]. This leads to a median of about 5000 to 9000 genes per cell being quantified[53]. Since solid tissues have to be disaggregated into a single cell suspension to perform a scRNA-Seq experiment, it is important that the procedure to do so do not introduce some sort of bias[52]. For example, dissociation by use of enzymes such as collagenase should not lead to the lysis of a specific cell type in the sample. Furthermore, the state of the cells should remain unaltered throughout the procedure as well as possible. Finally, scRNA-Seq data is not only many orders of magnitude more dimensional than bulk RNA-Seq data, but it is also much more variable due to technical and biological factors[52]. In addition to the missing replicates/normalization problem, these problems highlight the need for gold-standard of bioinformatic methods and pipelines such as normalization methods for the analysis of scRNA-Seq data[58].

To delineate causal factors that control the state and fate of a cell, current approaches mainly rely on comparing gene expression profiles between cell types or states[25,37,59]. Researchers try to identify genes that are only expressed in certain cell types and not in others, or at specific stages during development, to infer their importance for a certain cell type or state. Subsequently, the comparison between gene expression patterns is then used to identify factors or factor combinations that drive cell state transitions through overexpression experiments, on a trial and error basis[33,35,37]. These approaches yield valuable information and have been successful in identifying numerous factors and factor combinations. However, they might miss unexpected factors and can be relatively labor intense or even infeasible due to the vast solution space of potential combinations[37]. Moreover, they leave researchers without a systemic understanding of the entirety of factors contribute to the cell type or state transition. Functional genetic

approaches based on the CRISPR/Cas9 system could offer a reasonable scalability ideal for systematic interrogation of cell state transitions. While knock-out or knock-down of a gene by for example using CRISPR wild-type[60] or CRISPRi[61] systems, respectively, often helps overcome barriers or stabilize cell states and types through autoregulatory feedback loops[62], it has not yet been shown to drive a cell type or state transition on its own. Therefore, gain of function approaches like CRISPRa might be a more reasonable choice for perturbing the state or type of a cell since a state transition is almost always accompanied by gain of properties. For example, a recent study identified factors that drive both mouse ESCs and fibroblasts toward a neural fate by deploying a high-throughput CRISPR activation screening approach[63]. However, the path of differentiation of ESCs into various cell types has few obstacles, so transitions between states and types that do not have a pluripotent state remains unaddressed. Despite the scalability and programmability of the CRISPR technologies, the binding of Cas9 has been shown to be influenced significantly by chromatin state[64], and the extent of gene activation by CRISPRa is quite variable from gene to gene and has an upper limit dependent on the activator architecture[65]. Ectopic overexpression of open reading frames on the other hand does not offer the same programmability and throughput of CRISPR systems. The constituents of ORF libraries are normally cloned individually and their quality control and handling are more complex than that of pooled single guide RNA (sgRNA) libraries due to the increased size and the variation in size[66]. Nevertheless, overexpression of ORFs offers crucial benefits the current CRISPRa systems do not offer, such as strong expression of the gene of interest, possible silencing of the introduced gene and control over the sequence[67].

There is a clear need for a new and unbiased method to identify causal factors which transition cells between states and types, to ultimately help understand those transitions in a systematic manner. The objective of this thesis was to help overcome current limitations by laying the groundwork for the development of such a novel method.

To achieve the research objective, I devised a high-throughput screening method to decipher cell state transitions. The method combines a novel genetic screening paradigm called Phenosudoku, an ORF-based lentiviral library and a targeted bulk RNA-Sequencing approach.

Pooled genetic screening methods do not really allow for high-content readouts, require a high cell number and cannot be used easily in combination with complex ORF libraries[68]. Arrayed methods on the other hand can be quite costly and time-consuming[68]. The Phensoduoku screening paradigm combines the economic efficiency and wide net cast by pooled screening method while retaining most of the advantages of screens performed in an arrayed format. It is carried out in random sub pools of perturbagens whose composition is known before the screen (and needs to be determined only once). Through complex mathematical modeling the causative agents (i.e. an ORF or single guide RNA) of a phenotype can be found out after assigning each well either a ‘hit’ or ‘no hit’ label. This paradigm will be a powerful tool for discovery when combined with the developed sequencing approach and complex ORF libraries.

Many cell state transitions are subtle and differ only slightly from each other, which makes their study challenging. While it is possible to distinguish drastically different cell states from each other by qPCR, this method is limited to only a few genes and quickly finds its limits once the cell states in question do not have genes that are completely differently expressed (=signature

genes). Moreover, performing qPCRs for many wells of a screen is a very laborious and almost infeasible task. To overcome these limitations, I chose to adapt a single-cell RNA-Seq method for my purposes (CEL-Seq2[69]) and use it as a bulk RNA-Seq method, which would allow me to robustly detect and quantify cell states using an analog signature of not only a few, but many transcripts (Figure 3).

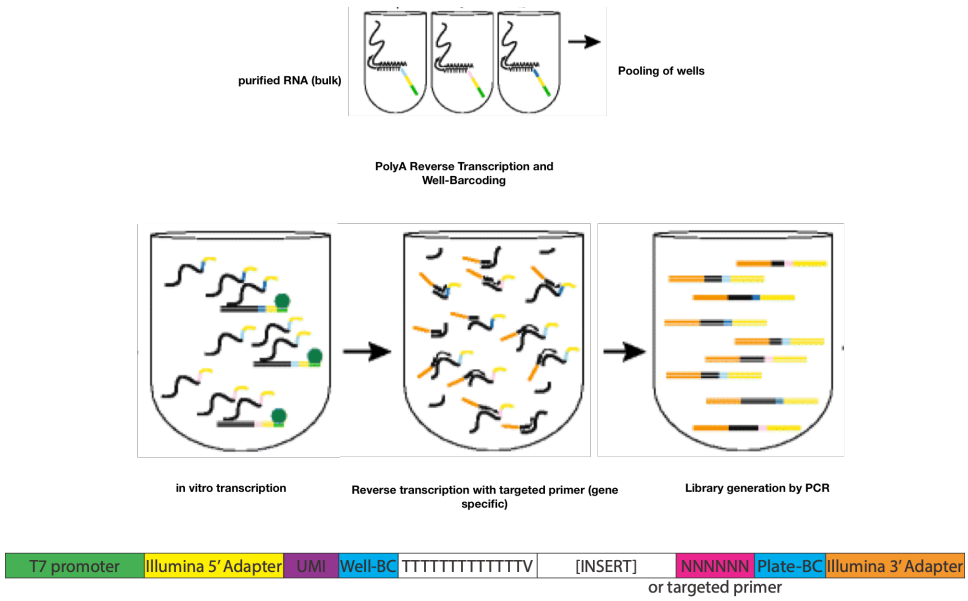


Figure 3 | Modification of the CEL-Seq2 single-cell RNA-Seq method[69] (Figure modified from original publication)

I also chose the transcriptome instead of the genome as a proxy for cell state, since the transcriptome is altered more drastically and quickly during cell state transitions as well as because it is easier to measure and offers higher multiplexing. To test my method, I chose the transition from fibroblast to reprogrammed state based on the Nobel Award discovery by Shinya Yamanaka in 2006[37]. In this system, fibroblast cells can be converted to induced pluripotent stem cells with the introduction of four specific transcription factors: Oct4, Sox2, Klf4 and c-Myc. Furthermore, the parental and the destination state are extremely different in their transcriptome in this case and the factors underlying this transition are quite well studied. This "positive control" assay would enable me to benchmark and evaluate performance prior to applying it to the more unknown mechanisms of cell state transition.

In this thesis I laid the groundwork for a proof of concept screen which will enable evaluating performance prior to applying the method to the more unknown mechanisms of cell state transitions such as the one from cancer to persister cell[30]. To do so, I identified a subset of genes, which are not or very lowly expressed in the parental cell state (i.e. the fibroblast state) and relatively high expressed in the destination cell state (i.e. the hiPSC state). Then, I designed and targeted primers against those genes and integrate them into the modified RNA-Seq protocol based on CEL-Seq2[69]. By assessing the expression of these genes by bulk RNA-Seq, I should be able to identify successful transitions to the parental cell state, even at very low frequencies, and with high accuracy. Furthermore, I evaluated the use of the DASH method[70] to deplete in this case 'irrelevant' transcripts which are abundant in fibroblasts (the initial cell state). This allowed me to increase sequencing depth while at the same time reducing irrelevant data. For the ORF

library, I explored pooled Gateway cloning reactions and ways to generate pooled ORF libraries and to convert them into the Phenosudoku format. Furthermore, after generating several ORF libraries in a pooled and Phenosudoku format and quality controlled them. The generated libraries are readily applicable to subsequent genetic screens dissecting other transitions of cell state or type.

The results of this endeavor offer a foundation to further develop and optimize a tool to decipher cell state transitions. This tool could have an important impact on our understanding of cell state transitions and transform the way we investigate them.

2 Materials and Methods

2.1 Cloning of vectors

The cloning of all vector plasmids was designed using Snapgene (v4.3.7). Most plasmids were cloned by HiFi DNA Assembly by mixing the vector backbone and the insert(s) in a 1:2 molar ratio and then adding deionized water and NEBuilder® HiFi DNA Assembly Master Mix (NEB, # E2621L) to obtain a 10 µl reaction (Table 1). If more than 2 inserts were used, the backbone and inserts were mixed in an equimolar ratio. The reaction was then incubated at 48°C for 15 minutes when 1 or 2 fragments were being inserted or for 60 minutes when 3–5 fragments were being inserted. Table 2 summarizes the cloning of vectors by HiFi DNA assembly. Vector plasmid VMS010 (pL-sEF1a-T7-att-V5-IRES-Puro) was cloned by introducing a V5-tag and three consecutive stop-codons by PCR and circularizing the resulting product in a KLD (NEB, M0554S) reaction (Table 3).

Component	**1-2 Inserts Assembly**	**3-5 Inserts Assembly**
DNA Molar Ratio	Vector: Insert = 1:2	Vector: Insert = 1:1
Total Amount of Fragments	0.03–0.2 pmol	0.2–0.5 pmol
	X µl	X µl
NEBuilder HiFi DNA Assembly Master Mix	5 µl	5 µl
Deionized H2O	5-X µl	5-X µl
Total Volume	10 µl	10 µl

Table 1 | HiFi reaction protocol

Vector plasmid	**Vector plasmid backbone**	**Restriction enzymes/primers for linearization**	**Insert(s)**	**PCR primers used**
VMS002 (pSicoR-sEF1a-O-K-M-S-aP2A-Blast)	MTM_672 (pSicoR-sEF1a-mCherry)	BmtI and PciI	hOKMS-aP2A cDNA amplified by PCR from FUW-tetO-hOKMS	OMS004 and OMS007
			aP2A-Blast amplified from MTM_277	OMS005 and OMS006
VMS005	V18034	XbaI and NsiI	Sox2 from FUW-tetO-hOKMS	OMS022 and OMS023

(pLN-sEF1a-SKM-Blast)	(pLN-sEF1a-spcas9-mtagbfp-blast)			
			hKM from FUW-tetO-hOKMS	OMS020 and OMS021
			E2A-Blast from MTM_1073	OMS019 and OMS024
VMS008_no_T7 (pL-sEF1a -att-IRES_Puro)	pSuperInf-IRES	StuI	Puro from pBID-Dest-pre-GFP-Puro	OMS029 and OMS030
VMS008 (pL-sEF1a-T7-att-IRES-Puro)	VMS008_no_T7 ((pL-sEF1a -att-IRES-Puro)	OMS031 and OMS032	NA	NA
VMS009 (pL-sEF1a-T7-att-STOP-IRES-Puro)	VMS008 (pL-sEF1a-T7-att-IRES-Puro)	NotI and PciI	STOP-Ires from VMS008	OMS035 and OMS036
			attR1-attR2 from VMS008	OMS033 and OMS034
VMS011 (pENTR11-attL1-GFP_nostop-attL2)	pENTR11	OMS45 and OMS046	GFP_nostop from MTM277	OMS050 and OMS051
VMS012 (pENTR11-attL1-Oct4_nostop-attL2)	pENTR11	OMS45 and OMS046	Oct4_nostop from FUW-tetO-hOKMS	OMS047 and OMS048
VMS013 (pENTR11-attL1-Oct4_stop-attL2)	pENTR11	OMS45 and OMS046	Oct4_stop from FUW-tetO-hOCT4	OMS047 and OMS049
VMS015 (pLN-sEF1a-SKM-GFP)	V18034 (pLN-sEF1a-spCas9-mTagBFP-Blast)	XbaI and NsiI	hSKM from VMS005	OMS054 and OMS055
			GFP from V18033	OMS056 and OMS057
VMS017 (pL-sEF1a-T7-att-STOP-IRES-mCherry)	VMS009 (pL-sEF1a-T7-att-STOP-IRES-Puro)	PciI and Bsu36I	mCherry from pSuperInf-IRES-mCherry	OMS058 and OMS059

Table 2 | Summary of HiFi reactions

Component	**Volume**
PCR Product	1 µl
KLD Reaction Buffer (2X)	5 µl

KLD Enzyme Mix (10X)	1 µl
Nuclease-free Water	3 µl
Total Volume	10 µl

Table 3 | KLD reaction protocol

Table 4 shows the PCR protocol used to amplify the DNA necessary for the cloning of the vector plasmids and Table 5 and Table 6 show the thermocycling conditions used for standard and 2-step PCRs, respectively. In Table 7 the PCR reactions, which were performed on a T100 Thermal Cycler (BioRad, # 1861096) are shortly recapitulated. A list of primers and oligonucleotides essential for the cloning steps, as well as a detailed description of the vector plasmids, can be found in the appendix of this thesis. Each of the plasmids that were cloned contained an ampicillin, kanamycin or spectinomycin resistance cassette, thus facilitating amplification of the constructs in *E. Coli*.

Component	**Input [µl]**	**Final Concentration**
Q5 High-Fidelity 2X Master Mix OR Phusion® Hot Start Flex 2X Master Mix	12.5	1X
10 µM Forward Primer	1.25	0.5 µM
10 µM Reverse Primer	1.25	0.5 µM
Template DNA	1	< 250 ng for Phusion < 1000 ng for Q5
Nuclease-Free Water	to 25	

Table 4 | PCR reaction protocol

STEP	**TEMP**	**TIME**
Initial Denaturation	98°C	30 seconds
25–35 Cycles	98°C	10 seconds
	50–72°C for Q5 45–72°C for Phusion	30 seconds
	72°C	see PCR table
Final Extension	72°C	2 minutes for Q5 10 minutes for Phusion
Hold	4°C	

Table 5 | Thermocycling condition for standard PCR

Step	Temperature	Time
Initial Denaturation	98°C	30 seconds
25–35 Cycles	98°C	10 seconds
	72°C	see PCR table
Final Extension	72°C	2 minutes
Hold	4°C	

Table 6 | Thermocycling conditions for 2-step PCR

insert	Primer1	Primer2	Template	Program	Annealing Temp. [°C]	Extension time [s]	Product
aP2A-Blast	OMS005	OMS006	MTM_277	Phusion Std.	67.7	15.00	yes
hOKMS-aP2A	OMS007	OMS004	FUW-tetO-hOKMS	Phusion Std.	64.4	90.00	no
hOKMS-aP2A	OMS007	OMS004	FUW-tetO-hOKMS	Phusion Std.	56 - 66	100.00	yes, more product at higher temperatures
hOKMS-aP2A	OMS007	OMS004	FUW-tetO-hOKMS	2step Q5 [1 + 25 cycles]	69 + 72	150.00	no
hOKMS-aP2A	OMS007	OMS004	FUW-tetO-hOKMS	Phusion Std.	69-65	100.00	small amount and at wrong size
hOKMS-aP2A	OMS007	OMS004	FUW-tetO-hOKMS	2step Q5 [3 + 25 cycles]	65 + 72	240.00	yes
aP2A-Blast	OMS005	OMS006	MTM_277	2step Q5 [1 + 25 cycles]	67.7 + 72	30.00	yes
Sox2	OMS022	OMS023	FUW-tetO-hOKMS	Phusion Std.	65.8	15.00	yes
hKM	OMS020	OMS021	FUW-tetO-hOKMS	Phusion Std.	67.3	45.00	yes

E2A-Blast	OMS019	OMS024	MTM_1073	Phusion Std.	66.4	10.00	yes
Puro	OMS029	OMS030	pBID-Dest-pre-GFP-Puro	Phusion Std.	66	15.00	yes
T7	OMS031	OMS032	VMS008_no_T7	Phusion Std.	66.1	135.00	yes
attR1-attR2	OMS033	OMS034	VMS008	2step Q5 [3 + 25 cycles]	65 + 72	60.00	yes
STOP-Ires	OMS035	OMS036	VMS008	Q5 Std.	66	30.00	yes
KLD	OMS039	OMS040	VMS008	2step Q5 [3 + 25 cycles]	65 + 72	270.00	yes
pENTR 11	OMS045	OMS046	pENTR11	Q5 Std.	67	65.00	yes
GFP_nostop	OMS050	OMS051	MTM_277	2step Q5 [1 + 25 cycles]	69 + 72	30 | 40	yes
Oct4_nostop	OMS047	OMS048	hOKMS	2step Q5 [3 + 25 cycles]	70 + 72	40 | 50	yes
Oct4_stop	OMS047	OMS049	FUW-tetO-hOCT4	2step Q5 [3 + 25 cycles]	71 + 72	40 | 50	yes
hSKM	OMS054	OMS055	VMS005	2step Q5 [2 + 25 cycles]	69 + 72	130.00	yes
GFP	OMS056	OMS057	V18033	2step Q5 [2 + 25 cycles]	50 + 72	30.00	yes
mCherry	OMS058	OMS059	pSuperInf-IRES-mCherry	Q5 Std.	70	60.00	yes

Table 7 | Summary of PCR reactions

To extract a specific PCR or restriction digest product a 1% agarose gel (w/v; in TAE-buffer) and a PowerPac™ Basic Power Supply (BioRad, #1645050) were used to perform gel electrophoresis. DNA was visualized using SYBR® Safe DNA Gel Stain (Invitrogen, # S33102). The desired band was then cut out and the DNA was extracted from the gel with a NucleoSpin® Gel and PCR Clean-up kit (Macherey-Nagel, #740609.250).

To amplify cloned plasmids, 50 µl of a competent *Escherichia Coli* strain dependent on the plasmid were transformed with 5-10 µl of the DNA HiFi assembly reaction by heat shock for 30 seconds at 42°C. When amplifying Gateway vectors containing the ccdB gene for example, ccdB resistant One Shot® ccdB Survival™ 2 T1R Chemically Competent Cells (Invitrogen, #A10460) were used. The bacteria were plated on LB-Agar plates containing Ampicillin (100 µg/ml), Kanamycin (50 µg/ml) or Spectinomycin (50 µg/ml), depending on the resistance cassette contained in the plasmid. Bacterial colonies were picked and grown in suspension culture in LB medium containing Ampicillin (100 µg/ml), Kanamycin (50 µg/ml) or Spectinomycin (50 µg/ml), depending on the resistance cassette contained in the plasmid. After 12-16 hours, amplified plasmids were extracted by performing plasmid minipreps with a QIAprep Spin Miniprep Kit (QIAGEN, #27106). The identity of all plasmids was verified by Sanger sequencing.

Component	**Input**
Entry clone	50 ng 1-7 µl
Destination vector	150 ng 1 µl
TE buffer, pH 8.0	to 8 µl
LR Clonase™ II Plus enzyme mix	2 µl
Proteinase K [to terminate reaction]	1 µl

Table 8 | Gateway LR reaction protocol

Entry clone	**Destination vector**	**Expression clone**
VMS011	VMS009	VMS026 pL-sEF1a-T7-GFP_nostop-STOP-IRES-Puro
VMS011	VMS010	VMS027 pL-sEF1a-T7-GFP_nostop-V5-IRES-Puro
VMS013	VMS009	VMS028 pL-sEF1a-T7-Oct4_stop-STOP-IRES-Puro
VMS013	VMS010	VMS029 pL-sEF1a-T7-Oct4_stop-V5-IRES-Puro
VMS012	VMS009	VMS030 pL-sEF1a-T7-Oct4_nostop-STOP-IRES-Puro
VMS012	VMS010	VMS031 pL-sEF1a-T7-Oct4_nostop-V5-IRES-Puro

Table 9 | Summary of vectors cloned by LR reaction

Some vector plasmids were cloned in a Gateway LR recombination carried out as specified in Table 8. Table 9 summarizes the LR reactions performed. Therefore, the 50 ng of the Entry clone and 150ng of the Destination vector were mixed with TE buffer of pH 8.0. Then, LR Clonase™ II Plus enzyme mix (Thermo Fisher Scientific, #12538120) was thawed on ice for 2 minutes. After adding 2 µl of LR Clonase™ II Plus enzyme mix, the reaction was mixed well by vortexing briefly and then spun down to collect the liquid at the bottom of the tube. The sample was incubated at

25°C for 1 hour. The reaction was terminated by adding 1 µl of Proteinase K and incubating at 37°C for 10 minutes. To amplify the resulting expression vector, I transformed 1 µl of the reaction into 50 µl of ccdB-sensitive *E. Coli* (DH5α E. Coli, QB3 MacroLab) by heat shock for 30 seconds at 42°C. I plated the bacteria on LB-Agar plates containing Ampicillin (100 µg/ml). I then picked bacterial colonies and grew them in suspension culture in LB medium containing Ampicillin (100 µg/ml). After 12-16 hours, I extracted the amplified plasmid by performing plasmid minipreps with a QIAprep Spin Miniprep Kit (QIAGEN, #27106). The identity of all plasmids was verified by Sanger sequencing.

2.2 Generation of ORFeome libraries

2.2.1 Generation of a lentiviral hORFeome v8.1 expression vector library

The 12787 Gateway ENTR vectors from the human ORFeome v8.1 (hORFeome v8.1) were obtained from the Human ORFeome V8.1 collection from the DNASU plasmid repository. The pool of all plasmids contained in this collection was prepared by Kol Jia Yong, PhD by growing mixed cultures of about 100 constructs, extracting the plasmids, normalizing the amount and pooling the resulting subpools into a plasmid pool containing all ENTR vector plasmids of the h ORFeome v8.1 collection. To generate a pool of lentiviral expression vectors, the ENTR vectors of the human ORFeome v8.1 collection were cloned into a pLEX307 Gateway destination vector in a pooled Gateway LR recombination reaction. Therefore, 1000 ng of the ENTR vector pool and 1000 ng of the pLEX307 destination vector were mixed with TE buffer of pH 8.0 (Table 10). As a negative control, 1000 ng of the pLEX307 destination vector were mixed with TE buffer of pH 8.0. All subsequent steps were also performed for the negative control unless otherwise indicated. Invitrogen LR Clonase™ II Plus enzyme mix (Thermo Fisher Scientific, #12538120) was thawed on ice for 2 minutes. After adding 20 µl of LR Clonase™ II Plus enzyme mix, the reaction was mixed well by vortexing briefly and then spun down to collect the liquid at the bottom of the tube. The sample was incubated at 25°C for 16 hours. The reaction was terminated by adding 10 µl of Proteinase K and incubating at 37°C for 15 minutes.

Component	**Input**
hORFeome V8.1 Entry clone pool	1000 ng
pLEX307 destination vector	1000 ng
TE buffer, pH 8.0	to 80 µl
LR Clonase™ II Plus enzyme mix	20 µl
Proteinase K [to terminate reaction]	10 µl

Table 10 | Summary of pooled LR reaction to generate lentiviral hORFeome v8.1 library

The DNA was then cleaned and concentrated by ethanol precipitation. Therefore, 2 µl tRNA at a concentration of 1 µg/µl (Thermo Fisher Scientific, #501005189) were added to the finished recombination reaction. Then, 11.2 µl Sodium Acetate (3M, pH 5.2, Thermo Fisher Scientific, # R1181) and 300 µl ethanol (≥99.8%, Sigma-Aldrich, 51976-500ML-F) were added as well in that order. The reaction was mixed by vortexing and then frozen overnight at -20°C. Then, the DNA was pelleted by centrifugation for 10 minutes at 20.000 rcf and 4°C and the supernatant was decanted. The pellet was washed with 70% ethanol at 4°C and the tube centrifuged again for 10 minutes at 20.000 rcf and 4°C. After decanting the supernatant, the pellet was dried by leaving

the tube open for about 15 minutes. The dried pellet was then resuspended in 40 µl of TE buffer of pH 8.0.

The concentrated plasmid library was then desalted using a Micro Bio-Spin™ P-30 Gel Column (Bio-Rad, #7326223). Therefore, the column was emptied by gravity flow and then centrifuged for 2 minutes at 1.000 rcf. Then, the column was refilled with 500 µl deionized water and subsequently emptied by centrifugation for 1 minute at 1.000 rcf. This step was repeated 5 times. Finally, the sample containing the plasmid library was applied to the column and the column was centrifuged for 4 minutes at 1.000 rcf. This procedure leads to an inflation of the sample volume by a factor of about 1.43 to 2.

To amplify the resulting expression vector library and assess the result of the pooled recombination reaction, the desalted sample was electroporated into MegaX DH10B™ T1R Electrocomp™ Cells using pre-cooled Gene Pulser/MicroPulser Cuvettes (BioRad #1652089), a Gene Pulser Xcell™ (BioRad, #165-2660) and an exponential decay pulse at 2000 V, 25 uF and 200 Ohm. Directly after electroporation, the bacteria were transferred to 5 ml prewarmed recovery media (Invitrogen, #46-0706) and allowed to recover by shaking at 37°C for 60 minutes. Then, 50 µl of a 1:5.000, a 1:50.000 and a 1:500.000 dilution was plated on LB-Agar plates containing Ampicillin (100 µg/ml) to assess success and efficiency of the LR reaction. The rest of the recovered bacteria were transferred to 250 ml LB media and grown at 37°C for 16 hours at 200 rpm in a bacterial shaker.

The next day, the amplified plasmid library was extracted by performing a plasmid maxi prep with a commercial plasmid isolation kit (QIAGEN, #12163). Moreover, colonies were counted on both the plates where the dilutions had been plated and the negative control plate. A certain number of bacterial colonies from the dilution plates were picked and grown in suspension culture in LB medium containing Ampicillin (100 µg/ml). After 12-16 hours, the plasmids were extracted from those colonies and their identity was determined by Sanger sequencing to estimate the successful recombination rate.

To quality control the qualitative and – most of all – quantitative composition of the ENTR and expression plasmid pools of the hORFeome v8.1, those pools were subjected to Tn5-based library generation and subsequent next generation sequencing. To do so, cDNA libraries were generated from the ENTR and expression plasmid pools using the Nextera DNA Flex Library Prep Kit (Illumina, #20018705). Details of this library generation can be found in Table 11. The sequencing libraries were then normalized to 1 nM, pooled and quality controlled using a DNA High Sensitivity Chip (Agilent, #5067-4626) and an Agilent 2100 Bioanalyzer. Then, they were sequenced on an Illumina NovaSeq sequencing platform (Table 12).

Sample ID	Concentration of plasmid pool [ng/µl]	DNA_input [µl]	DNA input [ng]	i5	i7
hORFeome v8.1 pENTR223	87	5.17	449.79	H505	H701
hORFeome v8.1 pLEX307	68	6.62	450.16	H517	H702

Table 11 | Library generation of hORFeome v8.1 ENTR and expression vector pools

Library ID	Concentration [ng/µl]	Mean size [bp]	Concentration [nM]	Reads [#]
hORFeome v8.1 pENTR223	9.9	758	19.7	2x151
hORFeome v8.1 pLEX307_1000	8.3	694	18.1	2x151

Table 12 | Sequencing of libraries generated from hORFeome v8.1 ENTR and expression vector pools

The resulting sequencing data was analyzed using various software tools. The reads were automatically demultiplexed according to their sample indices by BaseSpace. Reads were then quality trimmed using Trim Galore (v0.4.5) and their quality was assessed using FastQC (v0.11.7). The reference genome containing the ORFs of the hORFeome V8.1 as well as the gff file containing information about those ORFs were generated using custom python scripts (python v3.6). Then, the reads were aligned against the reference genome locally using bowtie2 (v2.3.5) and the resulting SAM file was sorted and converted into BAM format using SAMtools (v1.9). The BAM file was then translated into a file containing read counts per ORF using HTSeq (v0.11.2). Those read numbers were then normalized according to feature size and total reads by conversion to TPM values using a custom python script. Resulting data was analyzed using Microsoft Excel (v16.31), R (v3.6) and custom python scripts.

2.2.2 Generation of a lentiviral expression vector ORF library containing epigenetic modifiers and transcription factors

To reduce library size and adjust library composition to the goals of this thesis, the Epigenetic Factors and Epigene Histones Collection (565 ORFs) and the 90/90 Human ORFeome V1 Transcription Factor Subcollection (1500 ORFs) were obtained from DNASU as bacterial glycerol stocks. The size distribution and molar amounts of individual components of the combined Epigenetic Modifiers and Transcription Factor (=EMTF) ORF library were determined using custom python scripts. Furthermore, vectors were separated into two groups – one containing all plasmids with a Kanamycin resistance cassette and the other containing all plasmids with a Spectinomycin resistance cassette. Glycerol stocks were thawed on ice, mixed by careful pipetting and 20 µl of each clone was transferred into a pool of their respective group. About half of the volume of each pool (11.5 ml of the Kanamycin pool, 8.8 ml of the Spectinomycin pool) was frozen at -80°C as a backup. The remaining volume of the pools was pelleted separately by centrifugation at 6.000 rcf for 15 minutes at 4°C and plasmid pools were extracted using a commercial maxi prep kit (QIAGEN, #12163).

Component	Amount [ng]	fmole	Amount [µl]
ENTR clones pool [EMTF]	1000ng	382.22	32.3
VMS009	2200ng	396.88	5.05
LR Clonase™ II Plus enzyme mix	NA	NA	20
TE buffer, pH 8.0	NA	NA	42.7

Table 13 | Summary of pooled LR reaction to generate lentiviral EMTF library

The pooled recombination reaction was performed for with the amounts depicted in Table 13 and incubated at 25°C for 24 hours. The ethanol precipitation was performed analogously to the

one described for hORFeome V8.1. Subsequently, the recombined plasmid pool was resuspended in 30 µl deionized H2O. Then, it was desalted and electroporated into MegaX DH10B™ T1R Electrocomp™ Cells and amplified in liquid culture as described previously. 240 ml of the liquid culture were pelleted by centrifugation at 3000 rcf and 4°C for 10 min, resuspended in LB media containing 25% Glycerol, frozen in a bath of ethanol and dry ice and stored at -80°C. The next day, the glycerol stock was thawed to determine its CFU/ml by plating several dilutions onto LB plates containing Ampicillin (100 µg/ml).

Corresponding volume [µl]	Dilution factor	Colonies	CFU/ml
1	1.00E+03	too many to count accurately	NA
0.1	1.00E+04	too many to count accurately	NA
0.01	1.00E+05	1488	1.49E+08
0.001	1.00E+06	213	2.13E+08
0.0001	1.00E+07	16	1.60E+08
0.00001	1.00E+08	1	1.00E+08

Table 14 | Estimation of CFU/ml in glycerol stock of EMTF library

Then, the glycerol stock was thawed to again determine its CFU/ml by plating several dilutions onto LB plates containing Ampicillin (100 µg/ml) and to convert the library into a Phenosudoku format (Table 14). In this format, each well of a 96-well plate receives a random sub selection of the total pool of library plasmids in the form of one CFU (or bacteria). The number of constructs per well is determined by the concentration of CFUs in the initial dilution of the pool. To improve plasmid yields, Terrific Broth media (Sigma-Aldrich, #T0918) was used. By diluting the glycerol stock by a factor of 5.280.000 in Terrific Broth, the number of constructs was estimated to be around 37 per well. This was reasonably close to the number determined retrospectively by the plated dilutions: around 30 CFU/well (Table 15). Under constant stirring, the diluted pool of bacteria containing constructs of the EMTF library was aliquoted into 6 96-well plates at 1.2 ml per well. The plates were then incubated for 25 hours at 37°C and 250 rpm in a bacterial shaker. The liquid cultures in the 96-well plates were then plasmid prepped using a commercial 96 well mini prep kit (Macherey-Nagel, #740625.1). This was done according to the protocol of the manufacturer using a vacuum manifold, with all optional washing steps, drying of the NucleoSpin® Plasmid Binding Plate by centrifugation at 4122 rcf for 15 minutes and a final elution by centrifugation at 4122 rcf for 3 minutes after applying 100 µl Elution Buffer AE to each well.

Dilution factor	Colonies	CFU/ml (glycerol stock)	CFU/well	Volume plated
5,280,000	16	7.04E+07	16	1.2 ml
5,280,000	28	1.48E+08	28	50 µl
528,000	333	1.76E+08	33.3	50 µl

Table 15 | Estimation of CFU/well of EMTF library in Phenosudoku format

The pool of EMTF ENTR vectors as well as the pool of EMTF expression vectors were converted into sequencing libraries and sequenced analogously to the pools of the hORFeome v8.1 library. Details of this library generation can be found in Table 16. The sequencing libraries were then

normalized to 1 nM and pooled. Then, they were sequenced on an Illumina MiniSeq sequencing platform (Table 17).

Sample ID	Concentration of plasmid pool [ng/µl]	DNA_input [µl]	DNA input [ng]	i5	i7
EMTF expression plasmids pool	54	3.703703704	200	510	706
EMTF ENTR plasmids pool	68	2.941176471	200	522	712

Table 16 | Library generation of EMTF ENTR and expression vector pools

Library ID	Concentration [ng/µl]	Concentration [nM]	Concentration for loading [pM]	PhiX [%]	PhiX 1.3pM [µl]	MiniSeq Kit
EMTF expression plasmids pool	8.18	20.657	1.3	4%	13.2	300 cycles, High Output
EMTF ENTR plasmids pool	11.9	30.05	1.3	4%	13.2	300 cycles, High Output

Table 17 | Sequencing of libraries generated from EMTF ENTR and expression vector pools

The resulting sequencing data was analyzed using various software tools analogously to the sequencing data obtained from hORFeome V8.1 sequencing libraries.

To quality control the qualitative and quantitative composition of the wells, 24 of the wells were subjected to Tn5-based library generation and subsequent next generation sequencing. To do so, cDNA libraries were generated using the Nextera DNA Flex Library Prep Kit (Illumina, #20018705). Details of this library generation can be found in Table 18. The sequencing libraries were then pooled and sequenced on an Illumina MiniSeq sequencing platform (Table 19).

Plate ID	Well	Concentration of plasmid pool [ng/µl]	DNA_input [µl]	DNA input [ng]	i5	i7
EMTF_P1	A01	151	1	151	H503	H705
EMTF_P1	A02	107	1	107	H503	H706
EMTF_P1	A03	113	1	113	H503	H707
EMTF_P1	A04	104	1	104	H503	H710
EMTF_P1	A05	40	3	120	H503	H711
EMTF_P1	A06	60	2	120	H503	H714
EMTF_P1	A07	87	1.5	130.5	H505	H705
EMTF_P1	A08	74	1.5	111	H505	H706

EMTF_P1	A09	80	1.5	120	H505	H707
EMTF_P1	A10	92	1.5	138	H505	H710
EMTF_P1	A11	78	1.5	117	H505	H711
EMTF_P1	A12	92	1.5	138	H505	H714
EMTF_P3	D01	145	1	145	H506	H705
EMTF_P3	D02	152	1	152	H506	H706
EMTF_P3	D03	114	1	114	H506	H707
EMTF_P3	D04	125	1	125	H506	H710
EMTF_P3	D05	100	1	100	H506	H711
EMTF_P3	D06	117	1	117	H506	H714
EMTF_P3	D07	91	1.5	136.5	H517	H705
EMTF_P3	D08	145	1	145	H517	H706
EMTF_P3	D09	119	1	119	H517	H707
EMTF_P3	D10	110	1	110	H517	H710
EMTF_P3	D11	117	1	117	H517	H711
EMTF_P3	D12	139	1	139	H517	H714

Table 18 | Library generation of 24 EMTF Phenosudoku wells

Library ID	**Concentration [ng/µl]**	**Concentration [nM]**	**Concentration for loading [pM]**	**PhiX [%]**	**PhiX 1.3pM [µl]**	**MiniSeq Kit**
4P1_A1-12_4P3_D1-12	11.4	28.79	1.3	5%	26.4	300 cycles high output

Table 19 | Sequencing of libraries generated from EMTF Phenosudoku wells

The resulting sequencing data was analyzed using various software tools. The reads were automatically demultiplexed according to their sample indices by BaseSpace. To streamline analysis, a bash script was generated to analyze all 24 wells in parallel. Reads were then quality trimmed using Trim Galore (v0.4.5) and their quality was assessed using FastQC (v0.11.7). The reference genome containing the ORFs of the EMTF library as well as the gff file containing information about those ORFs were generated using custom python scripts (python v3.6). Then, the reads were aligned against the reference genome locally using bowtie2 (v2.3.5) and the resulting SAM file was sorted and converted into BAM format using SAMtools (v1.9). The BAM file was then translated into a file containing read counts per ORF using HTSeq (v0.11.2). Those read numbers were then normalized according to feature size and total reads by conversion to TPM values using a custom python script. Resulting data was analyzed using Microsoft Excel (v16.31), R (v3.6) and custom python scripts.

2.2.3 Generation of a lentiviral sub library containing 241 transcription factor ORFs (TF241)

A more focused sub library of transcription factor ORFs was then generated by individual LR recombination steps. First, 260 glycerol stocks containing selected ENTR vector plasmids of transcription factors were picked and transferred to wells in 96 square-well blocks (Macherey-Nagel, 740476.24). Those wells contained 1.2 ml of Terrific Broth growth media with either Kanamycin (50 µg/ml) or Spectinomycin (50 µg/ml), and the liquid cultures were then incubated for 24 hours at 37°C and 200 rpm. The liquid cultures in the 96-well plates were then plasmid prepped using a commercial 96 well mini prep kit (Macherey-Nagel, #740625.1). This was done according to the protocol of the manufacturer using a vacuum manifold, with all optional washing steps, drying of the NucleoSpin® Plasmid Binding Plate by centrifugation at 4122 rcf for 15 minutes and a final elution by centrifugation at 4122 rcf for 3 minutes after applying 80 µl Elution Buffer AE to each well. The extracted plasmids were used in 260 individual Gateway LR recombination reactions in wells of a 96-well plate using quantities outlined in Table 20.

Component	**Input**
Entry clone	~65 ng X µl
Destination vector	65 ng 0.5 µl
LR Clonase™ II Plus enzyme mix	0.5 µl
TE buffer, pH 8.0	to 2.5 µl
Proteinase K [to terminate reaction]	0.5 µl

Table 20 | Summary of LR reactions to generate lentiviral expression vectors

The reactions were incubated at 25°C for 2 hours, then the reaction was stopped by adding Proteinase K and an incubation step at 37°C for 10 minutes. 2 µl of each reaction were then transformed into 40 µl STBL3 E. Coli cells (QB3 MacroLab Berkeley) by heat shock for 10 seconds. After a brief incubation on ice for 2 minutes, 140 µl of recovery media (Invitrogen, #46-0706) were added to each well. The plates were incubated at 37°C for 30 minutes at 200 rpm in a bacterial shaker and then 180 µl from each well was plated onto 6x8 LB Agar plates containing Carbenicillin (100 µg/ml). Those 6x8 LB Agar plates were incubated at 37°C for 14 hours and then assessed for results of the Gateway LR reactions. 241 of the 260 LR reactions yielded colonies. From those apparently successful reactions, colonies of equal, medium size and normal morphology were picked into respective wells of 96 square-well blocks (Macherey-Nagel, 740476.24) containing 1.2 ml of terrific broth and Carbenicillin (100 µg/ml). Those liquid cultures were then incubated at 300 rpm and 37°C for 18 hours. Then, 100 µl of each well were added to the pool of expression vectors consisting of 241 transcription factor ORFs. 3 ml of this pool were used to extract the plasmid library by using a commercial plasmid mini prep kit. 20 ml of this pool were used to generate a glycerol stock (25% glycerol) for further analysis and use.

Corresponding volume [µl]	**Dilution factor**	**Colonies [#]**	**CFU/ml**
10	100	too many to count	NA

5	200	too many to count	NA
0.5	2,000	too many to count	NA
0.05	20,000	886	1.77E+07
0.005	200,000	81	1.62E+07

Table 21 | Estimation of CFU/ml in glycerol stock of TF241 ibrary

Dilution factor	**colonies**	**CFU/ml**	**CFU/well**
650.000	26	1.69E+07	26
130.000	146	1.90E+07	29.2
1.300.000	18	2.34E+07	36

Table 22 | Estimation of CFU/well of 241 library in Phenosudoku format

The amount of CFU/ml of the glycerol stock containing the expression vector plasmids was determined by plating several dilutions onto LB Agar plates containing Carbenicillin (100 µg/ml) and counting them (Table 21). Then, the glycerol stock was thawed to again determine its CFU/ml by plating several dilutions onto LB plates containing Carbenicillin (100 µg/ml) and to convert the library into a Phenosudoku format as described for the EMTF library pool. By diluting the glycerol stock by a factor of 650.000 in Terrific Broth, the number of constructs was estimated to be around 20 per well. This was reasonably close to the number determined retrospectively by the plated dilutions: around 30 CFU/well with a range of probably between 26 and 36 CFU/well (Table 22). Under constant stirring, the diluted pool of bacteria containing constructs of the TF241 library was aliquoted into 4 96-well plates at 1.2 ml per well. The plates were then incubated for 25 hours at 37°C and 300 rpm in a bacterial shaker. The liquid cultures in the 96-well plates were then plasmid prepped using a commercial 96 well mini prep kit (Macherey-Nagel, #740625.1). This was done according to the protocol of the manufacturer using a vacuum manifold, with all optional washing steps, drying of the NucleoSpin® Plasmid Binding Plate by centrifugation at 4122 rcf for 15 minutes and a final elution by centrifugation at 4122 rcf for 3 minutes after applying 100 µl Elution Buffer AE to each well.

Plate ID	**Concentration of plasmid pool [ng/µl]**	**DNA input [µl]**	**DNA input [ng]**	**Nextera CD index well**
TF241_P1_C01	78.77	2	157.54	D01
TF241_P1_C02	84.04	2	168.08	D02
TF241_P1_C03	92.58	2	185.16	D03
TF241_P1_C04	72.98	2	145.96	D04
TF241_P1_C05	53.2	2	106.4	D05
TF241_P1_C06	63.28	2	126.56	D06
TF241_P1_C07	81.18	2	162.36	D07
TF241_P1_C08	84.37	2	168.74	D08
TF241_P1_C09	87.88	2	175.76	D09
TF241_P1_C10	74.06	2	148.12	D10
TF241_P1_C11	66.99	2	133.98	D11

TF241_P1_C12	74.96	2	149.92	D12
TF241_P2_F01	90.94	2	181.88	E01
TF241_P2_F02	67.02	2	134.04	E02
TF241_P2_F03	101.86	2	203.72	E03
TF241_P2_F04	87.04	2	174.08	E04
TF241_P2_F05	65.06	2	130.12	E05
TF241_P2_F06	79.26	2	158.52	E06
TF241_P2_F07	88.72	2	177.44	E07
TF241_P2_F08	87.05	2	174.1	E08
TF241_P2_F09	87.35	2	174.7	E09
TF241_P2_F10	87.39	2	174.78	E10
TF241_P2_F11	53.54	2	107.08	E11
TF241_P2_F12	86.87	2	173.74	E12
TF241_culture_pool	101	2	202	A02

Table 23 | Library generation of TF241 expression vector pool and 24 Phenosudoku wells

To quality control the qualitative and quantitative composition of the initial library of expression vectors and the wells of the plates in Phenosudoku format, the initial library and 24 of the wells were subjected to Tn5-based library generation and subsequent next generation sequencing. To do so, cDNA libraries were generated using the Nextera DNA Flex Library Prep Kit (Illumina, #20018705). Details of this library generation can be found in Table 23. The sequencing libraries were then pooled in molar percentages of 72.73% for the pool of wells and 27.27% for the pool of expression vectors. The resulting library pool was sequenced on an Illumina MiniSeq sequencing platform (table).

Library ID	**Conc. [ng/µl]**	**Conc. [nM]**	**Loading conc. [pM]**	**PhiX [%]**	**PhiX 1.3pM [µl]**	**MiniSeq Kit**
TF241_Phenosudoku_wells_24_pooled	11.1	28.030303	1.35	1	5	300 cycles, High Output
TF241_expression_pool	9.2	23.23232	1.35	1	5	300 cycles, High Output

Table 24 | Sequencing of libraries generated from TF241 expression vector pool and Phenosudoku wells

The resulting sequencing data was analyzed analogously to the approach described for the EMTF library and Phenosudoku wells.

2.3 Spiking in human embryonic stem cells into fibroblasts

For the spike-in experiments of human embryonic stem cells (hESCs) into fibroblasts, human BJ fibroblasts (ATCC, CRL-2522) were cultured with DMEM medium (Corning, #10-017-CV) supplemented with 10% FBS (Corning, #35-072-CV), 1% Penicillin-Streptomycin solution (Corning, #30-002-CI), 1× nonessential amino acids (Corning, #25-025-CI), 1x sodium pyruvate

(Corning, #25-000-CI), and 0.06 mM 2-Mercaptoethanol (Thermo Fisher Scientific, #21985-023). After a few passages, they were seeded into wells of a 96-well plate at a concentration of 33,000 cells per well. The plates were incubated for about 8 hours to let the cells attach and then, the media was removed and hESCs at varying amounts were spiked into the wells (Figure 4). The hESCs were received from Gopika Nair, PhD (Dr. Matthias Hebrok lab) and adjusted to the appropriate amounts by the process of serial dilution. Subsequently, the plates were centrifuged at 300 rcf and the media was aspirated. Then, the total RNA was isolated from the cells in each well using the MagMAX™-96 Total RNA Isolation Kit (Thermo Fisher, #AM1830) according to the protocol of the manufacturer. RNA was eluted in 20 µl elution buffer. RNA concentration was measured on a Thermo Scientific™ NanoDrop 2000 and the quality and quantity of the RNA samples was also assessed using an Agilent RNA 6000 Nano Kit (Agilent, #5067-1511) and an Agilent 2100 Bioanalyzer.

	1	2	3	4	5	6	7	8	9	10	11	12
A	5 hESCs	5 hESCs	5 hESCs	50 hESCs	50 hESCs	50 hESCs	500 hESCs	500 hESCs	500 hESCs	5000 hESCs	5000 hESCs	5000 hESCs
B	5 hESCs	5 hESCs	5 hESCs	50 hESCs	50 hESCs	50 hESCs	500 hESCs	500 hESCs	500 hESCs	5000 hESCs	5000 hESCs	5000 hESCs
C	Fibroblasts	Fibroblasts	Fibroblasts	Fibroblasts	Fibroblasts	hESCs						
D												
E												
F												
G												
H												

Figure 4 | Plate layout of hESC spike-in experiments

2.4 RNA-Sequencing using Quantseq 3' mRNA Seq

The extracted RNA from the spike-in experiments was used to generate RNA-Seq libraries. Therefore, the QuantSeq 3' mRNA-Seq Library Prep Kit for Illumina (FWD) (Lexogen, #015) was used according to the manufacturer's recommendations. Table 25 outlines the library generation procedure. The PCR Add-on Kit for Illumina (Lexogen, #020) was used to quantify the libraries before amplification. The libraries were quality controlled using a DNA High Sensitivity Chip (Agilent, #5067-4626) and an Agilent 2100 Bioanalyzer (Table 26). Then, the libraries were pooled in an equimolar manner and sequenced on a MiniSeq sequencing platform (Table 27).

Sample	RNA conc. [ng/µl]	Volume [µl]	RNA [ng]	PCR cycles	i7 index
5 hESCs	51.3	4	205.2	14	A1
5 hESCs	43	4	172	14	B1
50 hESCs	39.3	4	157.2	14	C1
50 hESCs	41.64	4	166.56	14	D1
500 hESCs	40.76	4	163.04	14	E1
500 hESCs	38.01	4	152.04	14	F1
5000 hESCs	43.33	4	173.32	16 (eluted in 3x volume!)	G1
5000 hESCs	48.54	4	194.16	14	H1
hESCs only	226	0.71	160.46	14	A2

Fibroblasts only	43.51	4	174.04	14	B2

Table 25 | Summary of generation of RNA-Seq libraries from spike-in experiment

Sample condition	**Molarity [pmol/L]**
5 hESCs	22,676.75
5 hESCs	29,994.75
50 hESCs	22,980
50 hESCs	31,834.5
500 hESCs	38,513.75
500 hESCs	35,186.5
5000 hESCs	40,148.5
5000 hESCs	16,806.5
hESCs only	42,542
Fibroblasts only	34,655.5

Table 26 | Molarity of RNA-Seq libraries generated using the QuantSeq protocol

Library	**Loading conc. [pM]**	**PhiX [%]**	**PhiX 1.8pM [µl]**	**MiniSeq Kit**
Pool of QuantSeq libraries	1.8	1	5	150 cycles, High Output

Table 27 | Sequencing of RNA-Seq libraries generated using the QuantSeq protocol

The sequencing data was analyzed using various software tools. First, the raw sequencing data created by the sequencing run was converted from binary base call (BCL) format to FASTQ format using Bcl2fastq (v2.20.0.422). After trimming the reads to remove reads or sequence stretches containing base calls with low quality or adapters using Trim galore (v0.4.5) and cut to a length of 60 base pairs using cutadapt (v2.1 with Python 3.6.8), the quality of the reads was assessed using FastQC (v0.11.7). Then, genome indices were generated using STAR (v2.5.3a) and the sequencing reads were aligned to the genome and counted. The raw read counts were then analyzed using the DESeq package (v1.34.1) in R by finding differentially expressed genes and generating heatmaps and PCA plots.

2.5 RNA-Sequencing using a modified CEL-Seq2 method

In principle, CEL-Seq2[69] is a single cell RNA-Seq library generation method, but it is highly customizable, so I modified it for my purposes.

2.5.1 RNA-Seq library generation and sequencing of hESC spike-ins using targeted primers v1

Libraries were prepared from the RNA extracted from the hESC spike-in experiment. Due to the procedure of targeted library preparation, spike-in samples were prepared separately from samples containing only hESCs and only Fibroblasts. First, the library generation for the spike-in samples is described. The details for the initial steps can be found in Table 28.

Sample	RNA conc. [ng/µl]	Volume [µl]	RNA [ng]	RT Primer [#]	RT Primer [well]
5 hESCs	51.3	3.5	179.55	1	A1
5 hESCs	43	4	172	4	A4
50 hESCs	39.3	4	157.2	5	A5
50 hESCs	41.64	4	166.56	9	A9
500 hESCs	40.76	4	163.04	10	A10
500 hESCs	38.01	4	152.04	23	B11
5000 hESCs	43.33	4	173.32	25	C1
5000 hESCs	48.54	4	194.16	26	C2
5 hESCs	24.4	4	97.6	31	C7
50 hESCs	20.7	4	82.8	46	D10

Table 28 | Summary of first RT reaction

First, the 2x primer mix for each well was prepared. Therefore, the reverse transcription (RT) primer (20ng/µl) was mixed with the dNTP solution (4mM) in a 1:1 ratio to generate the 2x primer mix (10ng/µl RT primer; 2mM dNTP solution). After mixing 4 µl of RNA sample and 4 µl of the 2x primer mix, 1.2 µl of this mixture was then transferred to a new plate, incubated for 5 minutes at 65°C and then immediately moved on ice. Then it was centrifuged at 4000 rcf for 1 minute to collect the contents at the bottom of the well and stored on ice. 0.8 µl of the reverse transcription (RT) reaction mix (Table 29) were then added to each well and the samples were incubated for 1 hour at 42°C in the thermal cycler with the temperature of the lid set to 50°C. The reaction was then heat inactivated by incubation at 70°C for 10 minutes with the temperature of the lid set to 105°C.

First Strand buffer	0.4 µl
DTT 0.1M	0.2 µl
RNase Inhibitor	0.1 µl
Superscript II	0.1 µl

Table 29 | Composition of RT reaction mix

The samples were then moved to ice for about 2 minutes. 10 µl of the second strand reaction mix (Table 30) was added, the samples were flicked and centrifuged at 4,000 rcf for 1 minute to collect the contents at the bottom of the wells and the reaction was incubated at 16^{o}C for 2 hours in a thermal cycler with the lid turned off.

RNAse free Water	7 µl
Second strand buffer	2.31 µl
dNTP	0.23 µl
Ligase	0.08 µl
E. coli DNA Polymerase	0.3 µl
RNaseH	0.08 µl

Table 30 | Composition of the second strand reaction mix

The cDNA was then cleaned up using Ampure XP DNA beads (Beckman Coulter, A63880). All samples going into the same in vitro transcription (IVT) reaction were pooled in a 1.5 ml DNA LoBind microcentrifuge tube (Eppendorf, #0030108051) since they were already barcoded at this point and 120 µl Ampure XP DNA beads were added to 100µl of the sample pool. The sample pool was mixed by vortexing thoroughly and incubated at room temperature for 15 minutes. The tube was then placed on a DynaMag™-2 magnet (Thermo Fisher Scientific, #12321D) for 5 minutes until the liquid appeared clear. The supernatant was removed except for 5 µl and 200 µl freshly prepared 80% ethanol (Sigma-Aldrich, #E7023-500ML) were added. The sample was incubated for 30 seconds and then the supernatant was removed. This washing step was repeated. The beads were air dried for 15 minutes or until they were completely dry. The beads were resuspended in 20 µl water by pipetting the entire volume up and down 10 times. After incubation for 2 minutes the tube was placed on the magnetic stand for 5 minutes. The supernatant containing the cleaned-up cDNA was transferred to a new PCR tube (Eppendorf, EP0030124332-1PAK).

9.6 µl of the cDNA were transferred to a new PCR tube and mixed with 14.4 µl of the IVT mix (Table 31). The reaction was mixed by vortexing and the contents of the tube were collected at the bottom by centrifugation in a microcentrifuge. The IVT reaction was incubated in the thermal cycler at 37°C for 13 hours with the lid temperature set to 70°C. At the end of the incubation, the sample was kept at 4°C, thus ensuring the stability of the antisense RNA (aRNA) for several hours.

A	2.4 µl
G	2.4 µl
C	2.4 µl
U	2.4 µl
10x T7 buffer	2.4 µl
T7 Enzyme	2.4 µl

Table 31 | Composition of IVT mix

To hydrolyze primers and unincorporated nucleotides, 9.6 µl ExoSAP-IT PCR Product Clean-Up Reagent (Affymetrix, #78200) were added to the IVT reaction. The reaction was incubated in the thermal cycler for 15 minutes at 37°C with the lid temperature set to 45°C. Subsequently, the aRNA was fragmented to an average size of about 500 base pairs by incubation for 15 minutes at 80°C. However, the resulting size distribution was quite flat and broad.

The aRNA was then cleaned up. An equal volume of RNAClean XP beads (Beckman Coulter, #A63987) prewarmed to room temperature were added to the sample. After incubation at room temperature for 10 minutes, the PCR tube was placed on a 96-well magnetic stand for 5 minutes. All but 5 µl of the supernatant was removed and 200 µl freshly prepared 70% ethanol was added. The sample was incubated for 30 seconds and then the supernatant was removed. This wash was repeated two additional times. The beads were then air dried for 15 minutes or until completely dry and resuspended in 30 µl H2O. The sample was mixed by pipetting up and down ten times. After an incubation step of 2 minutes, the tube was placed on the magnetic stand for 5 minutes

and the supernatant was transferred to a new PCR tube. The aRNA clean-up was repeated with a final elution in 10 µl.

Additionally, the protocol was repeated up until this point without fragmentation by heat and the aRNA was fragmented by sonication using a Covaris S220 Focused-ultrasonicator and a microTube AFA 6x16mm (Covaris, #520045) under specified conditions (Table 32). The fragmented aRNA was then concentrated by repeating the clean-up protocol using RNAClean XP beads.

Volume [µl]	Temparature [°C]	Duty Factor [%]	Water level	Power [W]	cpb	Time(s)
120	8-10	10	12-14	175	200	10s, 20s, 30s, 45s
120	8-10	5	12-14	105	200	10s, 20s
55	6-8	10	10	75	1,000	30s

Table 32 | aRNA fragmentation conditions

The outcome of the fragmentation and the aRNA amount and quality were determined by running a part of the sample on the Agilent 2100 Bioanalyzer using an RNA Pico Kit (Agilent, #5067-1513).

Five library-RT reactions were set up with the details summarized in Table 33. For each reaction, 5 µl of the solution containing the specified amount of aRNA in ng were mixed with 1 µl of targeted primers v1 and 0.5 µl dNTPs (10mM). After incubation for 5 minutes at 65°C, the reaction was quick-chilled on ice. Then, 3.5 µl of the library-RT mix (Table 34) was added to the reaction. The reaction was incubated at 42°C for 2 minutes. Then, 0.5 µl SuperScript II Reverse Transcriptase (Invitrogen, 18064014) was added, the sample was mixed by vortexing and the contents of the tube were collected at its bottom by centrifugation for 5 seconds in a microcentrifuge. The library-RT reaction was incubated at 42°C for 1 hour in the thermal cycler with the lid temperature set to 50°C. The reaction was inactivated by incubation at 70°C for 15 minutes with the lid temperature set to 105°C.

Fragmentation condition	aRNA input for library-RT [ng]	Targeted_primers_v1 [pmol]
Chemical fragmentation	250	0.02
Chemical fragmentation	250	0.002
Covaris 120 µl 175W 15s 8-10°C	125	0.001
Covaris 55 µl 75W 30s 6-8°C	85	0.001
Unfragmented	125	0.002

Table 33 | Summary of library-RT reactions

First Strand buffer	2 µl

DTT 0.1M	1 µl
RNaseOUT	0.5 µl

Table 34 | Composition of library-RT mix

The number of PCR cycles necessary for library amplification was determined by subjecting a part of the sample to qPCR (Table 35). Real-time qPCR was performed on an ABI PRISM 7900HT Sequence Detection System with the program specified in Table 36.

Component	**Input [µl]**
Phusion® High-Fidelity PCR Master Mix with HF Buffer (2x)	12.5
10 µM RP1 Primer	1
10 µM RPIX Primer	1
Library-RT reaction	1
2.5x SYBR in DMSO	1
PCR water	8.5

Table 35 | qPCR reaction protocol

STEP	**TEMP**	**TIME**
Initial Denaturation	98°C	30 seconds
20 -25 Cycles	98°C	10 seconds
	60°C	30 seconds
	72°C	30 seconds
Final Extension	72°C	10 minutes
Hold	4°C	

Table 36 | qPCR thermocycling program

Component	**Input [µl]**
Phusion® High-Fidelity PCR Master Mix with HF Buffer (2x)	25
10 µM RP1 Primer	2
10 µM RPIX Primer	2
Library-RT reaction	X
PCR water	21-X

Table 37 | PCR reaction protocol for amplifying the libraries

Once cycle numbers were determined, libraries were amplified by PCR with the appropriate cycle number using a thermal cycler (Table 37). Those PCR reactions are summarized in Table 38.

Fragmentation condition	aRNA input for library-RT [ng]	Targeted primers v1 [pmol]	PCR cycles	Library -RT input [µl]	PCR volume [µl]
Chemical fragmentation	250	0.02	9	5	50
Chemical fragmentation	250	0.002	13	5	50
Covaris 120µl 175W 15s 8-10°C	125	0.001	15	5	50
Covaris 55µl 75W 30s 6-8°C	85	0.001	15	10	50
Unfragmented	125	0.002	13	5	50

Table 38 | Conditions for library amplification by PCR

PCR reactions were cleaned up using AMPure XP Beads. 1 volume (50 µl) of AMPure XP Beads were added to the sample and the sample was mixed thoroughly by pipetting up and down ten times. After incubation at room temperature for 15 minutes the tube was placed on a DynaMag™-2 magnet (Thermo Fisher Scientific, #12321D) for 5 minutes until the liquid appeared clear. 95µl of the supernatant were removed and 200 µl freshly prepared 80% ethanol (Sigma, #E7023-500ML) were added. The sample was incubated for 30 seconds and then the supernatant was removed. This washing step was repeated. The beads were air dried for 15 minutes or until they were completely dry. The beads were resuspended in 25 µl water by pipetting the entire volume up and down 10 times. After incubation for 2 minutes the tube was placed on the magnetic stand for 5 minutes. The supernatant containing the cleaned-up cDNA was transferred to a new PCR tube. The clean-up procedure was repeated by adding 25 µl AMPure XP Beads and eluting in a final volume of 10 µl.

To assess amount and quality of the libraries, 1 µl was analyzed using the Agilent 2100 Bioanalyzer and a DNA High Sensitivity Chip (Agilent, #5067-4626). Libraries were stored at -20°C until sequencing.

The library which underwent chemical fragmentation with an input of 250 ng aRNA and 0.002 pmol targeted primer v1 into the library-RT reaction was sequenced on the MiniSeq platform (Table 39).

Library description	Concentration for loading [pM]	PhiX [%]	MiniSeq Kit
Chemical Fragmentation, 250 ng aRNA, 0.002 pmol targeted primers v1	1.8	20%	300 cycles, High Output

Table 39 | Sequencing of RNA-Seq library generated using modified CEL-Seq2 protocol and targeted primers v1

The sequencing data was analyzed using various software tools. Reads were quality trimmed using Trim Galore (v0.4.5) and their quality was assessed using FastQC (v0.11.7). The reference

genome was generated using STAR (v2.5.3a). Read 1 was trimmed to a length of 12 base pairs and Read 2 was trimmed to a length of 99 base pairs using cutadapt (v2.1 with Python 3.6.8). Then, the Celseq2 software tool from the Yanai lab (v0.5.3.3, https://github.com/yanailab/celseq2) was used to demultiplex the samples according to their indices, align them to the reference genome, generate read counts and collapse UMIs. The raw read counts were then analyzed using the DESeq package (v1.34.1) in R. Resulting data was analyzed using Microsoft Excel (v16.31), R (v3.6) and custom python scripts.

2.5.2 RNA-Seq library generation and sequencing of hESCs and fibroblasts using random hexamer primers

RNA Seq libraries were prepared from the RNA extracted from samples containing only hESCs and only BJ fibroblasts (Table 40).

Sample	RNA conc. [ng/µl]	Volume [µl]	RNA [ng]	RT Primer [#]	RT Primer [well]
hESCs	45	4	180	17	B5
Fibroblasts	43.51	4	174.04	37	D1
hESCs	45	4	180	1	A1
Fibroblasts	43.51	4	174.04	4	A4
hESCs	45	4	180	5	A5
Fibroblasts	43.51	4	174.04	9	A9

Table 40 | Summary of first RT reaction

The libraries were generated as described in the modified CEL-Seq2 protocol for the spike-in samples. Some changes were made to the protocol, which are described below.

The aRNA was fragmented by chemical fragmentation and sonication as indicated in Table 41. Furthermore, a random hexamer (rHex) RT primer was used to reverse transcribe the aRNA. In this RT reaction, 1 µl of dNTP solution was used instead of 0.5 µl. After addition of the RT primer and the dNTP solution, the reaction was incubated at 25°C for 2 minutes. Then, 1 µl of Superscript II reverse transcriptase was added, and the sample was incubated at 25°C for 10 minutes. The reaction was then incubated at 42°C for 1 hour as previously described.

Fragmentation condition	aRNA input for library-RT [ng]	Library-RT primer (rHex) [ng]	PCR cycles	Library-RT input [µl]	PCR volume [µl]
Chemical fragmentation	250	1000	4	5	50
Chemical fragmentation	250	250	5	5	50
Covaris 175W 15s 120 µl 8-10°C	125	250	6	5	50
Covaris 75W 30s 55 µl 6-8°C	90	250	7	10	50

Table 41 | aRNA fragmentation conditions

The library which underwent sonication with an input of 90 ng aRNA and 250 ng rHex RT primer into the library-RT reaction was sequenced on the MiniSeq platform (Table 42).

Library ID	Concentration [nM]	Concentration for loading [pM]	PhiX [%]	MiniSeq Kit
90ng aRNA (75W 40s 55ul 6-8deg) + 0.25ug CelSeq-RT (rHex)	17	1.8	20%	150 cycles, High Output

Table 42 | Sequencing of RNA-Seq library generated using modified CEL-Seq2 protocol and rHex primers

2.5.3 RNA-Seq library generation and sequencing of hESC spike-ins using targeted primers v2 and random hexamer primers

RNA-Seq libraries were generated from RNA extracted from the spike-in samples using the modified CEL-Seq2 protocol as previously described with some changes (Table 43).

Sample	RNA conc. [ng/µl]	RT Primer [#]	RT Primer [well]
5 hESCs	28.65	1	A1
5 hESCs	39.6	4	A4
5 hESCs	35.85	5	A5
50 hESCs	42.75	9	A9
50 hESCs	31.55	10	A10
50 hESCs	38.05	23	B11
500 hESCs	35.85	25	C1
500 hESCs	38.5	26	C2
500 hESCs	32.45	31	C7
5000 hESCs	35.6	46	D10
5000 hESCs	37.4	54	E06
5000 hESCs	38.95	68	F08

Table 43 | Summary of first RT reaction

For the first RT reaction, RT primer (50ng/µl) was mixed with the dNTP solution (10mM) in a 1:1 ratio to generate a 5x primer mix (25ng/µl RT primer; 5mM dNTP solution). 1 µl of this 5x RT primer mix was then mixed with 4 µl RNA and 1.3 µl of this RNA/Primer/dNTP mix was added to the RT reaction. Superscript IV Reverse Transcriptase (Invitrogen, #18090050) was used instead of Superscript II Reverse Transcriptase (Table 44). The aRNA was fragmented under conditions indicated in Table 45.

First Strand buffer	0.4 µl
DTT 0.1M	0.1 µl
RNase Inhibitor	0.1 µl
Superscript IV	0.1 µl

Table 44 | Condition of RT mix with Superscript IV

Volume [µl]	aRNA input [ng]	Temparature [°C]	Duty Factor [%]	Water level	Power [W]	cpb	Time [seconds]
55	1,000	6-8	10	10	50	200	30
55	1,000	6-8	10	10	75	200	40

Table 45 | Conditions of aRNA fragmentation

Two library-RT reactions were set up with the details summarized in Table 46. For the reaction using targeted primers, 5 µl of the solution containing the specified amount of aRNA in ng were mixed with 1 µl of targeted primers v2 and 1 µl dNTPs (10mM). After incubation for 5 minutes at 65°C, the reaction was quick-chilled on ice. Then, 3 µl of the library-RT mix (Table 47) were added to the reaction. 1 µl SuperScript IV Reverse Transcriptase was added, the sample was mixed by vortexing and the contents of the tube were collected at its bottom by centrifugation for 5 seconds in the microcentrifuge. The library-RT reaction was incubated at 55°C for 1 hour in the thermal cycler with the lid temperature set to 50°C. The reaction was inactivated by incubation at 80°C for 10 minutes with the lid temperature set to 105°C.

Fragmentation condition	aRNA input for library-RT [ng]	Targeted primers v2 [pmol]	Library-RT (rHex) primers [ng]
55 µl, 7°C, 50W, 30s	486	15	NA
55 µl, 7°C, 75W, 40s	441	NA	250

Table 46 | Summary of library-RT reactions

First Strand buffer	2 µl
DTT 0.1M	0.5 µl
RNaseOUT	0.5 µl

Table 47 | Composition of library-RT mix using Superscript IV

For the reaction using random hexamer primers, 5 µl of the solution containing the specified amount of aRNA in ng were mixed with 1 µl of library-RT random hexamer primers and 1 µl dNTPs (10mM). After incubation for 5 minutes at 65°C, the reaction was quick-chilled on ice. Then, 3 µl of the library-RT mix (Table 47) were added to the reaction. 1 µl SuperScript IV Reverse Transcriptase was added, the sample was mixed by vortexing and the contents of the tube were collected at its bottom by centrifugation for 5 seconds in the microcentrifuge (??). After an incubation step at 23°C for 10 minutes, the library-RT reaction was incubated at 55°C for 1 hour in the thermal cycler with the lid temperature set to 50°C. The reaction was inactivated by incubation at 80°C for 10 minutes with the lid temperature set to 105°C.

qPCR reactions were scaled down to a reaction volume of 12.5 µl and performed on an Applied Biosystems® QuantStudio® 5 Real-Time PCR System. The PCR reactions were scaled down to a reaction volume of 25 µl (Table 48).

Fragmentation condition	aRNA input for library-RT [ng]	PCR cycles	Library-RT input [µl]	PCR volume [µl]	RPIX primer
55 µl, 7°C, 50W, 30s	486	8	5	25	RPI8
55 µl, 7°C, 75W, 40s	441	6	5	25	RPI2

Table 48 | Conditions for library amplification by PCR

One library was sequenced on the MiniSeq platform (Table 49) and sequencing data was analyzed as previously described.

Library ID	Concentration Qubit [ng/µl]	Concentration Bioanalyzer (150-800bp) [nM]	Conc. for loading [pM]	PhiX [%]	MiniSeq Kit
Spike-ins, 55 µl, 7°C, 50W, 30s + targeted primers v2	2.26	12.8	1.8	1%	150 cycles, High Output

Table 49 | Sequencing of RNA-Seq library generated using modified CEL-Seq2 protocol and targeted primers v2

2.5.4 RNA-Seq library generation and sequencing of hESCs and fibroblasts using targeted primers v2 and random hexamer primers

RNA-Seq libraries were again generated from the RNA extracted from samples containing only Fibroblasts or hESCs using the modified CEL-Seq2 protocol analogously to the library preparation described for the refined approach for the spike-in samples (Table 50). Unless otherwise specified, the same protocol was used.

Sample	RNA conc. [ng/µl]	RT Primer [#]	RT Primer [well]
hESCs	22	1	A1
Fibroblasts	27.2	4	A4
hESCs	22	5	A5
Fibroblasts	18.95	9	A9
hESCs	22	10	A10
Fibroblasts	21	23	B11

Table 50 | Summary of first RT reaction

The details for the library-RT step are specified in Table 51and the details of the PCR amplification in Table 52.

Fragmentation condition	aRNA input for library-RT [ng]	Targeted primers v2 [pmol]	Library-RT (rHex) primers [ng]

55 µl, 7°C, 50W, 30s	435	15	NA
55 µl, 7°C, 75W, 40s	440	NA	250

Table 51 | Summary of library-RT reactions

Fragmentation condition	aRNA input for library-RT [ng]	PCR cycles	Library-RT input [µl]	PCR volume [µl]	RPIX primer
55 µl, 7°C, 50W, 30s	435	8	5	25	RPI4
55 µl, 7°C, 75W, 40s	440	6	5	25	RPI1

Table 52 | Conditions for library amplification by PCR

One of the libraries was sequenced on the MiniSeq platform (Table 53) and sequencing data was analyzed as previously described. The untargeted library generated using random hexamer primers was used for the DASH approach.

Library ID	Concentration Qubit [ng/µl]	Concentration Bioanalyzer (150-800bp) [nM]	Conc. for loading [pM]	PhiX [%]	MiniSeq Kit
hESCs and fibroblasts, 55 µl, 7°C, 50W, 30s + targeted primers v2	2.1	8.4	1.8	1%	150 cycles, High Output

Table 53 | Sequencing of RNA-Seq library generated using modified CEL-Seq2 protocol and targeted primers v2

2.6 Depletion of Abundant Sequences by Hybridization (DASH) using Cas9

2.6.1 Design of single guide RNA library

Sequencing data from the samples containing only Fibroblasts and hESCs was analyzed to design a library of single guide RNAs (sgRNAs) to deplete reads mapping to genes abundant in fibroblasts from sequencing libraries and thereby increase the coverage of hESC-relevant genes. Therefore, the DASHit software tool was used[70]. The reads were trimmed to remove poly A stretches and poly G stretches using cutadapt to avoid unspecific sgRNAs. The sequencing data was converted from FASTQ to FASTA format using seqtk (v1.3). Then, candidate sgRNA targets were identified in the reads. Out of those candidates the top 5000 sgRNAs that hit the largest number of reads were extracted. The sequencing data of a sample of only fibroblasts and a sample of only hESCs were DASHed *in silico* using the top 5000 sgRNAs and also the top 1000 sgRNAs. The resulting reads (DASHed and not DASHed) were aligned to the reference genome using STAR and compared.

2.6.2 Preparing the DASH library

For each sgRNA, a DNA oligo was purchased which included the T7 promoter (for IVT, 22bp), the spacer sequence (20 bp) and the first 22 base pairs of the tracr RNA. Furthermore, a DNA oligo representing the 3' end of the sgRNA (90 bp) was purchased (Table 54). The library of oligos representing the 5’ end of the sgRNAs was mixed with the DNA oligo representing the 3’ end of the sgRNA the two primers and the PCR reagents in a PCR tube (Table 55). The DNA template library for IVT was then generated and amplified in a pooled reaction using the PCR program specified in Table 56.

Description	**Sequence (5’-3’)**	**Length [bp]**
5’ end of sgRNA (DASH library)	TAATACGACTCACTATAGNNNNNNNNNNNNNNNNNNNGTTTAAGAGCTATGCTGGAAAC	60
3’ end of sgRNA	AAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTAAACTTGCTATGCTGTTTCCAGCATAGCTCTTA	60
Primer FWD	TAATACGACTCACTATAG	18
Primer REV	AAAAAAAGCACCGACTCGGTGC	22
DNA template for IVT	TAATACGACTCACTATAGNNNNNNNNNNNNNNNNNNNGTTTAAGAGCTATGCTGGAAACAGCATAGCAAGTTTAAATAAGGCTAGTCCGTTATCAACTTGAAAAAGTGGCACCGAGTCGGTGCTTTTTTT	131

Table 54 | Components to generate DNA template for IVT of DASH library

Component	**Concentration**	**Volume [µl]**	**Amount [pmol]**
5’ end of sgRNA (DASH library)	17.92 ng/µl	1	0.965
3’ end of sgRNA	10 µM	0.5	5
Primer FWD	100 µM	0.5	50
Primer REV	100 µM	0.5	50
Q5 High-Fidelity 2x Master Mix	2x	25	NA
PCR water	NA	22.5	NA

Table 55 | PCR conditions to amplify DNA template of DASH library

STEP	**TEMP**	**TIME**
Initial Denaturation	98°C	30 s
35 cycles	98°C	10s
	47°C	30s
	72°C	30s
Final Extension	72°C	2 minutes
Hold	4°C	

Table 56 | Thermocycling conditions to amplify DNA template of DASH library

A part of the PCR reaction was analyzed on a 1% Agarose Gel to confirm unique product at the expected size. The rest of the reaction was cleaned up using a PCR cleanup kit (NEB, T1030S). The IVT reaction was then set up according to the specifications in TABLE and using the MEGAscript® T7 Transcription Kit (Invitrogen, #AM1334). The reaction was thoroughly mixed by vortexing, contents were collected at the bottom of the tube by centrifugation in the minicentrifuge and the reaction was incubated at 37°C for 13 hours (Table 57). Then, 1 µl TURBO DNase was added and the sample was incubated at 37°C for 15 minutes. The IVT product was then cleaned up using the MEGAclear™ Kit Purification for Large Scale Transcription Reactions (Invitrogen, #AM1908) according to the recommendations of the manufacturer. The transcribed RNA was eluted twice in 50 µl of preheated Elution Buffer yielding a concentration of 322 ng/µl. The RNA was then concentrated by precipitation using 5 M Ammonium Acetate and resuspension in 20 µl H2O, leading to a concentration of 1369 ng/µl. The sample was aliquoted into 5 µl aliquots and stored at -80°C until use.

Component	**Volume [µl]**
ATP solution	2
UTP solution	2
GTP solution	2
CTP solution	2
10X Reaction Buffer	2
Enzyme Mix	2
linear template DNA (from PCR)	1.6 (300 ng)
Nuclease-free water	6.4

Table 57 | IVT reaction conditions

2.6.3 DASHing of sequencing libraries

The generated libraries of sgRNAs were used to DASH out fibroblast-abundant genes from generated CEL-Seq2 libraries. The quantities of the components used in those reactions can be found in Table 58. As a control, the same procedure was followed with H20 instead of the DASH library of sgRNAs.

Component	**Concentration [ng/µl]**	**molarity**
Cas9-NLS	NA	40 µM
DASH-library of sgRNAs	1369 ng/ul	37.34 pmol/ul
untargeted library (hESCs and Fibroblasts only)	2.4 ng/ul	14.5 nM
untargeted library (hESCs spiked into Fibroblasts)	3.24 ng/ul	15.5 nM

Table 58 | Quantities of components used in DASH reactions

Cas9 protein was mixed with the DASH sgRNA library in Buffer 3.1 (NEB, #B7203S) to a volume of 4 µl and incubated at 37°C for 10 minutes. Then, the cDNA library was added to each reaction and the reaction was incubated at 37°C for 3 hours. The quantitative details of each DASH reaction can be found in Table 59, Table 60, Table 61and Table 62.

Component	Molarity [nM]	volume [µl]	Final concentration [µM]
Cas9-NLS	275	1.00	2.75
DASH-library of sgRNAs	2750	1.00	27.5
Untargeted library (hESCs and fibroblasts)	2.75	5.00	0.0055
H2O		2.00	
NEBuffer 3.1 (10x)		1.00	
Total		10.00	

Table 59 | DASH reaction conditions – untargeted library (hESCs and fibroblasts), 10x

Component	Molarity [nM]	volume [µl]	Final concentration [µM]
Cas9-NLS	27.5	1.00	0.275
DASH-library of sgRNAs	275	1.00	2.75
Untargeted library (hESCs and fibroblasts)	0.275	5.00	0.0055
H2O		2.00	
NEBuffer 3.1 (10x)		1.00	
Total		10.00	

Table 60 | DASH reaction conditions – untargeted library (hESCs and fibroblasts), 1x

Component	Molarity [nM]	volume [µl]	Final concentration [µM]
Cas9-NLS	260	1.00	2.6
DASH-library of sgRNAs	2600	1.00	26
Untargeted library (hESCs spiked into fibroblasts)	2.6	5.00	0.0052
H2O		2.00	
NEBuffer 3.1 (10x)		1.00	
Total		10.00	

Table 61 | DASH reaction conditions – untargeted library (spike-ins), 10x

Component	Molarity [nM]	volume [µl]	Final concentration [µM]
Cas9-NLS	26	1.00	0.26
DASH-library of sgRNAs	260	1.00	2.6
Untargeted library (hESCs spiked into fibroblasts)	0.26	5.00	0.0052

H2O		2.00	
NEBuffer 3.1 (10x)		1.00	
Total		10.00	

Table 62 | DASH reaction conditions – untargeted library (spike-ins), 1x

At the end of the incubation, 1 µl of Proteinase K was added and the sample was incubated at 37°C for 10 minutes. 89 µl water was added to the sample and then 100 µl of phenol: chloroform: isoamyl (25:24:1) alcohol (Thermo Fisher Scientific, 15593031) were also added. After 20 seconds of vortexing, the sample was centrifuged at room temperature for 5 minutes at 20.000 rcf. The aqueous phase (~95 µl) was transferred to a new tube and 1 µl of glycogen (20 µg/µl, Invitrogen, #10814010) was added, as well as 9.5 µl Sodium-Acetate (3M, pH 5.2) and 285 µl 100% ethanol. The reaction was incubated at -20°C overnight. The next day, the sample was centrifuged at 20.000 rcf and 4°C for 30 minutes and the supernatant was removed. 150 µl of 70% ethanol were added and the sample was again centrifuged at 20.000 rcf and 4°C for 30 minutes. The supernatant was removed, and the sample was again centrifuged at 20.000 rcf and 4°C for 30 minutes. As much of the supernatant as possible was removed and the pellet was air dried until it appeared clear. It was resuspended in 13 µl H2O.

1 µl of the samples was used to determine the quantity of the library by qPCR (Table 63). Then, the DASHed and not DASHed libraries were amplified in a 25 µl PCR reaction as described in the CEL-Seq2 refined protocol. Amplified libraries were analyzed with an Agilent 2100 Bioanalyzer and a DNA High Sensitivity Chip (Agilent, #5067-4626). Then, the indicated libraries were sequenced on an Illumina MiniSeq sequencing platform (Table 64).

Sample ID	**qPCR max(fluorescence) /2 [cycle]**	**PCR cycles**	**RPIX primer**
F+E _dashed_10x	9	7	RPI1
F+E _undashed_10x	9	7	RPI1
F+E _dashed_1x	13	11	RPI1
F+E _undashed_1x	13	11	RPI1
SI_dashed_10x	9	7	RPI2
SI_undashed_10x	8	6	RPI2
SI_dashed_1x	13	11	RPI2
SI_undashed_1x	13	11	RPI2

Table 63 | qPCR results and PCR conditions for library amplification

Library ID	**Concentration (150-800bp) [nM]**	**Conc. for loading [pM]**	**PhiX [%]**	**MiniSeq Kit**
Fibroblasts and hESCs, 55 µl, 7°C, 75W, 40s, DASHed_1x	10.6	1.8	1%	150 cycles, High Output

Fibroblasts and hESCs, 55 µl, 7°C, 75W, 40s, unDASHed_1x	9.6	1.8	1%	150 cycles, High Output
Spike-ins, 55 µl, 7°C, 75W, 40s, DASHed _1x	10.7	1.8	1%	150 cycles, High Output
Spike-ins, 55 µl, 7°C, 75W, 40s, unDASHed _1x	8.4	1.8	1%	150 cycles, High Output

Table 64 | Sequencing of DASHed and not-DASHed RNA-Seq libraries

2.7 Finding Lowly Abundant Sequences by Hybridization (FLASH)

Sequencing data obtained from sequencing hESCs and fibroblasts, 164 genes were selected that are specific to the hESC state while being not expressed in the fibroblast state. Using the FLASHit software tool (https://github.com/czbiohub/flash), 836 sgRNAs were designed which target those 164 genes. However, this library of sgRNAs was never experimentally validated.

2.8 Tissue culture

For my experiments, I kept BJ fibroblasts, HEK293T and HEK293 cells in culture and maintained them. For adherent cell lines I passaged by trypsinization and subsequent dilution in fresh media. Depending on the cell line, different cell culture media with different additives were required (Table 65).

Name	**Organism**	**Tissue**	**Morphology**	**Culture properties**	**Culture media**
BJ fibroblasts (CRL-2522)	*Homo sapiens*	Skin; foreskin	fibroblast	adherent	DMEM medium (with 4.5 g/L Glucose and L-Glutamine, without Sodium Pyruvate) supplemented with 10% FBS (v/v), 100U/ml Penicillin, 100mg/ml Streptomycin, 1× nonessential amino acids, 1x Sodium Pyruvate, and 0.06 mM β-Mercaptoethanol
HEK293T	*Homo sapiens*	colon	epithelial	adherent	DMEM medium (with 4.5 g/L Glucose and L-Glutamine, without Sodium Pyruvate) supplemented with 10% FBS (v/v)
HEK293	*Homo sapiens*	kidney	epithelial	adherent	DMEM medium (with 4.5 g/L Glucose and L-Glutamine, without Sodium Pyruvate) supplemented with 10% FBS (v/v), 100U/ml Penicillin, 100mg/ml Streptomycin

Table 65 | Summary of cultured cell lines

To freeze cells, 1x freezing media was prepared consisting of 20% culture media, 20% FBS and 10% DMSO. Adherent cells were trypsinized and resuspended in 1.5 ml freezing media. Then they were transferred to a cryovial which was put into a Mr. Frosty and at -80°C for at least 2 hours. Then the cryovial was transferred to liquid nitrogen.

2.9 Reprogramming fibroblasts in 96-well plates

For the reprogramming experiments, wells of a 96-well plate and of a 6-well plate were covered with Geltrex™ LDEV-Free, hESC-Qualified, Reduced Growth Factor Basement Membrane (Thermo Fisher Scientific, #A1413302). Therefore, 45 µl per well (96-well plate) and 1.5 ml per well (6-well plate) were dispensed into the wells and the plates were incubated in the tissue culture incubator at 37°C for at least 60 minutes. The Geltrex solution was aspirated and BJ-Fibroblast cells (ATCC, CRL-2522, Passage number 9) were added immediately to the wells in media without antibiotics – 15,000 or 30,000 per well in the 96-well plate and 100,000 per well in the 6-well plate (Day -1). On day 0, the cells were infected with the indicated amount of concentrated lentivirus (Figure 5). 0.42 µl VMS028 and 0.7 µl VMS005 concentrated lentivirus was used in the well of the 6-well plate. After 48 hours on day 2, the media was changed to culture media without antibiotics containing 2 µg/ml Puromycin and 5 µg/ml Blasticidin to select for cells infected by both lentiviruses. The media was changed daily and on day 8, the media was switched to TeSR™-E7™ media (Stemcell Technologies, #05914) still containing Puromycin and Blasticidin. On day 9, the media was changed to TeSR™-E7™ media without Puromycin and Blasticidin. The media was changed daily until day 24 for the 6-well experiment and day 20 for the 96 well experiment. Pictures were taken using a microscope to assess reprogramming progress.

	1	2	3	4	5	6	7	8	9	10	11	12
A	0.05µl VMS028 0.08ul VMS005	0.05µl VMS028 0.04ul VMS005	0.05µl VMS028 0.02ul VMS005	0.1µl VMS030 0.08ul VMS005	0.05µl VMS030 0.04ul VMS005	0.05µl VMS030 0.02ul VMS005	0.1ul VMS028 0.16ul VMS005	0.1ul VMS028 0.08ul VMS005	0.1ul VMS028 0.04ul VMS005	0.1ul VMS030 0.16ul VMS005	0.1ul VMS030 0.08ul VMS005	0.1ul VMS030 0.04ul VMS005
B												
C												
D												
E												
F												
G												
H												

15,000 cells seeded on day-1
30,00 cells seeded on day-1

Figure 5 | Plate layout for reprogramming fibroblasts in 96-well plate

2.10 Production of lentiviral supernatant

To obtain cells that stably express the target proteins I used lentivirus which stably integrates into the genome. Lentivirus was produced either by the ViraCore facility provided by the McManus lab or by me.

For the production of lentiviral supernatant in 10 cm dishes, HEK-293T packaging cells were used. 5E+06 cells were seeded into a 10 cm tissue culture plate. The next day, when the cells were about 70% to 80% confluency, they were transfected using jetPRIME® DNA and/or siRNA transfection reagent (VWR, #89129-924). Therefore, 100 µl 5x jetPRIME® transfection buffer was mixed with 4.5 µg 2nd generation packaging plasmid mix in 45 µl water. Then 5.5 µg transfer

plasmid was added in 55 µl water. 280 µl water were added, the mixture was vortexed and incubated at room temperature for 2 minutes. Finally, 20 µl of jetPRIME® transfection reagent was added, the mixture was vortexed thoroughly and incubated at room temparature for 15 minutes. Then, the mixture was added dropwise to the seeded cells. The next day after transfection I changed the media to DMEM. 3 and 4 days after transfection, the lentiviral supernatant was harvested.

Analogously, virus was produced in 96-well plates with quantities indicated in Table 66. In this case, 20,000 cells were seeded per well.

Component	**Volume [µl]**
5x jetPRIME® transfection buffer	1.5
2nd generation packaging plasmid mix	1.5 (150 ng)
Transfer plasmid	2.5 (250 ng)
H2O	1.3
jetPRIME® transfection reagent	0.7

Table 66 | Quantities of transfection reaction components per well when transfecting in a 96-well plate

3 Results

3.1 Gene expression profile comparison between iPSCs and BJ fibroblasts

First, I investigated how the transcriptome of the fibroblast state differed from the transcriptome of the reprogrammed state. To do so, I used RNA-Seq data from the Gene Expression Omnibus to compare gene expression levels of human Fibroblasts and induced pluripotent stem cells (iPSCs)[71–73]. As expected, the transcriptomes of the two cell states differ drastically, with a lot of genes being upregulated or downregulated in the reprogrammed state (Figure 6A). Conversely, the gene expression levels in two different reprogrammed cell lines were almost identical (Figure 6B). This demonstrates that this cell state transition offers a suitable model to benchmark my novel method.

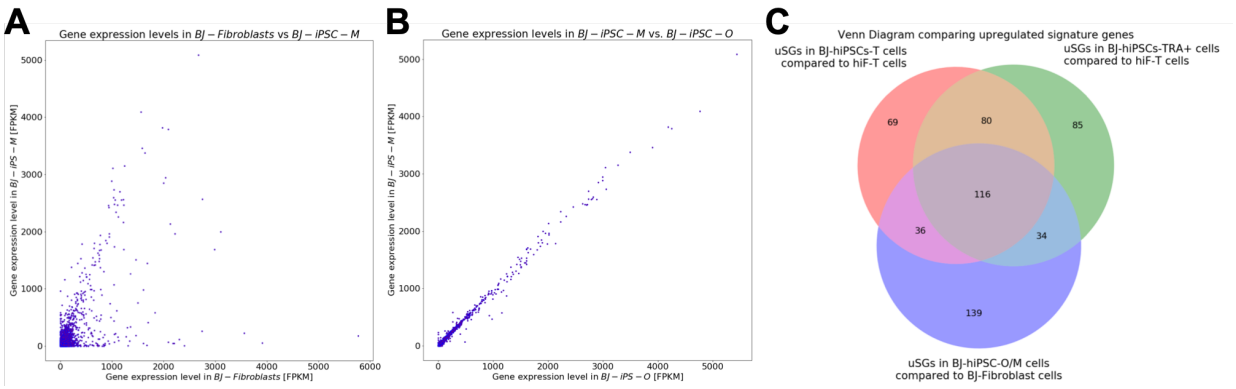


Figure 6 | A Comparing gene expression levels in BJ fibroblasts and BJ-iPSC-M cells **| B** Comparing gene expression levels in two different iPSC cell lines **| C** Comparing upregulated signature genes between three similar cell state transitions

3.2 Finding suitable target gene set

I then analyzed the RNA-Seq data of two different publications[71,72] on reprogramming BJ fibroblasts to iPSCs to find common genes that are upregulated in the reprogrammed state as

well as lowly or not expressed in the fibroblast state. I called those genes upregulated signature genes (uSGs). I reasoned that RNA-Seq enriched for this gene set would allow me to reliably identify and quantify reprogrammed cells in a mixture of fibroblast cells, even if those reprogrammed cells are very lowly abundant. At the same time, I would reduce unnecessary sequencing data and costs by enriching for those genes. Using a script wrote in python, I identified uSGs for three almost identical cell state transitions: human BJ fibroblasts to iPSC-O or M clones[71], human secondary induced BJ fibroblasts immortalized with T antigen (hiF-T) to human induced pluripotent stem cells immortalized with T-antigen (BJ-hiPSCs-T), and hiF-T cells to the TRA-1-60+ fraction of BJ-hiPSCs-T cells (BJ-hiPSCs-T-TRA+)[72]. TRA-1-60 is a pluripotency-associated marker in human stem cells and iPSCs[72]. Unexpectedly, those transitions only shared a relatively small part (116 genes) of their uSGs between them, demonstrating the sometimes stochastic and variable nature of reprogramming specifically and cell state transitions in general (Figure 6C). However, it must be noted here that while the cutoffs for defining upregulated signature genes were applied consistently, they were also chosen relatively arbitrary (reference python code). Thus, while it is clear that these 116 “core” genes are consistently upregulated in all three independent state transitions, there might be more that this analysis missed due to filtering that is too strict. On the other hand, if the filtering was too tolerant, there might only be a small subset of those “core” genes that are consistently upregulated during reprogramming. Either way, my analysis yielded a gene set of 116 uSGs shared between all analyzed cell state transitions that are characteristically upregulated in the reprogrammed state compared to the fibroblast state, while being lowly or not expressed in the fibroblast state. Thus, this gene set should be suitable to detect lowly abundant reprogrammed cells in a mixture with fibroblast cells. Furthermore, this analysis method can easily be applied to find uSGs of other cell state transitions, as long as RNA-Seq data from the parental and destination state is available.

3.3 Designing targeted primers v1

To incorporate the 116 uSGs of the transition from fibroblast to reprogrammed state into an enriching RNA-Seq library generation method, I first had to identify an RNA-Seq method suitable for my needs. It would have to allow for high multiplexing while being relatively fast, straight forward, economically efficient and customizable. The CEL-Seq2 method ticked all of those boxes despite being a single-cell RNA-Seq method[69], which represents a strength since I was working with rather few cells and did not need complete resolution of all genes in each cell, but rather could view each well as a composite transcriptome averaged across all cells in it. So, I designed CEL-Seq2-compatible primers against the 3’ end of all transcripts of the 116 identified uSGs using python scripts and primer3, allowing a maximum of 2 primers per transcript. After eliminating redundant primers, the design yielded 235 targeted RT primers which could be used instead of random hexamer primers in the second reverse transcription step of the CEL-Seq2 protocol.

3.4 Spiking hESCs into fibroblasts

To simulate the rare event of reprogramming and assess the performance of the modified CEL-Seq2 method, I spiked in varying amounts of human embryonic stem cells (hESCs) into human BJ fibroblasts. Therefore, I seeded 33.000 BJ fibroblasts into wells of a 96 well tissue culture dish. After attachment, I added 5, 50, 500 and 5000 hESCs by serial dilution to the respective wells of the dish. Then, I extracted total RNA from the wells. To make sure that the extracted RNA was

adequate for further analysis, I assessed its quality and quantity for most of the wells. During the first trial of the spike-in experiment, the RNA was of good integrity for the most part. While the concentration was quite consistent between samples it was overall quite low (Table 67). Interestingly, there was much more RNA in the wells containing only hESCs. This might be explained by the fact that hESCs are much smaller than fibroblast[74] and thus there was probably a higher number of cells in a given volume. While the RNA quality and quantity were generally sufficient for generating libraries using the modified CEL-Seq2 method, the still rather low concentration could lead to problems in detecting especially the lower amounts of spiked in hESCs. That is why I repeated the spike-in experiment and extraction method with slight modifications to increase the concentration of the extracted RNA and improve its integrity. When I extracted the RNA from the wells, I used a lower elution volume in order to increase the concentration of the eluted RNA. Indeed, I managed to almost double the concentration of the extracted RNA (Table 68). Moreover, the concentrations were again highly consistent between wells and the RNA integrity number (RIN) was quite high when analyzed on the bioanalyzer (Table 68). This shows that this RNA extraction method yields RNA of high quality and concentration which is ideal for all kinds of library generation methods. Also, the consistency and reproducibility of this kind of quality and concentration demonstrates the method's suitability for large scale applications such as the planned proof of concept screen.

Sample condition	**RNA conc. [ng/µl]**	**RIN**	**rRNA ratio [28S/18S]**
5 hESCs	41.60	8	1.43
5 hESCs	28	8.8	2.2
5 hESCs	22.00	1.8	9.2
50 hESCs	24	9	2.2
50 hESCs	20	9.1	1.9
50 hESCs	23.40	8.6	1.52
500 hESCs	23.24	7.70	1.93
500 hESCs	24	8.7	1.7
500 hESCs	NA	NA	NA
5000 hESCs	48.57	8.6	1.48
5000 hESCs	62.17	7.3	1.42
5000 hESCs	NA	NA	NA
hESCs	301	9.6	2
hESCs	513	9.4	1.9
Fibroblasts	41.85	8.7	1.8
Fibroblasts	45.10	NA	1.66

Table 67 | Quality and quantity of extracted total RNA – first repetition

Sample condition	**RNA conc. [ng/µl]**	**RIN**	**rRNA ratio [28S/18S]**
5 hESCs	51.3	9	2.2

5 hESCs	43	8.8	2.2
5 hESCs	24.4	9.1	1.9
50 hESCs	39.3	8.7	1.7
50 hESCs	20.7	9.3	1.7
50 hESCs	41.64	9.2	1.8
500 hESCs	40.76	9.6	2
500 hESCs	38.01	9.4	1.9
500 hESCs	37.56	NA	NA
5000 hESCs	42.23	8.7	1.8
5000 hESCs	43.33	9.5	2.1
5000 hESCs	48.54	9.5	2.1
hESCs	9.1	9.4	2.1
Fibroblasts	43.51	9.4	2.2

Table 68 | Quality and quantity of extracted total RNA – second repetition

3.5 Library generation using QuantSeq kit

To identify the performance and limits of RNA-Sequencing as a means to detect low amounts of hESCs in a mixture of fibroblasts, I set out to analyze the RNA samples extracted from the hESCs spiked into the fibroblasts. Therefore, I first utilized the QuantSeq 3' mRNA FWD library preparation kit to generate RNA-Seq libraries for the different RNA samples. Before library amplification, the amount of the libraries was highly consistent except for one sample (the first replicate of the sample containing 5000 hESCs), which was however explained by elution in three times the elution volume of the other samples (Figure 7A). As expected, this sample needed about two cycles more to reach half-maximal fluorescence. After library amplification, the libraries were assessed for quality and quantity on a bioanalyzer machine. None of the libraries was overamplified and they all showed a size distribution with a peak at around a length of 250 base pairs and a mean size of around 350 base pairs, as expected (Figure 7B). After pooling the libraries in an equimolar ratio, they were ready to be sequenced.

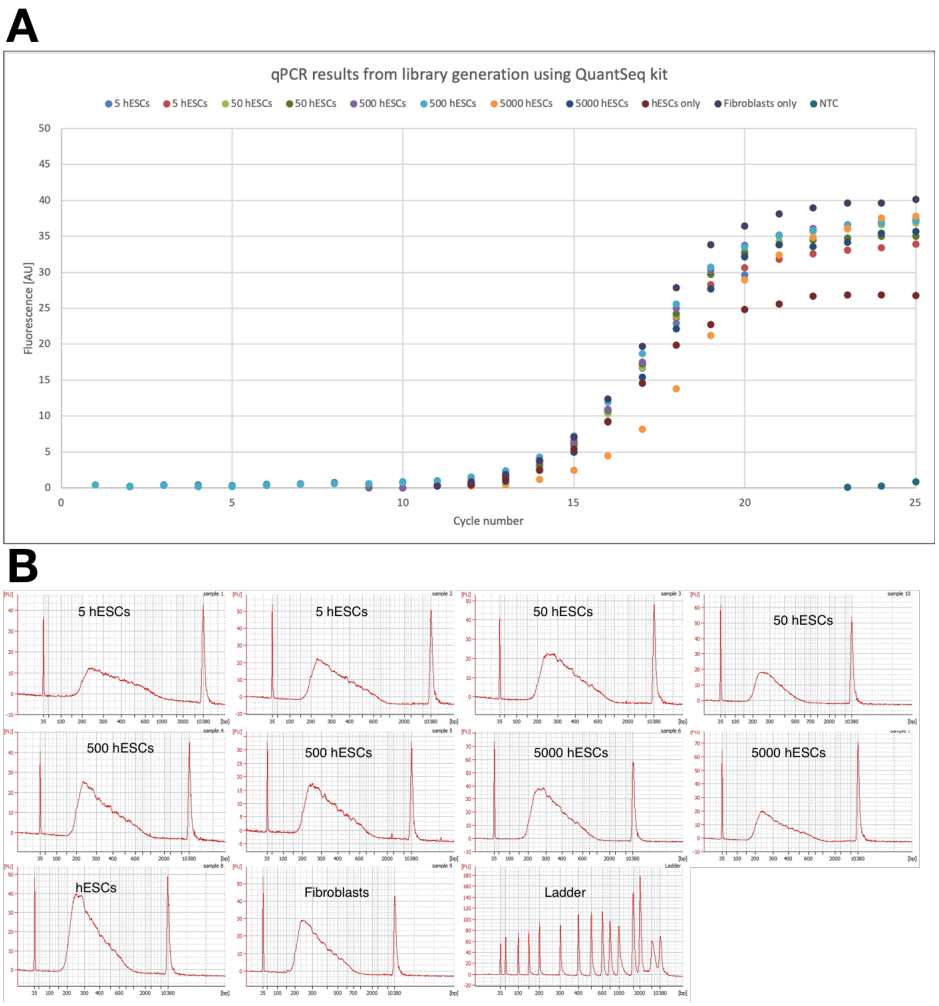


Figure 7 | A Bioanalyzer EGRAM traces of QuantSeq cDNA libraries **| B** qPCR results from library generation using QuantSeq kit

3.6 Sequencing analysis of QuantSeq RNA-Seq libraries

I then analyzed the QuantSeq libraries of the RNA of hESCs spiked into fibroblasts. Therefore, I sequenced the pooled libraries on a MiniSeq system and analyzed the data using various software tools. The quality of the sequencing run was good, so I proceeded with analyzing the data. While the samples containing 5000 or 500 hESCs are quite well distinguishable by principle component analysis (PCA), the samples containing 50 or 5 hESCs clustered together quite closely as well as with the fibroblast sample (Figure 8A). When excluding the fibroblast and hESC samples from the PCA, samples containing different amounts of hESCs spiked in cluster a bit further away from each other (Figure 8B). However, the 'intersample' and intrasample distances between the samples containing 50 and 5 hESCs spiked in are about the same, making those samples not well distinguishable (Figure 8B). Differential gene expression analysis showed that only the genes ESRG, LIN28A and L1TD1 were differentially expressed when comparing the sample containing only fibroblasts with the sample containing 500 hESCs spiked into the fibroblasts (Table 69). DPPA4 was just not significantly differentially expressed (adjusted p-value = 0.07582768). There were no significantly differentially expressed genes in the samples containing 50 or 5 hESCs, consistent with their close clustering with the sample containing only fibroblasts in the PCA

analysis. These results demonstrate that commercial 3’ mRNA sequencing can detect about 500 hESCs spiked into 33,000 fibroblasts (1.5%), even when sequenced at low depth (around 1-1.5 million reads/sample). This limit of detection depends of course on sequencing depth.

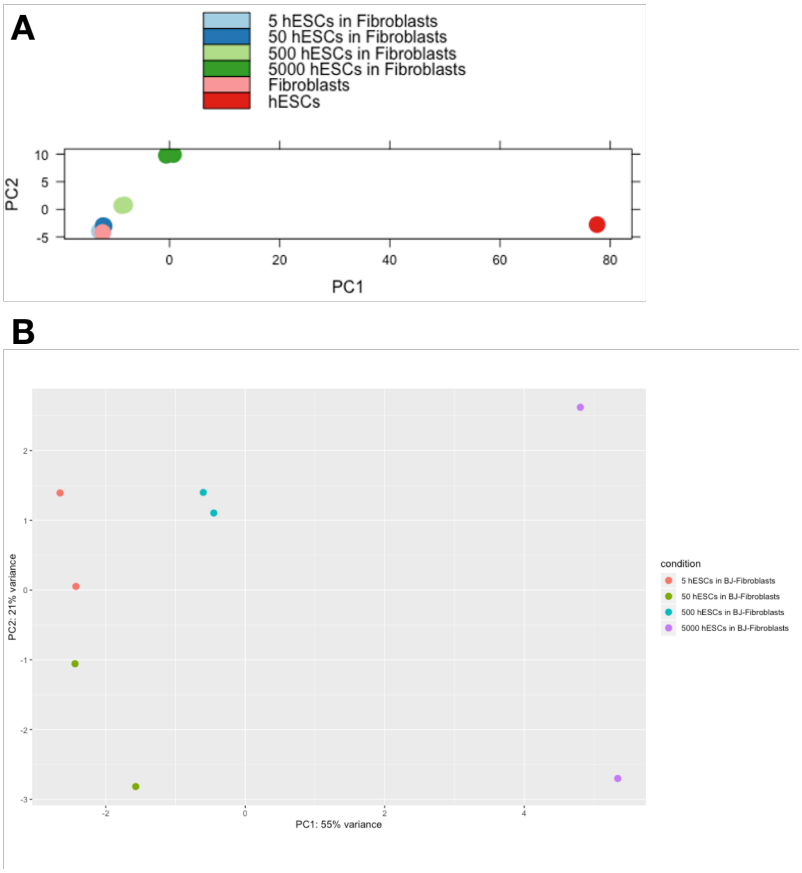


Figure 8 | A PCA of hESC spiked into fibroblasts (QuantSeq) **| B** PCA of hESC spiked into fibroblasts (QuantSeq) without samples containing only fibroblasts and only hESCs

Gene symbol	baseMean fibroblasts	baseMean 500 hESCs	log2FoldChange	pval	padj
ESRG	0	22.8520101	Inf	9.98E-08	0.00186583
LIN28A	1.01663941	25.4735545	4.64712032	2.99E-07	0.00279089
L1TD1	1.01663941	22.3499097	4.45838904	2.11E-06	0.01312043
DPPA4	0	16.2136627	Inf	1.62E-05	0.07582768

Table 69 | Differential Gene Expression analysis between sample containing only fibroblasts and sample containing 500 hESCs spiked into fibroblasts

3.7 Library generation using modified CEL-Seq2 approach with targeted primers v1

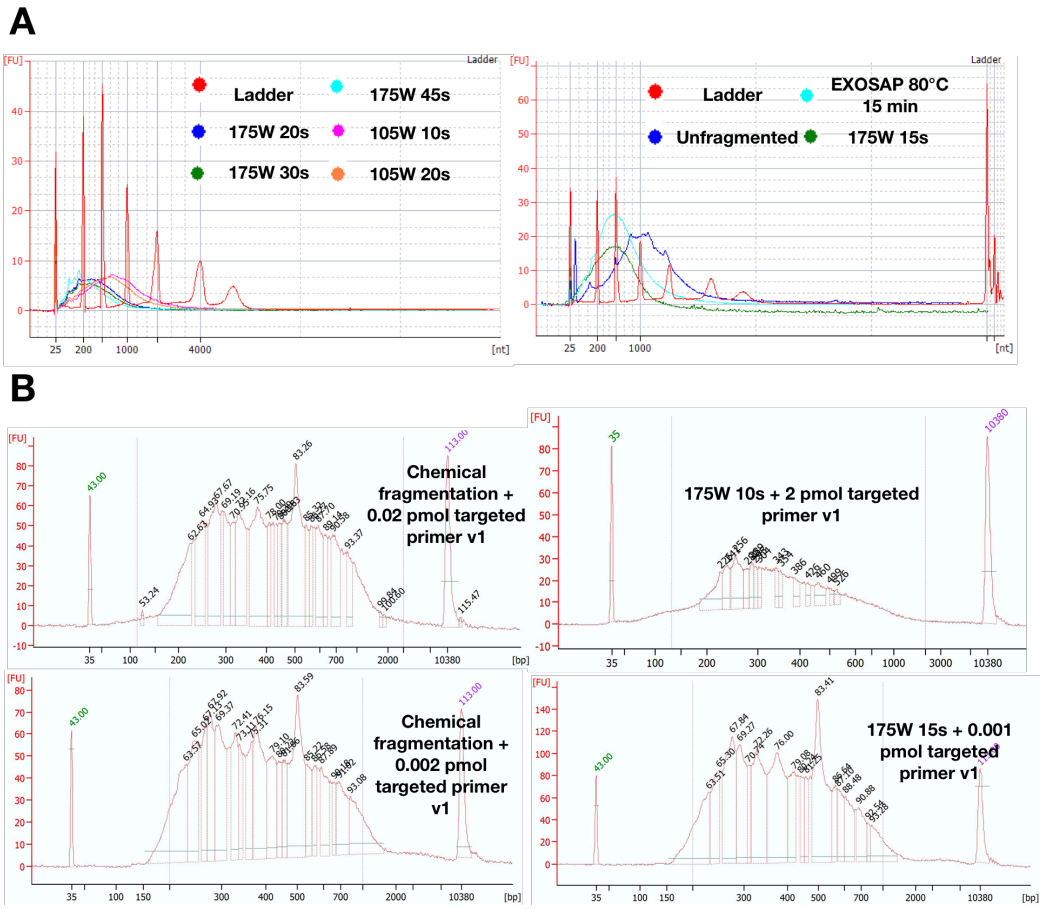


Figure 9| A EGRAM overlay of different fragmentation conditions**| B** Comparing size distribution of libraries when using different amounts of targeted primer v1 input after fragmentation of the aRNA to a length of about 500 base pairs

Then, I tested the performance of the designed targeted primers. To do so, I analyzed the extracted RNA from the spike-in of hESCs into fibroblasts by preparing RNA-Seq libraries using a modified CEL-Seq2 protocol (see materials and methods). After the *in vitro* transcription step to amplify the antisense RNA (aRNA), fragmentation of the aRNA using sonication was optimized (Table 70, Figure 9A).

Duty factor	Watt	cpb	temperature [°C]	water level	volume [µl]	tube	time [s]	peak [bp]
10%	175	200	8	12-14	120	6x16 AFA	10	500
10%	175	200	8	12-14	120	6x16 AFA	15	450
10%	175	200	8	12-14	120	6x16 AFA	20	350
10%	175	200	8	12-14	120	6x16 AFA	30	300
10%	175	200	8	12-14	120	6x16 AFA	45	200

5%	105	200	8	12-14	120	6x16 AFA	10	650
5%	105	200	8	12-14	120	6x16 AFA	20	525
5%	105	200	8	12-14	120	6x16 AFA	40	275
5%	105	200	8	12-14	120	6x16 AFA	50	225
5%	105	200	8	12-14	120	6x16 AFA	80	150
10%	50	1,000	6-8	10	55	microTube-50	20	500
10%	50	1,000	6-8	10	55	microTube-50	30	400
10%	75	1,000	6-8	10	55	microTube-50	30	250
10%	75	1,000	6-8	10	55	microTube-50	20	300

Table 70 | RNA fragmentation conditions using Covaris S220 Ultrasonicator

Unexpectedly, when the mixture was heated after adding ExoSAP-IT™ PCR Product Cleanup Reagent also fragmented the aRNA to a length of around 500 base pairs (Figure 9A). The size distribution of unfragmented aRNA had a peak around 1000 base pairs (Figure 9A). Using the optimized sonication conditions and chemical shearing conditions, I proceeded with the preparation of the libraries. Varying the amount of targeted primer input for the second reverse transcription had no noticeable effect on the size distribution of the library (Figure 9B). However, it did – as expected – have an effect on the amount of library product (Figure 10A). Unexpectedly, not shearing the aRNA at all and reverse transcribing using the targeted primers did not yield a cDNA library with an acceptable size distribution (Figure 10B). Similarly, the size distribution of the library using targeted primers did not look as clean as expected, since it had a long right tail and irregularities even after shearing to a size of about 500 base pairs (Figure 9B). Importantly, when the aRNA was sheared to a size peak of only around 200 base pairs, the quality of the size distribution of the corresponding cDNA library looked significantly better (Figure 10C). Collectively, these data demonstrate that both chemical shearing and shearing by sonication can aid the generation of a high-quality RNA-Seq library, while the effect of primer amount for reverse transcription is negligible, as long as it is within a certain range.

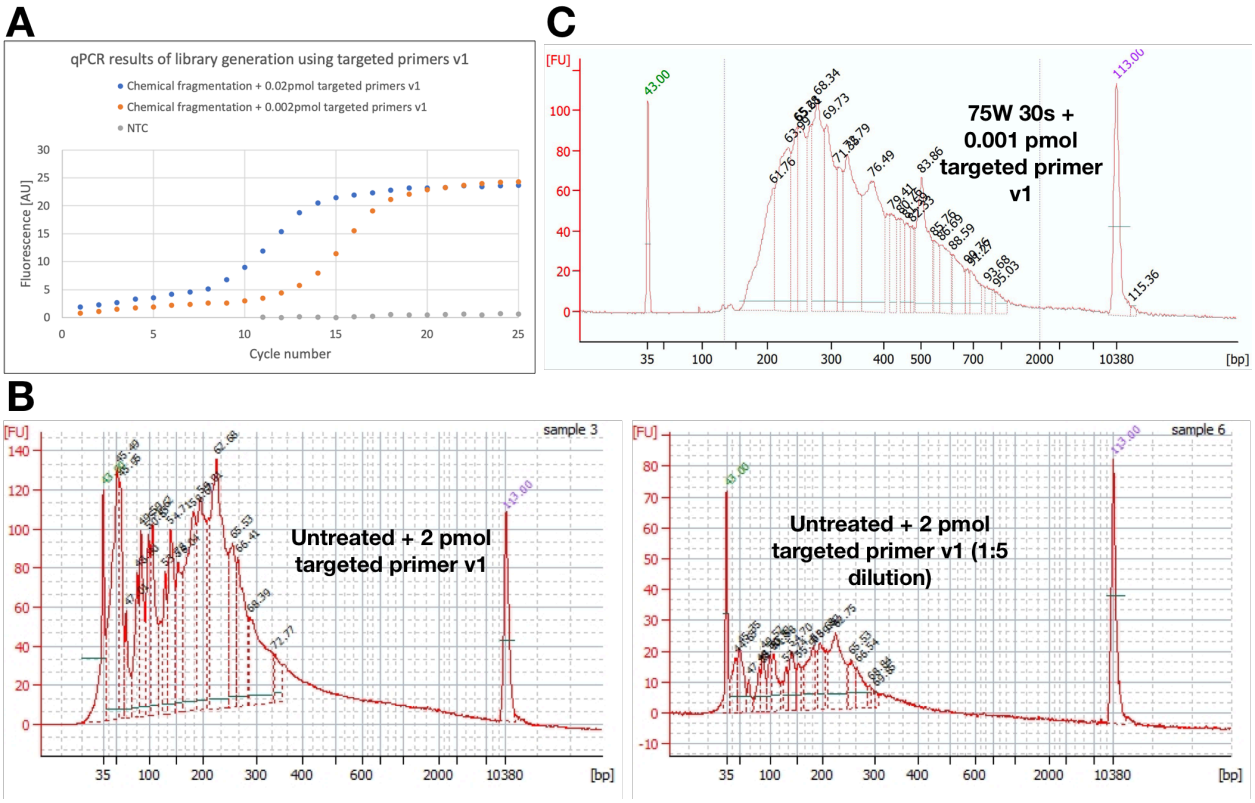


Figure 10 | A Comparing library amounts after reverse transcription using different amounts of targeted primers v1**| B** cDNA library generated using targeted primers v1 without fragmentation of the aRNA **| C** cDNA library generated using targeted primers v1 and fragmentation of the aRNA to about 200 base pairs

3.8 Sequencing analysis of libraries generated using modified CEL-Seq2 approach with targeted primers v1

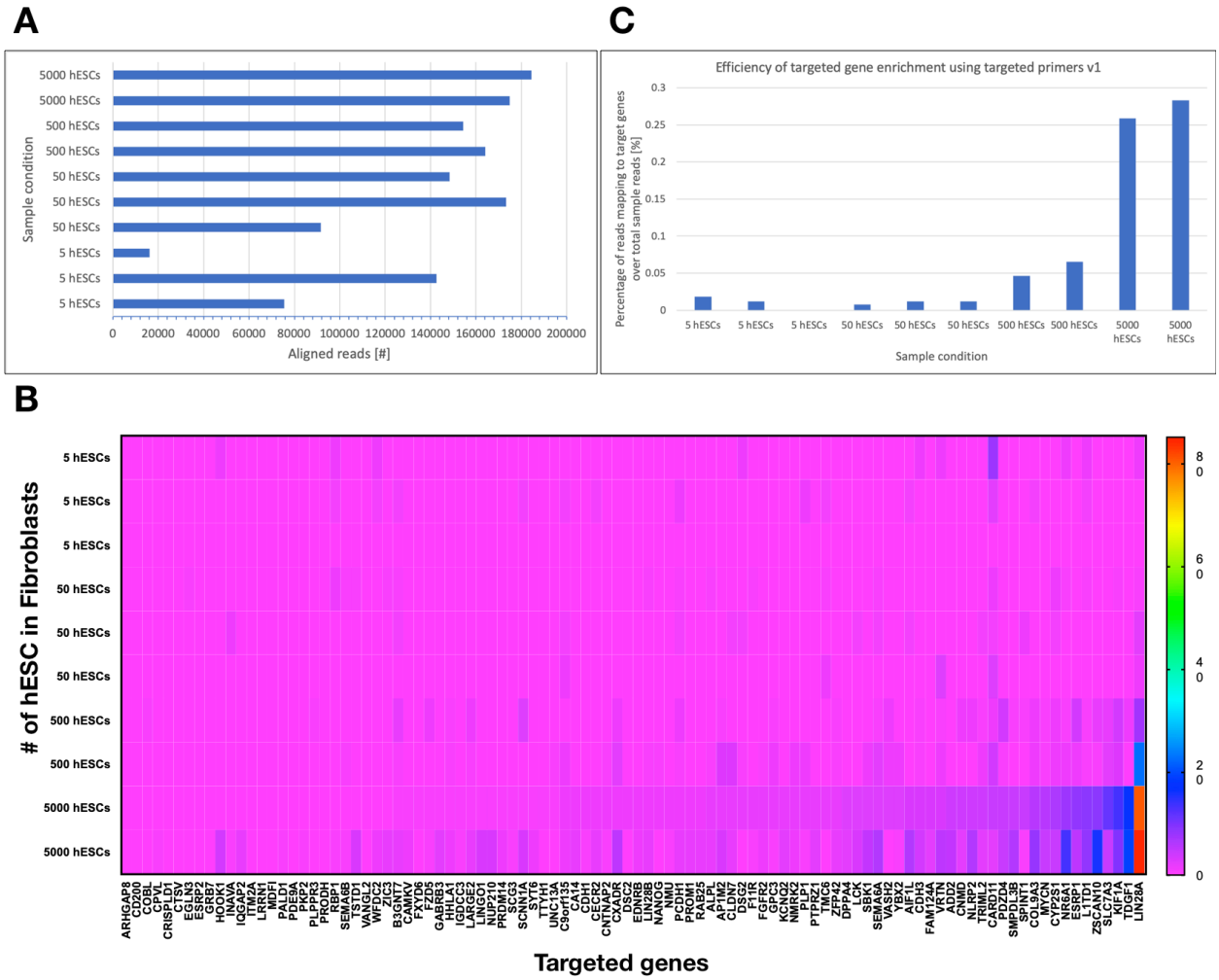


Figure 11 | A Comparison of the number of aligned reads per sample for libraries generated using modified CEL-Seq2 approach with targeted primers v1 **| B** Heatmap depicting normalized read counts of genes targeted by targeted primers v1 **| C** Efficiency of targeted gene enrichment by targeted primers v1

I then assessed the performance of my custom sequencing strategy as well as to which extent the enrichment of target genes was successful. To do so, I sequenced the generated libraries on the MiniSeq system and analyzed the data using various software tools. For this run, I used basespace to automatically convert the BCL data to FASTQ format. After trimming the reads, I checked their quality. I then used the celseq2 tool using STAR as alignment software to align the reads to the human genome and generate the UMI count matrix. I then analyzed the data using both custom python scripts and DESeq2. The sequencing run contained 52.73% PhiX spike-in control and yielded only around 5.8 million reads for all samples, which corresponded to between 16.000 and 184.000 reads per sample (Figure 11A). One sample containing 5 hESCs spiked in had only 16.000 reads. After trimming, the quality of the reads seemed acceptable overall. After making sure that the distribution of the non-target genes was similar between samples, I used

those genes to normalize the different samples. The genes that were targeted had low read counts overall, with many of them not having reads at all (Figure 11B). Also, the portion of reads mapping to targeted genes was quite low across all conditions, although it was substantially higher for the samples containing 500 or 5000 hESCs spiked in (Figure 11C). When I then separated the samples using principal component analysis and taking into account the information on all genes, all samples clustered quite closely together except for one sample that contained 5 hESCs and had the lowest number of reads (Figure 12A). Interestingly, also another sample that contained 5 hESCs clustered a bit further away from the others. After removing the sample with low read counts, the samples clustered quite closely again except for one sample containing 5 hESCs and one sample containing 50 hESCs (Figure 12B). I repeated this analysis to consider only targeted genes, which lead to a visible separation of both of the samples containing 5000 hESCs as well as both of the samples containing 500 hESCs from the rest of the samples (Figure 12C). Exclusion of the sample with low read counts did not alter the result (Figure 12D). Collectively, these results indicate that while the enrichment for target genes was not as successful as expected, the samples can be separated at least until the threshold of 500 hESCs when taking into account information on only the targeted genes. This separation is not apparent when considering all genes.

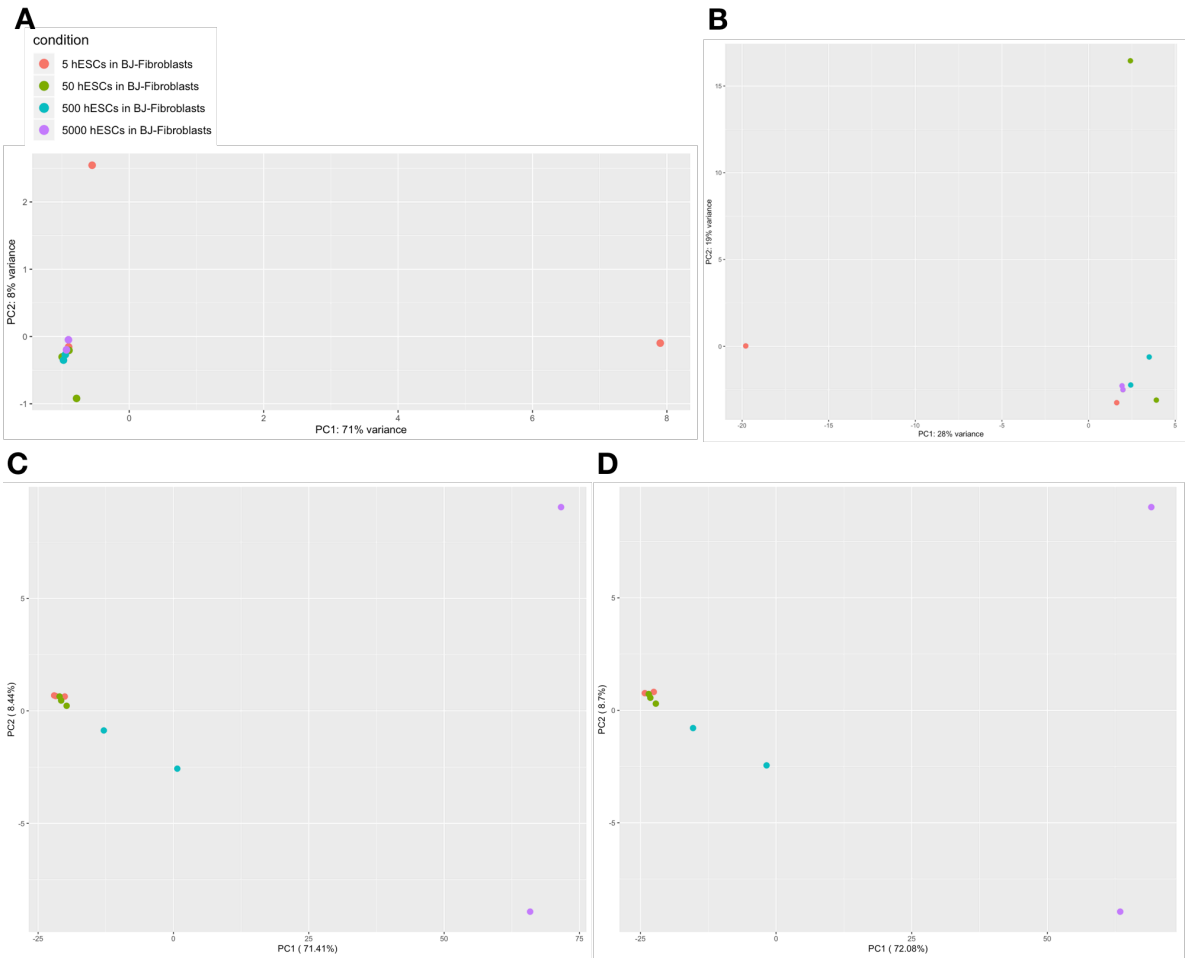


Figure 12 | A PCA taking into account sequencing data of all genes **| B** PCA taking into account sequencing data of all genes after exclusion of sample with low read counts **| C** PCA taking into account sequencing data of genes targeted by targeted primers v1

| D PCA taking into account sequencing data of genes targeted by targeted primers v1 after exclusion of sample with low read counts

3.9 Library generation from samples containing only fibroblasts or hESCs using modified CEL-Seq2 approach

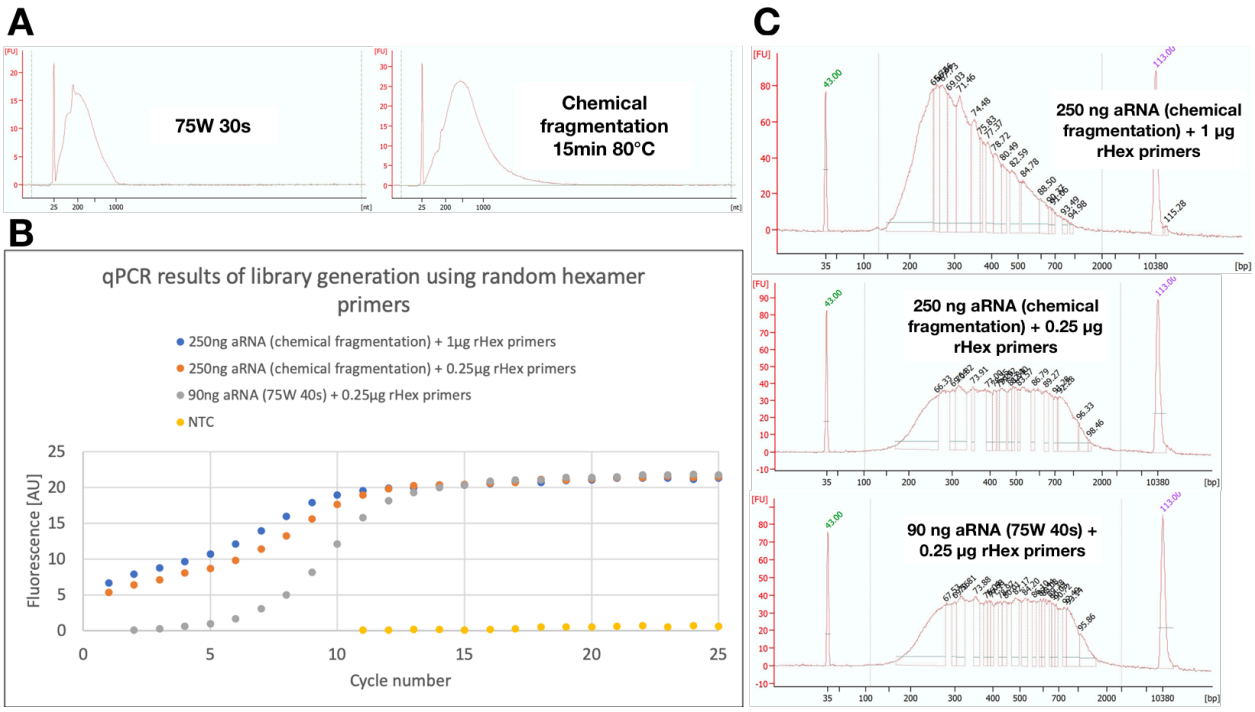


Figure 13 | A EGRAM of aRNA fragmented under indicated conditions **| B** qPCR results for generating libraries using the modified CEL-Seq2 protocol and random hexamer primers **| C** cDNA libraries generated after reverse transcribing indicated amount of fragmented aRNA using indicated conditions with indicated amount of random hexamer primers

Untargeted RNA-Seq libraries from RNA samples containing only fibroblasts or only hESCs were generated. This was done in order to generate sequencing data necessary for Depletion abundant sequences by hybridization (DASH) and finding lowly abundant sequences by hybridization (FLASH) approaches. Furthermore, this data could also be used to improve the design of the primers targeting the hESC-specific transcripts. Therefore, RNA-Seq libraries from RNA extracted from hESCs and BJ fibroblasts were generated using the modified CEL-Seq2 protocol. The best size distributions were obtained by either chemical shearing – although this is concentration dependent and less reproducible – and sonication at 75 W for 30 seconds in a volume of 55 µl at 6-8°C. Chemical shearing resulted in a size distribution with a peak at around 500 nucleotides, while sonication at 75 W for 30 seconds resulted in a size distribution with a peak at around 250 nucleotides (Figure 13A). After finding the optimal conditions for shearing the aRNA, libraries were generated and the cycle number for the PCR step was determined by qPCR (Figure 13B). Varying the amount of primers or of aRNA input into the reverse transcription reaction altered the amount of resulting library only slightly. Expectedly, more primers lead to more library and less aRNA to less library (Figure 13B). The libraries generated from aRNA which was sheared to a size of about 500 base pairs showed a homogenous and broad size distribution indicating poor quality (Figure 13C). The library resulting from aRNA sheared to a length of around 250 base pairs was of high quality in terms of quantity and size distribution, with a peak at around 275

nucleotides (Figure 13C). This library was used for further analysis by sequencing. Collectively, these results show that the CEL-Seq2 approach can be adapted to incorporate aRNA shearing by sonication and is quite robust in terms of primer and aRNA amounts used.

3.10 Using sequencing data from hESCs and fibroblasts for the design of targeted primers v2

I then assessed the composition of the CEL-Seq2 library generated from hESCs and Fibroblasts only to improve the primers used in the modified CEL-Seq2 approach. To do so, I sequenced the library on the MiniSeq platform and analyzed the data using various software tools. Using this data, I then deployed custom python scripts to incorporate information from the sequencing run into the design of new primers. The sequencing run yielded about 36 million reads in total, with aligned reads per sample ranging from about 1 million to 1.6 million (Figure 14A). Expectedly, differential gene expression analysis yielded 8445 differentially expressed genes between fibroblasts and hESCs with an adjusted p-value of less than 0.05 (padj < 0.05). Table 71 shows the top 20 differentially expressed genes with the lowest adjusted p-value from this analysis.

Gene symbol	**log2FoldChange over hESCs**	**pval**	**padj**
CAV1	-3.9508537	0	0
COL1A2	-3.8961292	0	0
COL3A1	-5.2893194	0	0
IGFBP7	-5.8926402	0	0
LGALS1	-5.0843229	0	0
MT2A	-4.1259938	0	0
S100A6	-4.7375952	0	0
COL6A3	-5.78103	2.85E-300	6.12E-297
COL6A2	-3.9471541	2.06E-299	3.93E-296
THBS1	-4.8193597	1.39E-297	2.38E-294
AHNAK	-4.9416762	7.43E-293	1.16E-289
TIMP1	-3.568835	1.17E-286	1.68E-283
VIM	-3.1959162	1.78E-285	2.36E-282
IGFBP4	-4.2335555	2.82E-285	3.45E-282
CD99	-3.5815645	1.35E-274	1.54E-271
CTSB	-4.7287408	9.98E-273	1.07E-269
DKK3	-5.1479672	9.97E-270	1.01E-266
CDKN1A	-4.5410731	6.89E-265	6.57E-262
SERPINE2	-3.3173204	1.95E-264	1.76E-261
TIMP3	-4.9731997	7.87E-261	6.75E-258

Table 71 | Top 20 differentially expressed genes when comparing fibroblasts and hESCs

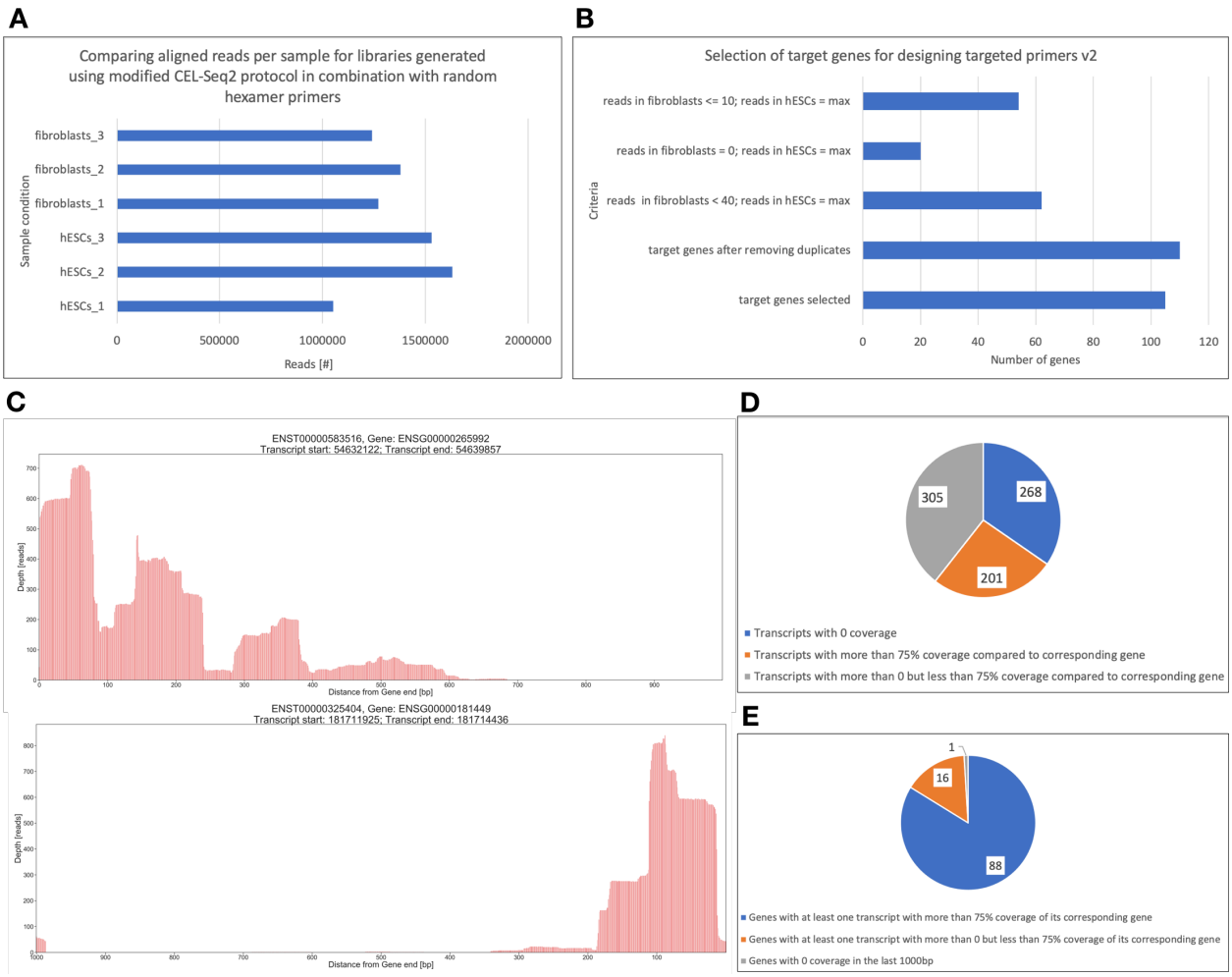


Figure 14 | A Comparison of the number of aligned reads per sample for libraries generated using modified CEL-Seq2 approach with random hexamer primers **| B** Selection of target genes for the design of targeted primers v2 **| C** Example graphs from the analysis of coverage of target transcripts at 3' end **| D** Filtering of transcripts of target genes according to coverage at 3' end **| E** Filtering of target genes according to coverage at 3' end

To select the new genes to enrich for, I filtered the data for 20 genes that had the highest read count in the hESC samples while having 0 reads aligned in the fibroblast samples. Furthermore, I added 54 genes that had the highest read count in the hESC samples while having less than 10 reads aligned in the fibroblast samples, as well as 62 genes that had the highest read count in the hESC samples while having less than 40 reads aligned in the fibroblast samples (Figure 14B). After further filtering, the final target gene list included 105 genes, which corresponded to 774 transcripts. Since those included a lot of unexpressed or lowly expressed transcripts, I performed further filtering by comparing the maximum coverage of the end of each transcript to the maximum coverage of its corresponding gene (Figure 14C). I excluded a gene without coverage in the last 1000 base pairs and 8 genes without transcripts of high enough coverage (Figure 14D, E). Then, I chose all transcripts with more than 75% coverage of their corresponding gene (201) and added the most abundant transcript for all target genes where no transcript was chosen yet. This led to a selection of 209 target transcripts of 96 target genes. When designing the primers for those transcripts, identical primers were also removed. This reduced the number of primers

from 544 for all transcripts to only 159. These results show that by incorporating coverage at the 3' end of transcripts into primer design for targeted sequencing approaches, the number of primers can be significantly reduced. Furthermore, the coverage per gene is not an ideal indicator for coverage of a specific transcript.

3.11 Library generation using modified CEL-Seq2 approach with targeted primers v2

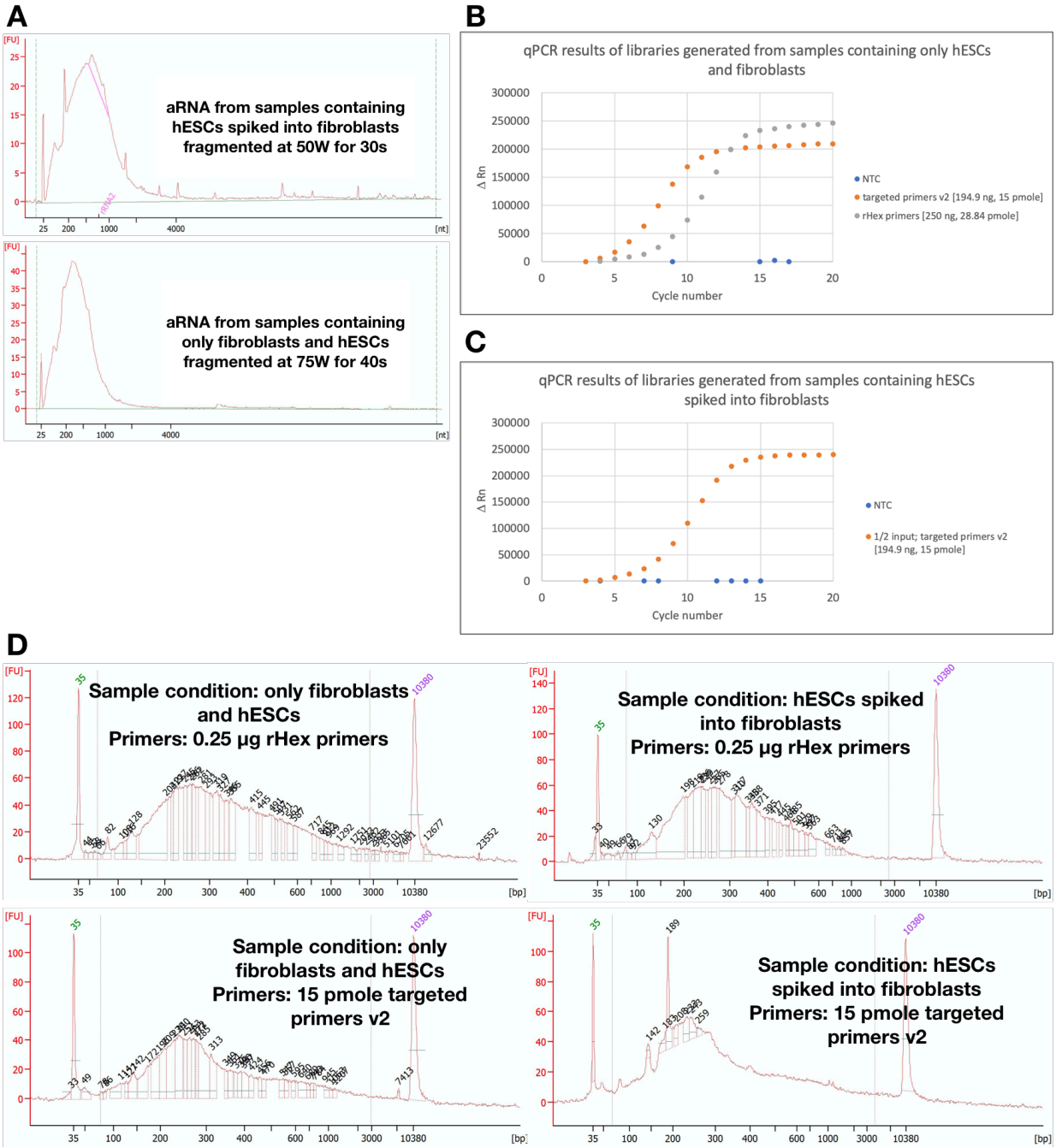


Figure 15 | A EGRAM of aRNA fragmented under indicated conditions **| B** qPCR results of libraries generated from samples containing only hESCs and fibroblasts **| C** qPCR results of libraries generated from samples containing hESCs spiked into fibroblasts **| D** cDNA libraries generated from indicated samples under the indicated conditions

I evaluated the performance of the newly designed primers to enrich for hESC-relevant transcripts. Therefore, I generated RNA-Seq libraries using the modified CEL-Seq2 protocol and the new gene specific primers. Specifically, I separately generated libraries for samples containing only hESC and Fibroblasts and the samples containing hESCs spiked in so that the targeted primers would not be absorbed by the hESC samples. As a control, I generated libraries following the conventional CEL-Seq2 protocol. Shearing of the aRNA yielded a size distribution with a peak around 500 nucleotides for the targeted samples and a peak at around 250 nucleotides for the untargeted samples (Figure 15A). The irregularities in the size distribution of the sample fragmented to a size of about 500 base pairs are probably caused by overloading. As expected, the cDNA library generated from samples containing only hESCs and fibroblasts using targeted primers was about fourfold less abundant than their untargeted counterpart (Figure 15B), probably mainly due to the amount of primers used (250 ng/28.84 pmol untargeted rHex primers and 194.9 ng/15 pmols targeted primers v2). Unfortunately, only the targeted library for the samples containing hESCs spiked into fibroblasts was quantifiable by qPCR, possibly due to a technical error (Figure 15C). However, it can be assumed that the quantity of the untargeted library was very similar to that of the library for the samples containing hESCs and fibroblasts only, since aRNA input and primer input were almost exactly the same. The bioanalyzer traces after PCR amplification strengthened this assumption, since both libraries were of about the same quantity after being amplified for 6 and 8 cycles, respectively (Figure 15D). The size distributions of targeted and untargeted libraries differed only slightly with the peak of the targeted libraries being sharper compared to the flat peaks of the untargeted libraries (Figure 15D). The size distribution of the targeted library generated from samples containing hESCs spiked into fibroblasts showed irregularities, which could be explained by the fact that some targeted genes amplified better than others (Figure 15D). The peaks of the size distributions of the libraries were all at around 250 nucleotides and the average sizes of the libraries were between 346 and 424 nucleotides. These results show that the newly designed primers in combination with the modified CEL-Seq2 protocol allow for robust generation of high-quality libraries, which are remarkably similar in terms of size distribution and quantity compared to conventional, untargeted libraries.

3.12 Sequencing analysis of libraries generated using modified CEL-Seq2 approach with targeted primers v2

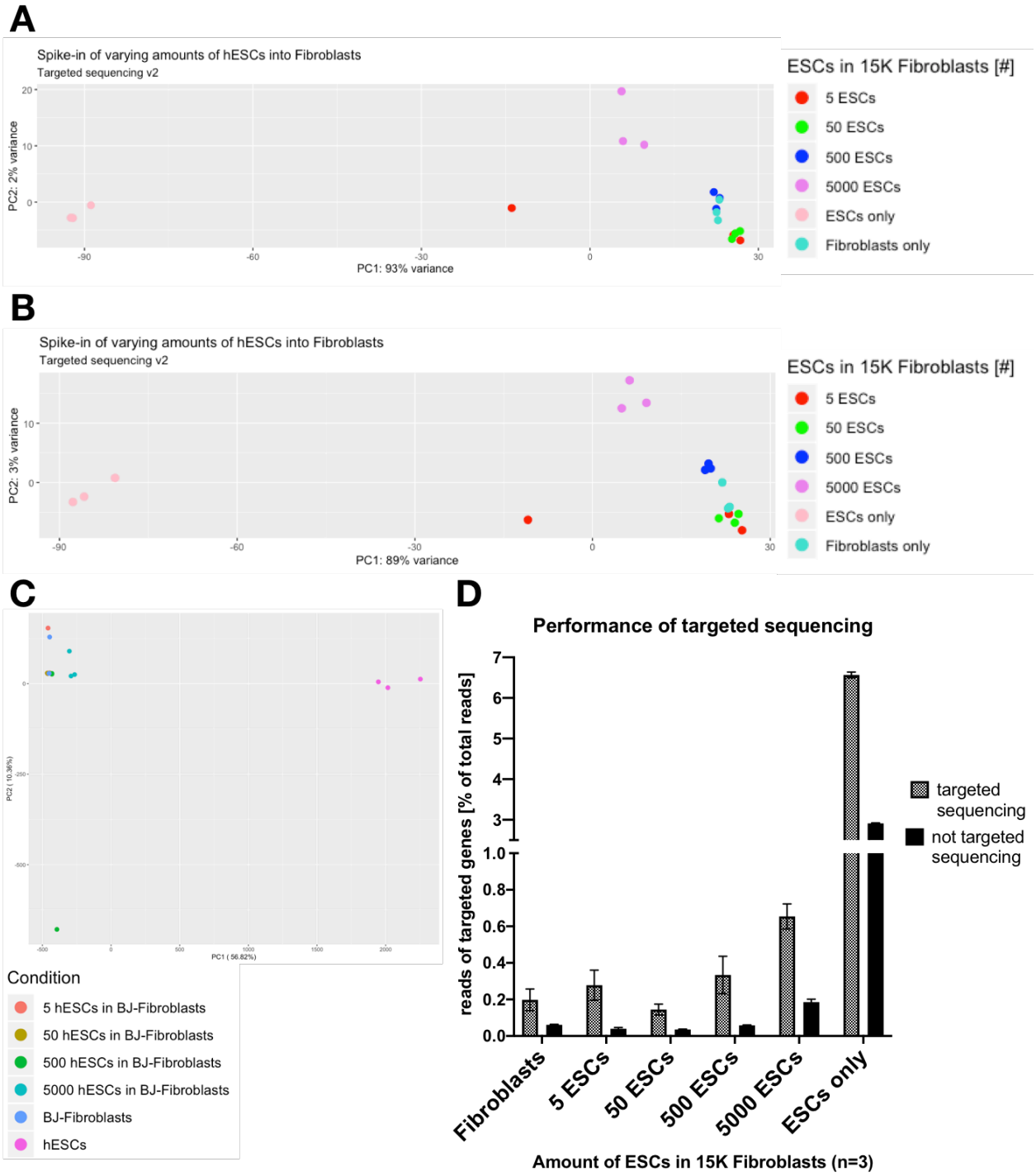


Figure 16 | A PCA using data of untargeted libraries taking into account all genes **| B** PCA using data of targeted libraries taking into account all genes **| C** PCA using data of targeted libraries taking into account only targeted genes **| D** Efficiency of targeted gene enrichment using targeted primers v2

I then tested whether the new primers and protocol enabled an improved resolution between different amounts of spike-ins of hESCs into fibroblasts. Therefore, I sequenced the libraries on the MiniSeq platform and analyzed the sequencing data using various software tools and custom python scripts. When taking into account all genes for principal component analysis, the samples prepared using the targeted primers (Figure 16B) showed better separation than the untargeted control (Figure 16A). This separation was especially pronounced between the samples containing 500 hESCs and the samples containing only fibroblast. However, there was no clear separation between samples containing less than 500 hESCs, indicating that there was not enough information to distinguish those samples from each other. In other words, the limit of detection – which is of course dependent on the sequencing depth – seems to be between 500 and 50 hESCs spiked into about 33,000 BJ fibroblasts when using targeted sequencing, and higher when using untargeted sequencing. Unexpectedly, the separation was less pronounced when taking into account only the targeted genes. In this case, only the samples containing 5000 hESCs or only hESCs were clearly separated from the rest of the samples, and one replicate containing 500 hESCs clustered very far away from the other replicates (Figure 16C). This could indicate that there might not be enough dimensionality contained in the data due the low number of target genes and the relatively low read counts when comparing with all genes. The enrichment for target genes was at least twofold and for some samples almost fourfold (Figure 16D). However, the overall portion of reads mapping to target genes was still quite low in absolute terms. It ranged from 0.2% for the fibroblast samples to about 7% in the hESC samples. Unsupervised clustering using only the information on targeted genes correctly separated samples only hESCs and 5000 hESCs when using untargeted sequencing data (Figure 17A). For the libraries generated by the targeted method, the samples containing only hESCs, 5000 hESCs and 500 hESCs are correctly clustered (Figure 17B). The enrichment for target genes is also apparent at the individual gene level. An overall increase of read counts per gene can be observed when increasing the number of hESCs in the sample, although this pattern is not very consistent, especially for lower amounts of hESCs such a 5 or 50 (Figure 17C). Collectively, these results show that the samples containing 500 hESCs could only be separated from the samples containing only fibroblasts using the new, targeted approach, which was not possible when using an untargeted approach at similar coverage. What is more, there was a significant enrichment for the targeted genes when considering aligned reads. However, the information contained in only the targeted genes did not seem enough to separate the samples containing fewer than 500 hESCs.

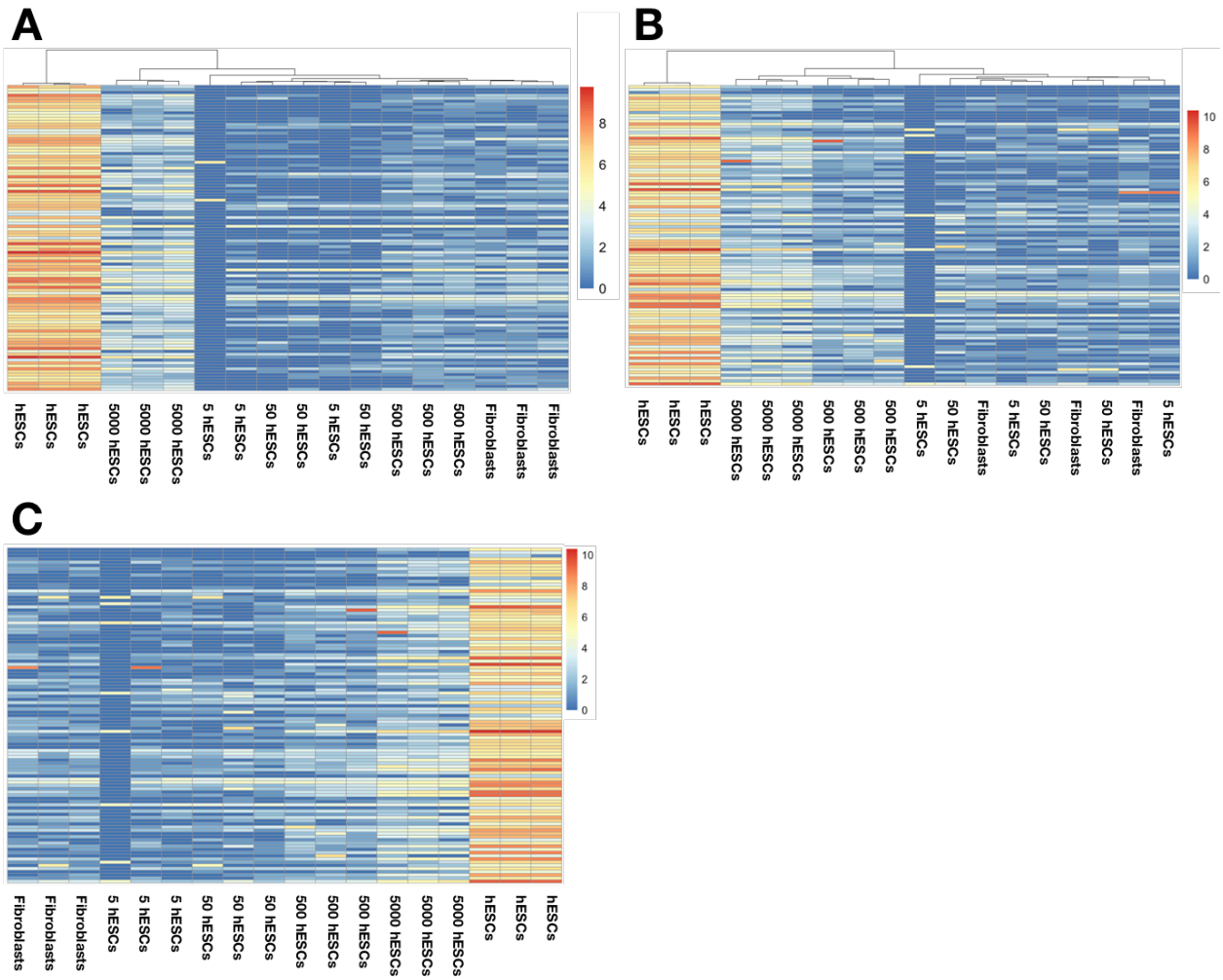
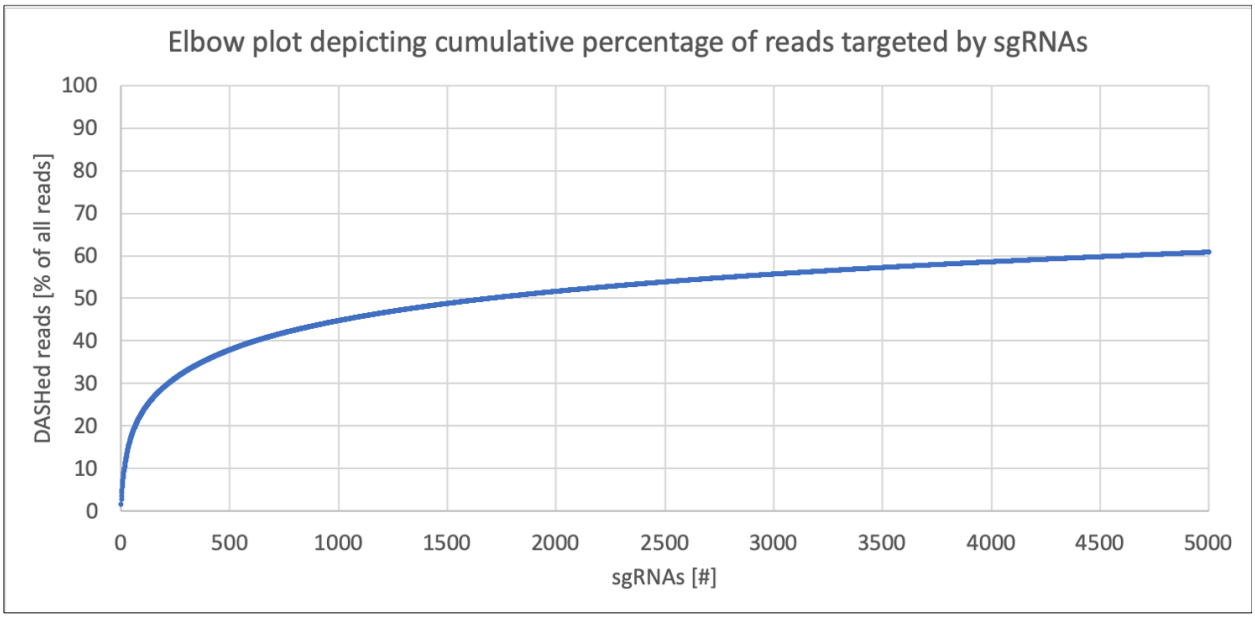


Figure 17 | A Unsupervised clustering using information of targeted genes obtained by untargeted sequencing approach. Read counts are log2(x+1) transformed. **| B** Unsupervised clustering using information of targeted genes obtained by targeted sequencing approach. Read counts are log2(x+1) transformed. **| C** Heatmap depicting log2(x+1) transformed normalized read counts obtained using targeted approach

3.13 Design of DASH sgRNA library

A



B

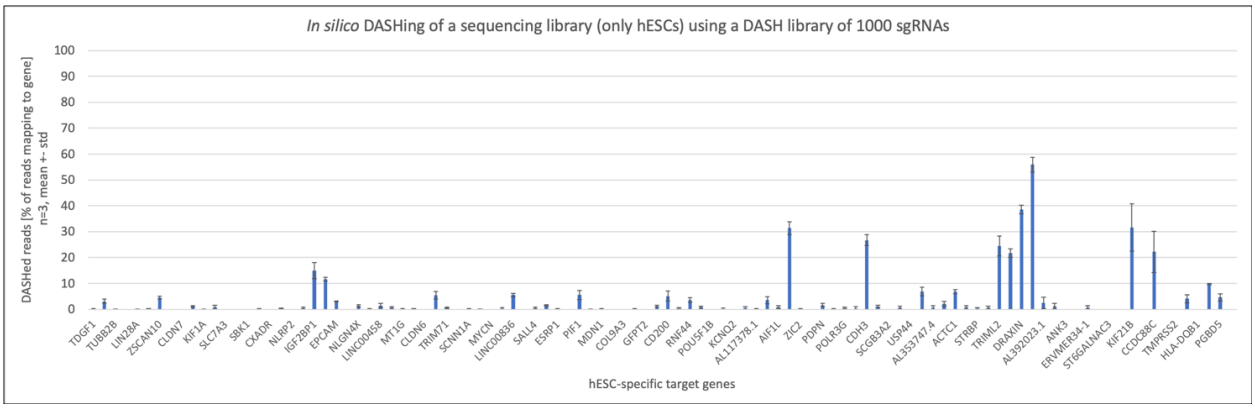


Figure 18 | A Elbow plot depicting the cumulative percentage of reads targeted by sgRNAs **| B** Assessing off-target effects of DASH library by in silico DASHing of a sequencing library (only hESCs) using a DASH library of 1000 sgRNAs

I then designed a library of sgRNAs to deplete reads mapping to genes abundant in fibroblasts and thereby increase the coverage of hESC-relevant genes. Therefore, I leveraged the sequencing data of samples containing only fibroblasts to design sgRNAs targeting the most abundant reads. I then performed DASH *in silico* against a sequencing library of a sample containing only hESCs to assess off-target effects against hESC genes of interest. The elbow plot depicting percentage of dashed reads against the number of sgRNAs ranked by number of reads hit shows that after about 500-1000 sgRNAs the curve begins to flatten (Figure 18A). This indicates that additional sgRNAs will only hit a relatively small number of reads, namely only around 0.01% of all reads per additional sgRNA. When selecting the first 1000 sgRNAs, theoretically 44.78% of all reads will be DASHed. In comparison, 5000 sgRNAs are in theory capable of DASHing 60.85% of all reads. While

the first 1000 sgRNAs did have off-target effects on some of the previously selected, hESC-specific genes, the relative number of DASHed reads per gene was quite low for most of the genes ranging from 0% to 58.27% and an average of 3.65% (Figure 18B). Collectively, these results show that it is not necessarily useful to design as many sgRNAs as possible to deplete fibroblast-specific reads in the sample, with the optimal tradeoff between number of reads dashed and chance of unwanted off-target hits being probably between 500 and 2000. When selecting 1000 sgRNAs, the number of off-targets hits seems to be negligible, while at the same time targeting 44.78% of all reads in the sample. This offers a promising approach to increase the coverage of hESC-relevant genes.

3.14 Depleting abundant fibroblast genes using DASH library

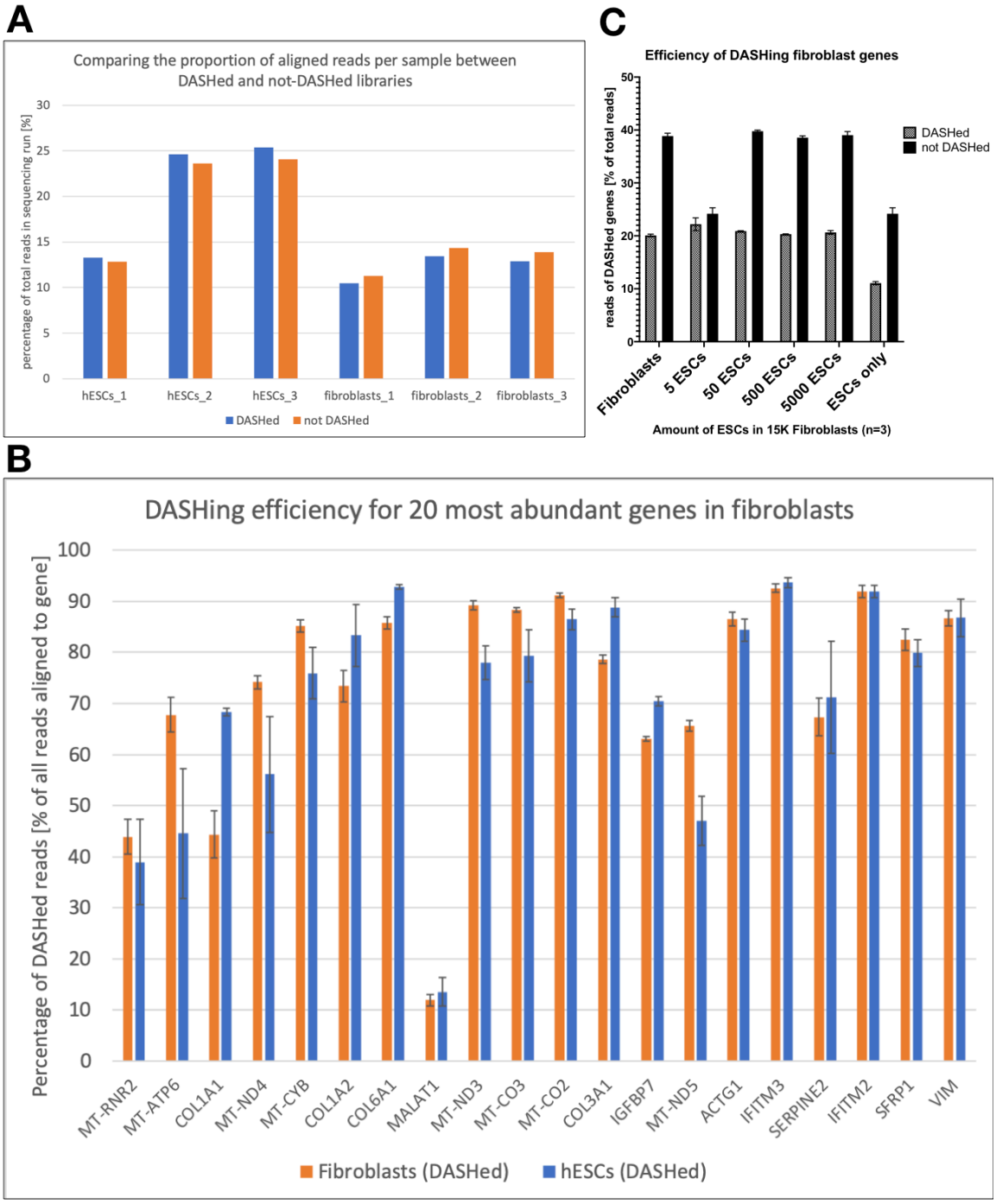


Figure 19 | A Comparing the proportion of aligned reads per sample between DASHed and not-DASHed libraries **| B** Efficiency of DASH approach on the 20 most abundant genes in fibroblasts **| C** Efficiency of DASHing DASH targets in indicated samples

I then tested the performance of the DASH library to improve the detection and quantification of hESCs spiked into fibroblasts using RNA-Seq. Therefore, I depleted libraries containing hESCs spiked into fibroblasts at varying amounts of reads abundant in fibroblasts using 1000 sgRNAs. After sequencing the libraries on the MiniSeq platform, I analyzed the sequencing data using

various software tools and custom python scripts. Unexpectedly, there was no noticeable difference in the proportions of reads between samples containing only fibroblasts or only hESCs after DASHing (Figure 19A). The depletion of reads was substantial when looking at the 20 most highly expressed genes in fibroblasts, showing a reduction of aligned reads to values between 12% and 93% in the libraries obtained from fibroblasts and between 14% and 93% in the libraries obtained from hESCs (Figure 19B). This indicates that the library of sgRNAs was successful in DASHing out fibroblast-abundant reads from the library and at a high efficiency. Furthermore, the most highly expressed genes that had about 40% of the total reads aligned to them I will from here on call 'DASH targets', since the DASH library was predicted to target about 40% of the reads *in silico*. It is important to note that this assumption could lead to an underestimation of DASHing efficiency, since not all of those genes have to be *bona fide* DASH targets. It should however give a good approximation. The percentage of reads aligned to those DASH targets was reduced about twofold from 40% to 20% after DASHing the library containing only fibroblasts (Figure 19C). Reductions of a similar extent were observed for the libraries of 50, 500 and 5000 hESCs spiked into fibroblasts, but not for the library containing 5 hESCs spiked into fibroblasts. For this sample, no significant reduction could be observed, possibly indicating technical faults. Unexpectedly, there was also a twofold reduction after DASHing from 25% to about 12% of the reads aligned to the DASH targets in the sample containing only hESCs. This could indicate that some DASH targets are not specific to fibroblasts since they are for example of mitochondrial origin. The number of reads aligned to selected, hESC specific genes did not change significantly in any of the samples when comparing DASHed (Figure 20A)and not-DASHed (Figure 20B) samples, indicating that unwanted off-target effects of the DASH library seem to be negligible. However, principal component analyses on both the gene expression data obtained from not-DASHed (Figure 20C) and DASHed (Figure 20D) libraries yielded no clear separation between samples containing 500 or less hESCs spiked in. Surprisingly, the samples containing 500 hESCs spiked into fibroblasts and the ones containing only fibroblasts clustered closely together, as well as the samples containing 5 hESCs and the ones containing 50 hESCs. There was no notable difference in the PCA plot on the DASHed samples, except slightly tighter intra-sample clustering of the fibroblast samples and the samples containing 500 hESCs spiked in. Together, these results show that the DASH approach using a library of 1000 sgRNAs can successfully deplete a large portion of specific reads from a cDNA library. However, this depletion did not improve the separation of samples according to the amount of hESCs spiked into fibroblasts using transcriptomic data.

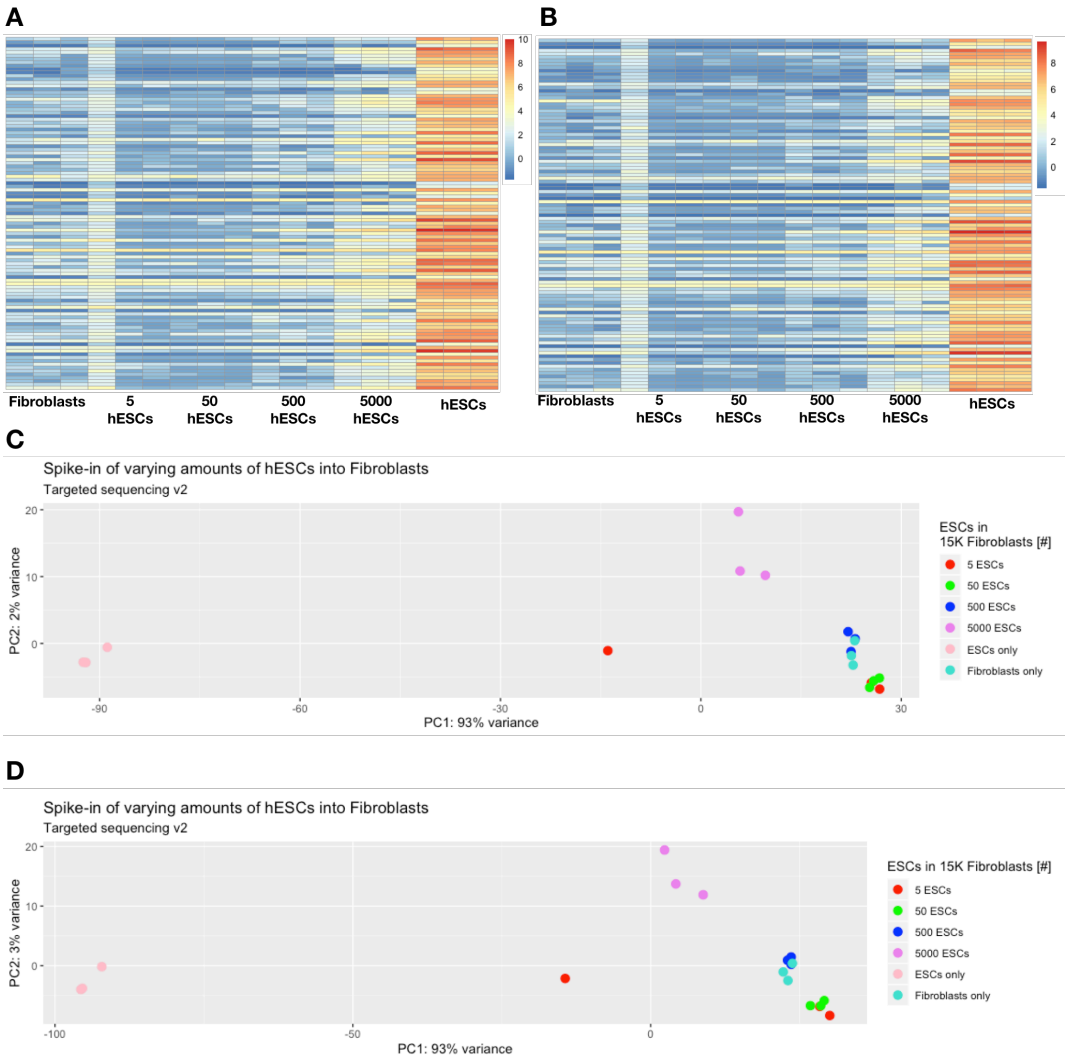


Figure 20 | A Heatmap depicting regularized log transformed read counts of hESC-specific genes in DASHed samples **| B** Heatmap depicting regularized log transformed read counts of hESC-specific genes in not-DASHed samples **| C** PCA using data of not-DASHed libraries taking into account all genes **| D** PCA using data of DASHed libraries taking into account all genes

3.15 Reprogramming fibroblasts in a 96-well plate using a two-component vector system

I then tested the limits and performance of reprogramming human BJ fibroblasts under feeder-free conditions. Therefore, I seeded different amounts of human BJ fibroblasts into a coated 96-well plate, transduced them with my lentiviral vectors and reprogrammed them. As a control, I also seeded 100.000 BJ fibroblasts into a 6-well plate under the same conditions. Double selection using puromycin and blasticidin left only GFP+ cells in the GFP-control wells, indicating that the fluorescence and selection markers of the generated vectors are functional (Figure 21).

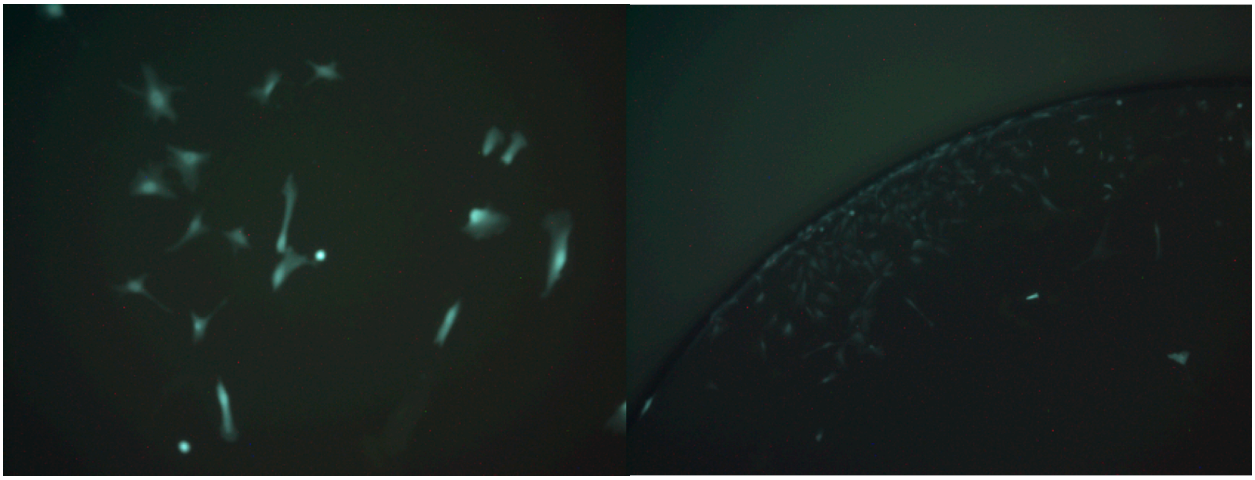


Figure 21 | BJ fibroblasts transduced with VMS026 expressing GFP after 7 days of selection with blasticidin and puromycin

Furthermore, the cells transduced with the two vectors expressing the Yamanaka factors started changing their morphology between day 4 and day 6 (Figure 22). They first became more elongated and then rounder. This is in agreement with previous results about the process of reprogramming[75].

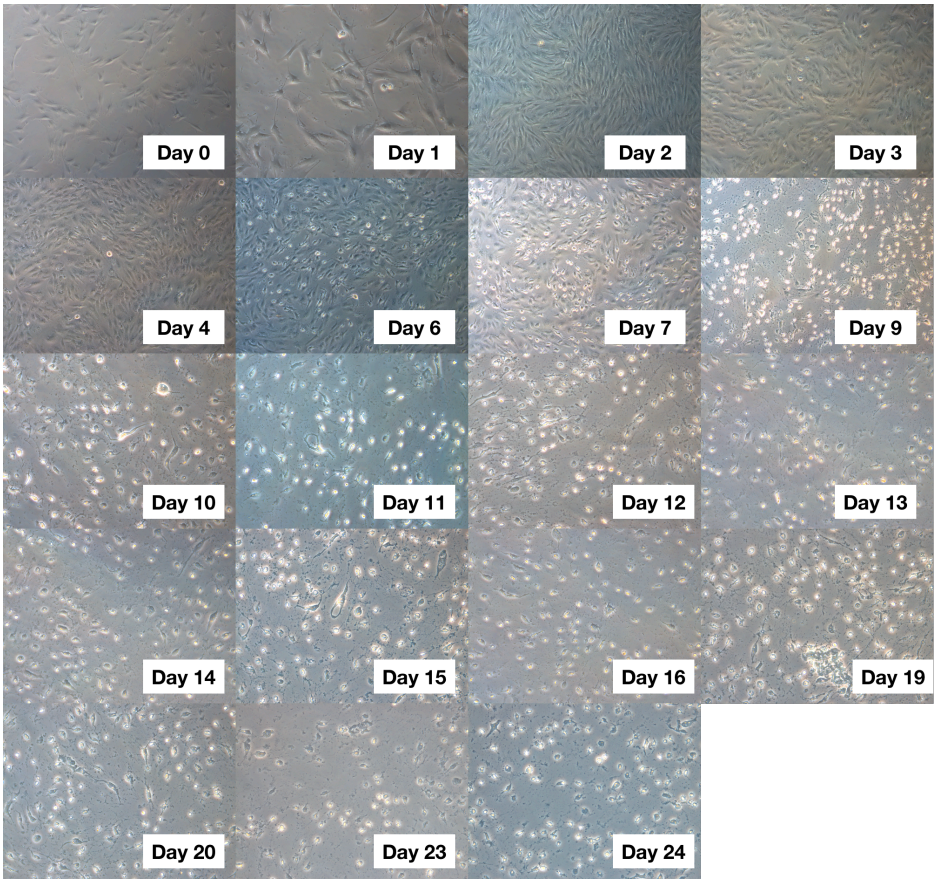


Figure 22 | Representative light-microscope pictures of reprogramming fibroblasts in 6-well plate under feeder-free conditions. Fibroblasts are transduced with VMS028 and VMS005.

Unexpectedly, almost all cells were dead between day 7 and day 9 of the reprogramming procedure in the 96-well plates, independent of the vectors used and the initial seeding density (Figure 23). This could indicate that the multiplicity of infection, the cell number and other conditions were not optimal in the 96-well plate wells.

Figure 23 | Representative light-microscope pictures of reprogramming fibroblasts in 96-well plate under feeder-free conditions. Fibroblasts were seeded at indicated densities and transduced with VMS005 and indicated vector.

There were many cells alive in the well of the 6-well plate and they kept changing their morphology as expected (Figure 22). After 24 days, despite there being small, round cells in the well, there were no characteristic iPSC colonies observable. This might indicate that the conditions for reprogramming might have been suboptimal in one way or another. Cells transduced with only the VMS005 vector – which contains all Yamanaka factors except Oct4 – and stably selected with blasticidin also showed an elongated and somewhat mesenchymal morphology (Figure 24). Together, this data indicates that while the generated vectors are functional, the detailed conditions for reprogramming in a high-throughput manner using few cells in 96-well plates need further optimization.

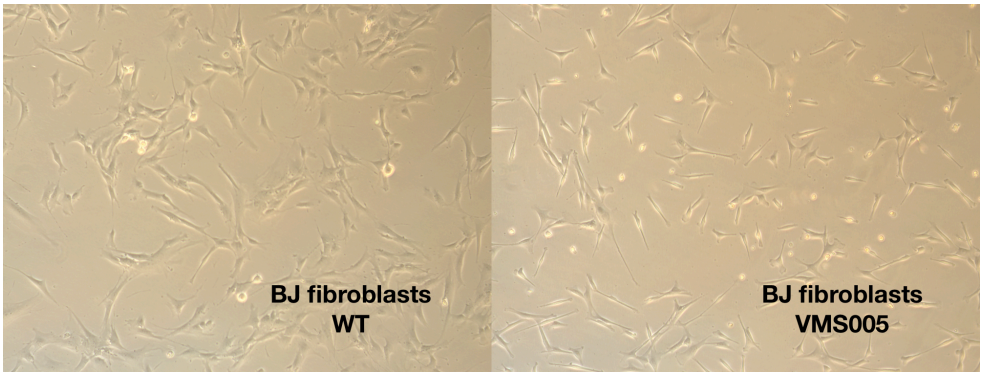


Figure 24 | Representative light-microscopy pictures showing non-transduced BJ fibroblasts and BJ fibroblasts transduced with VMS005 and stably selected

3.16 Generating a pooled lentiviral hORFeome v8.1 library

To prove the concept of my new screening system, I also needed to generate a library of perturbations. Therefore, I shuttled all ORFs of the human ORFeome v8.1 collection into a suitable destination vector (pLEX307) via Gateway cloning. I first attempted this by shuttling them in a pooled manner using only one cloning step (Figure 25).

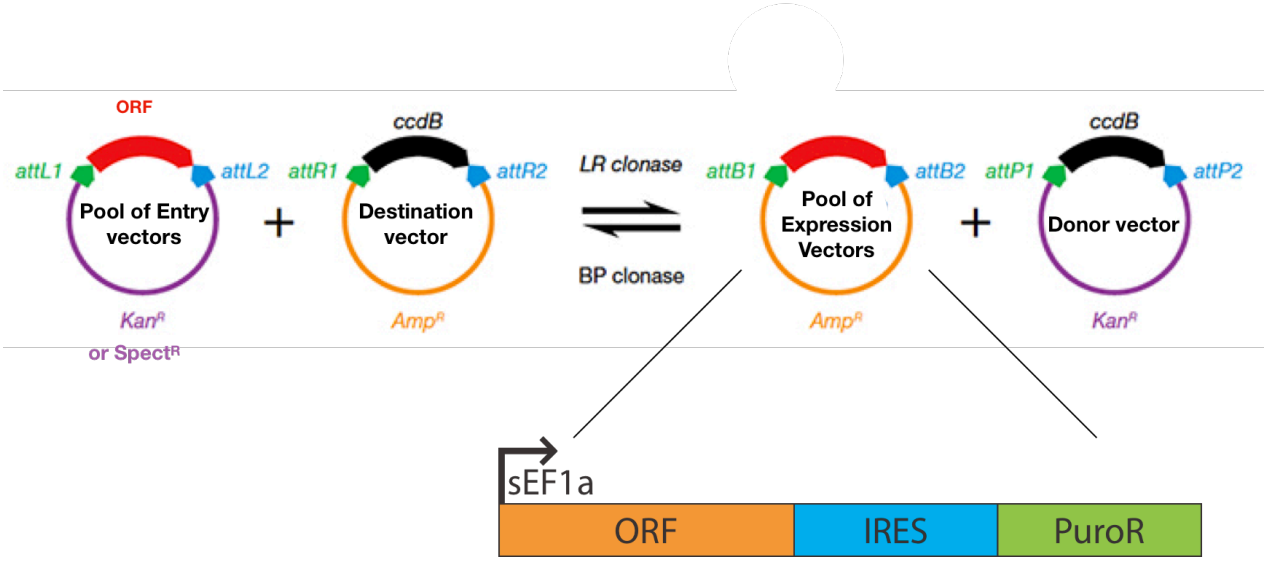


Figure 25 | Summary of pooled Gateway LR reaction

I checked 30 colonies by Sanger sequencing to estimate the unwanted recombination rate. Then, I generated cDNA libraries from the initial plasmid pool of entry vectors as well as from the cloned expression vectors using a Tn5 based protocol and sequenced them on a MiniSeq system. Out of the 30 colonies, all of them had successfully recombined and received an insert, suggesting that less than 3.33% of the colonies were background colonies without a proper insert. The efficiency of the Gateway reaction was highly similar between 660 ng and 1000 ng destination vector input, namely around 50.000.000 colonies or plasmid molecules (Table 72). This efficiency corresponds to a coverage of about 3900x the 12790 ORFs that were shuttled.

Sample	Colonies in 1:500,000 dilution	Colonies in 1:50,000 dilution	Colonies in 1:5,000 dilution
1000ng pENTR223/hORFeome 8.1 + 660ng pLEX307	100	~900	too many to count accurately
1000ng pENTR223/hORFeome 8.1 + 1000ng pLEX307	96	824	too many to count accurately

Table 72 | Initial quality control data of the pooled Gateway cloning reaction of hORFeome v8.1

Next generation sequencing data of the pool of ENTR vectors containing the ORFs showed the quantitative composition of the initial plasmid pool used in the Gateway reaction (Figure 26A). Out of the 12790 ORFs, 114 (0.89%) had no reads aligned, suggesting that they were not represented in the plasmid pool. Furthermore, only 31 ORFs (0.24%) made up about 6.3% of all TPM and were within a tenfold range of the ORF with the highest TPM. 11260 ORFs (88.04%)

were within a tenfold range between 0.2 and 0.02 TPM, with only 200 ORFs (1.56%) being above 0.2 TPM. This data suggests that already the initial plasmid pool of ENTR vectors was substantially skewed. Expectedly, the quantitative composition of the resulting pool of expression vectors was somewhat more skewed (Figure 26B). In this pool, 120 ORFs (0.94%) had no reads aligned. 38 ORFs made up about 12.8% of all TPM and were within a tenfold range of the ORF with the highest TPM. Moreover, 18 of the 31 ORFs (58.06%) which were within the top tenfold range of the ENTR vector pool were also contained within those 38 ORFs (18/31 = 47.37%). 10088 ORFs (78.87%) had a TPM value between 0.2 and 0.02, and only 428 ORFs (3.35%) had a TPM value above 0.2. Unexpectedly, there was no correlation between the size of an ORF and its abundance in the plasmid pool of ENTR vectors (Figure 26C) and a very slight and negative correlation between size and abundance in the plasmid pool of destination vectors (Figure 26D). However, there was a clear and positive correlation between abundance of an ORF in the plasmid pool of ENTR vectors and the plasmid pool of destination vectors (Figure 26E). Together, these data suggest that a pooled Gateway reaction promotes differences in abundances of an initial library. However, relative differences are mostly retained. Furthermore, the abundance of a plasmid in a plasmid pool does not seem to depend on the size of the ORF it contains.

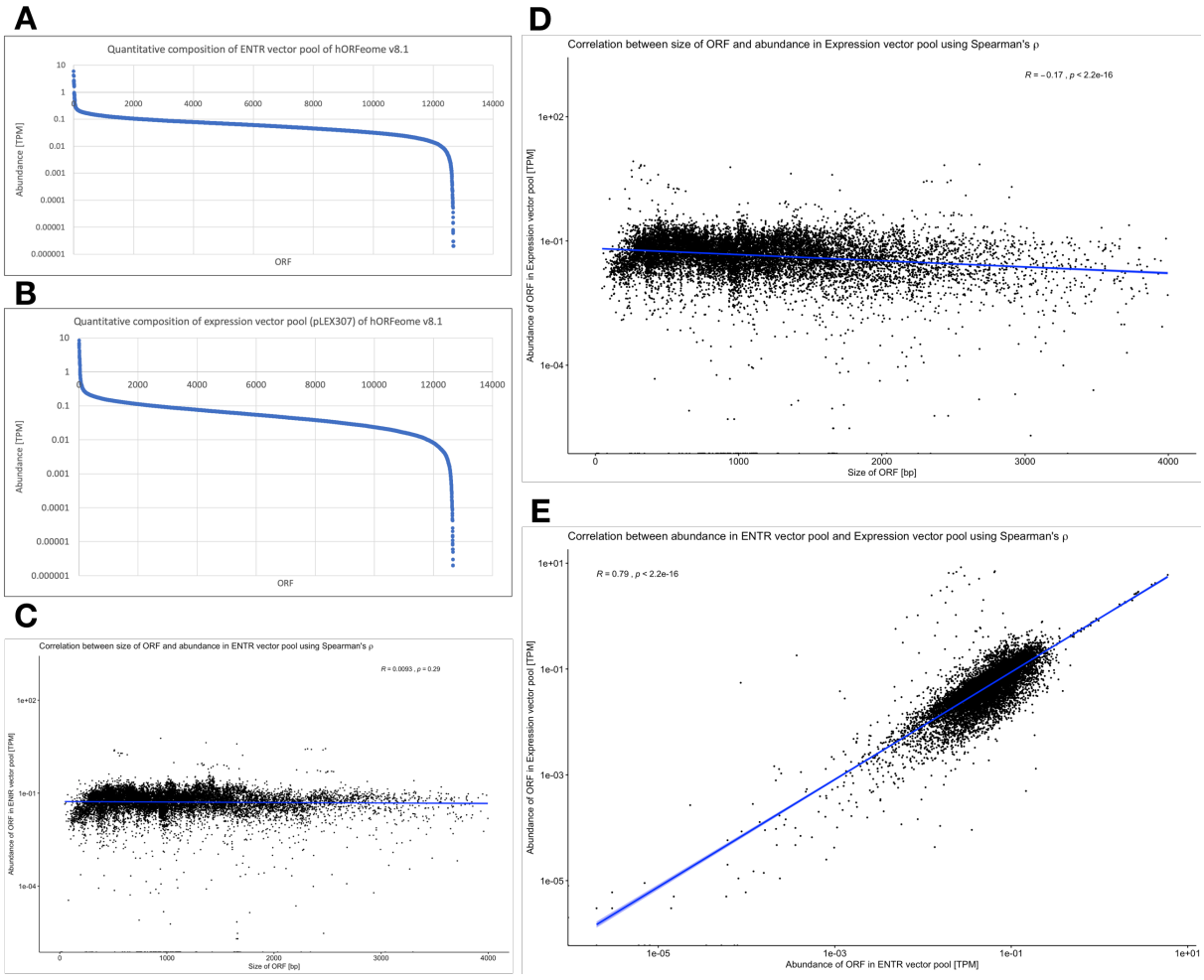


Figure 26 | A Quantitative composition of ENTR vector pool of hORFeome v8.1 **| B** Quantitative composition of expression vector pool of hORFeome v8.1 **| C** Scatter plot depicting correlation between the size of an ORF and its abundance in the ENTR vector

pool | **D** Scatter plot depicting correlation between the size of an ORF and its abundance in the expression vector pool | **E** Scatter plot depicting correlation between abundance of ORFs in the ENTR vector pool and abundance in the expression vector pool

3.17 Generating a pooled lentiviral ORFeome library consisting of epigenetic modifiers and transcription factors

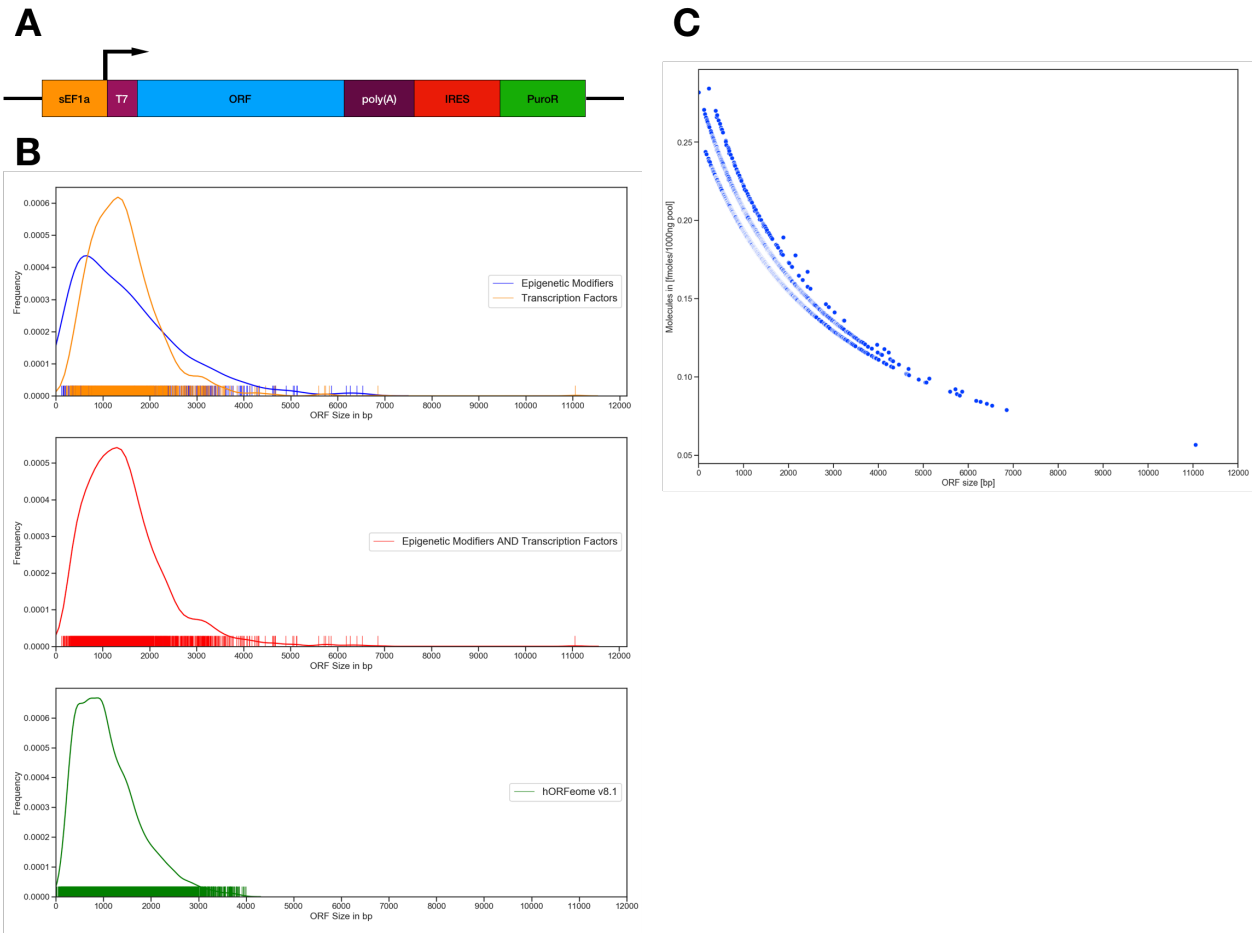


Figure 27 | A Conceptual vector map of VMS009 **| B** Comparison of the size distributions of ORFs in the epigenetic modifiers subcollection, the transcription factor subcollection, the union of the epigenetic modifiers subcollection and the transcription factor subcollection and the hORFeome v8.1 collection **| C** Theoretical molecular abundance of each ORF in the EMTF library when assuming equal masses

To potentially decrease skewing of the library and keep random side effects of the pooled gateway reaction to a minimum, I scaled down the number of ORFs included from 12,970 to 2,065. To do so, I combined a sub library consisting of epigenetic modifiers and a sub library consisting of transcription factors, since almost all transitions of cell state or type are primarily driven by epigenetic modifiers and transcription factors[26]. I designed and cloned a new Gateway destination vector containing new useful features such as a short EF1a promoter for reduced size, an IRES driving the expression of a puromycin-resistance cassette for selection directly linked to the expression of the ORF and a poly(A) stretch and a T7 promoter that can potentially be used for more efficient and economic sequencing of plasmid mixtures (Figure 27A). Then, I cloned the plasmid pool of ORFs into this new vector backbone in a pooled Gateway cloning reaction. I investigated 14 colonies by Sanger sequencing to estimate the unwanted recombination rate.

Finally, I prepared sequencing libraries from the plasmid pools, sequenced them and analyzed the sequencing data to assess their quantitative composition. The size distributions of the new epigenetic modifiers and transcription factor (EMTF) library and the hORFeome v8.1 library were quite similar, with the EMTF library having a few ORFs larger than 4,000 base pairs and in general being a bit more broad in its size distribution (Figure 27B). This leads to a roughly fivefold range of molar amounts of the individual plasmids in the EMTF library when assuming equal masses (Figure 27C). The 14 colonies I assessed by Sanger sequencing all had ORFs inserted, suggesting an insertion success rate of at least 92.86% (Table 73). Interestingly, the size of the ORFs in those colonies was between 351 and 2988 base pairs, with an average of 1,605.92 base pairs, which is just above the average size of the EMTF library (1,473 base pairs). This data argues against a potential preference in recombination for smaller ORFs.

Colony [#]	ORF symbol	ORF size [bp]
1	GFI1	1266
2	MED11	351
3	FHL2	1011
4	PIAS2	1719
5	PEPD	1482
6	KCNIP4	753
7	GSG2	2394
8	SPRY1	958
9	ZFP28	2605
10	TLE3	2317
11	TRIM38	1395
12	FOXO3	2020
13	ELK3	1224
14	PHC3	2988

Table 73 | Identity of colonies picked from plates resulting from the pooled Gateway reaction. The size of the respective ORF is indicated.

The pooled gateway reaction yielded around 3.45E+07 to 3.52E+07 bacteria that took up a recombined construct, which translates to a coverage of the library between 16,690.8563x and 17,029.5114x (Table 74).

Dilution factor	Colonies [#]	efficiency (bacteria with successful uptake of construct)	coverage
5000	too many to count accurately	NA	NA
50000	704	3.52E+07	17029.5114
500000	69	3.45E+07	16690.8563

Table 74 | Initial quality control data of the pooled Gateway cloning reaction of the EMTF ORF library

The sequencing data of the pool of ENTR vectors showed a surprisingly equal distribution of individual constituents of the library (Figure 28A). 1,857 ORFs (89.93%) were within a tenfold

range of the most abundant ORF. Only 30 ORFs (1.45%) had no reads aligned to them, indicating that they were not included in the library for some reason. Furthermore, only 131 ORFs (6.34%) had a TPM of less than 0.1, which is about the point when the curve falls of significantly. However, the sequencing data of the pool of expression vectors showed a more skewed distribution of abundances of individual ORFs in the mixture (Figure 28B). The TPM value of only 121 ORFs (5.86%) was within a tenfold range of the highest TPM value. Moreover, 194 ORFs (9.39%) had a TPM value of less than 0.01, which is where the curve begins to fall off. 46 ORFs (2.23%) had no reads aligned to them. In agreement with previous results from the pooled shuttling of hORFeome v8.1, there was a correlation between abundance of an ORF in the pool of ENTR vectors and in the pool of expression vectors (Figure 28C). Furthermore, there was a slightly negative correlation between the size of an ORF and its abundance in both the library of ENTR vectors (Figure 28D) and the library of expression vectors (Figure 28E). Together, these data indicate that even a smaller, more tightly distributed library of ENTR vectors will become skewed by a pooled Gateway reaction. Moreover, it suggests that in addition to a small influence of the size of an ORF, the skewing seems to be dependent on the initial abundance of a plasmid.

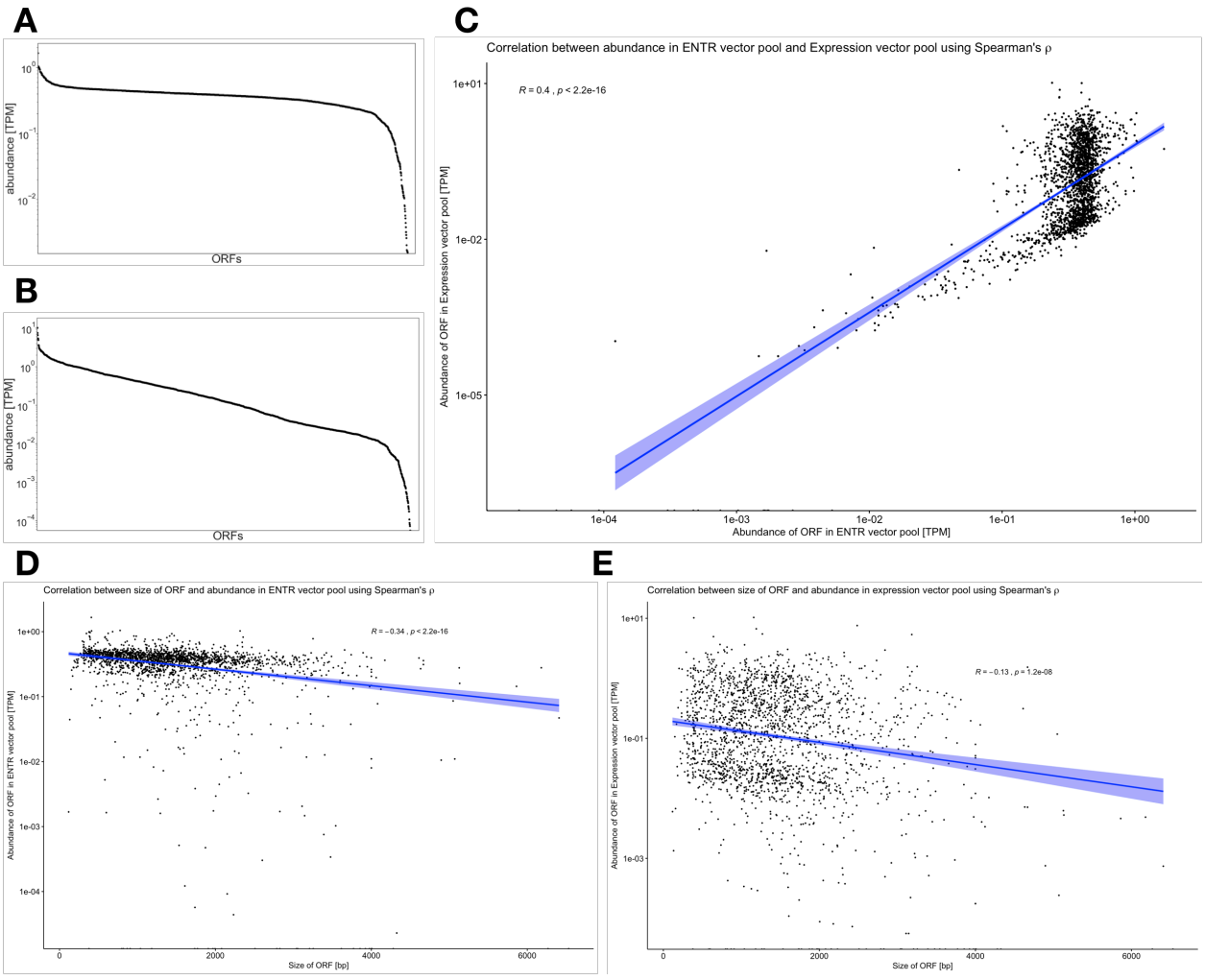


Figure 28 | A Quantitative composition of ENTR vector pool of EMTF library **| B** Quantitative composition of expression vector pool of EMTF library **| C** Scatter plot depicting correlation between abundance of ORFs in the ENTR vector pool and abundance

in the expression vector pool **|** **D** Scatter plot depicting correlation between the size of an ORF and its abundance in the ENTR vector pool **|** **E** Scatter plot depicting correlation between the size of an ORF and its abundance in the expression vector pool

3.18 Converting a pooled lentiviral ORFeome library consisting of transcription factors and epigenetic modifiers to the Phenosudoku format

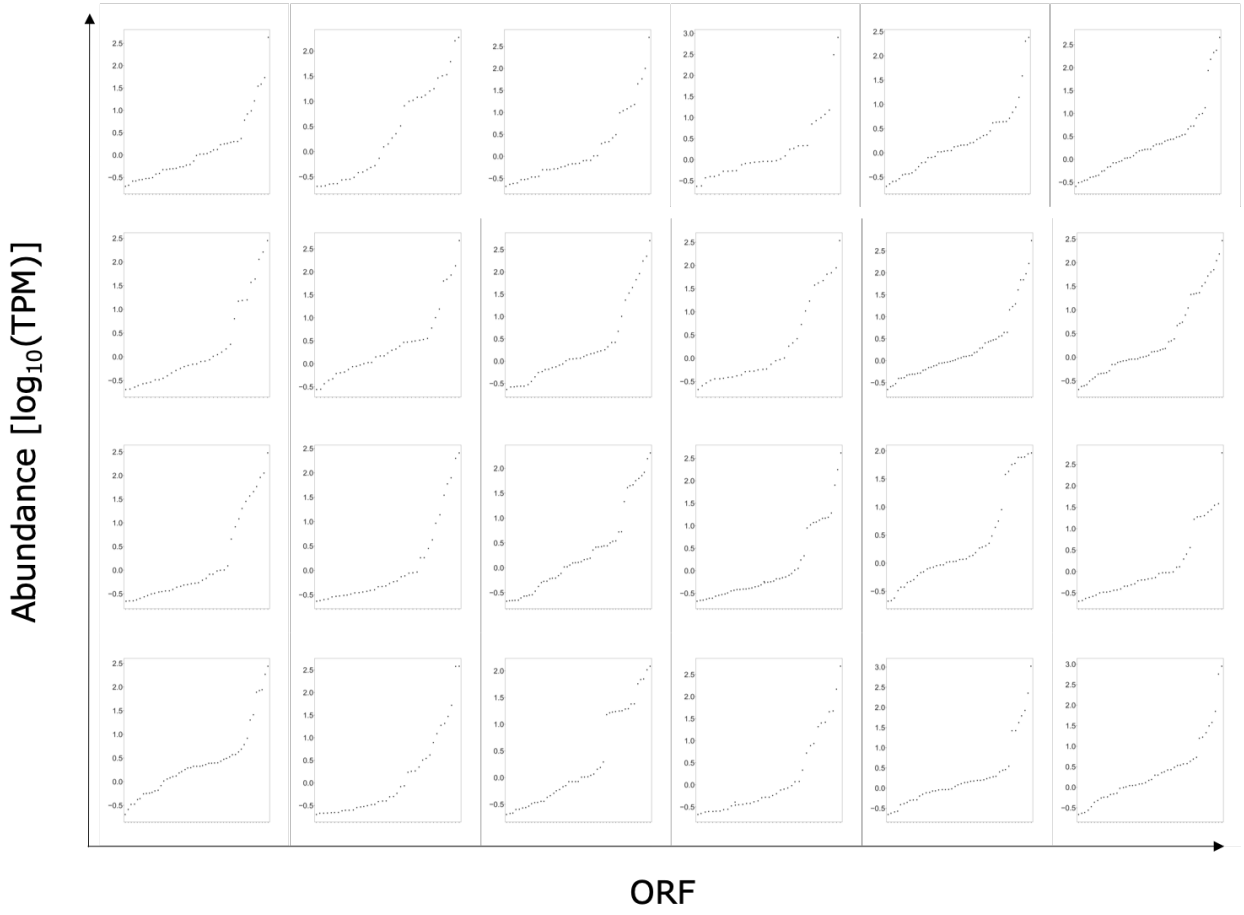


Figure 29 | Quantitative composition of the 24 analyzed wells of the EMTF library in the Phenosudoku format. Only ORFs with TPM > 0.2 are shown.

The library of expression vectors was then converted into a format compatible with the novel screening paradigm called Phenosudoku and assessed the quality of this new library. To do so, I randomly allocated sub pools of the library to wells of 96-well plates and investigated their quantitative composition by next generation sequencing of 24 wells. Expectedly, the composition of the wells was quite skewed, with very few plasmids making up most of the well and sometimes major differences in abundance of the individual constructs (Figure 29). This result is in agreement with the fact that the initial library of expression vectors also showed a skewed distribution of abundances. By counting colonies after plating different dilutions of the glycerol stock of the initial library of expression vectors, I estimated the average number of constructs per well to be about 31 (Table 75).

Dilution factor	Volume plated [µl]	colonies	CFU/ml (stock)	CFU/well
5,280,000	1,200	16	7.04E+07	16
5,280,000	50	28	1.48E+08	28
528,000	50	333	1.76E+08	33.3

Table 75 | Estimation of number of ORFs/well by estimating CFU

However, the sequencing data of the wells showed that there was an average of 9.29 constructs per well, with a maximum value of 15 and a minimum value of 4 (Figure 30).

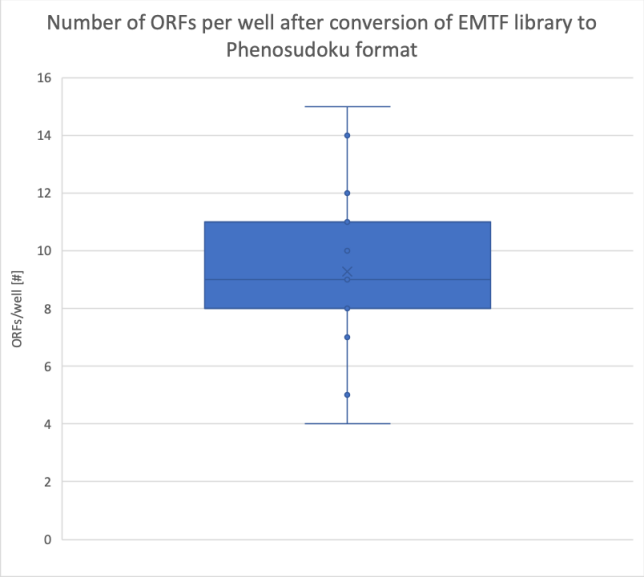


Figure 30 | Number of ORF constructs per well after conversion of EMTF library to Phenosudoku format. 24 wells were analyzed.

It must be noted that setting the threshold to determine which plasmids were or were not in a well was quite challenging for some wells, due to the sometimes-gradual decrease of reads (Figure 29). This was possibly caused by PCR artifacts. For the shown analyses, the threshold for presence of a construct was set at around 1 TPM or when there was a steep decrease in abundance. For the sake of completeness, the graphs shown in Figure 29 depict all ORFs with a TPM of at least 0.2. Some ORFs were much more prevalent than others. Only 21 ORFs were among the top 2 most abundant in the 24 analyzed wells (Table 76).

ORF	Size [bp]	ENTR pool [TPM]	Expr. Pool [TPM]	ENTR pool [rank]	Expr. Pool [rank]	Occurrence in top 2
NR2E1	1158	0.237959	10.334811	1727	1	2
PPP1R13L	2484	0.271785	7.57467	1626	3	1
LHX9	1191	0.395305	6.681825	889	4	12
GFI1	1269	0.270035	5.482373	1632	5	2
TCF23	642	0.454356	4.947352	415	7	8

HMG20A	1041	0.424944	3.61954	632	9	1
KRCC1	777	0.464834	2.973542	347	13	4
HOXA6	702	0.521225	2.701848	138	18	1
ASCL3	543	0.29089	2.641488	1552	21	1
ZNF689	1500	0.416288	2.516174	697	24	1
CTNND1	1830	0.388114	2.433659	954	30	3
TSG101	1173	0.36563	1.970884	1158	53	1
PMF1	615	0.732738	1.738263	31	67	1
RAI1	3132	0.314061	1.537983	1467	88	1
CHMP3	669	0.460585	1.53603	378	89	1
POLE4	351	0.403186	1.475827	823	96	2
NANOG	918	0.356882	1.300096	1223	126	1
TBX22	1662	0.354276	0.990918	1239	217	1
C17orf49	519	0.447126	0.81174	468	286	2
TFAP4	1017	0.424812	0.643125	633	362	1
MSS51	1380	0.431948	0.459199	574	506	1

Table 76 | Summary of ORFs which are among the top 2 most abundant in the phenosudoku wells.

For example, LHX9 was the most abundant ORF in 9 wells and was one of the top 2 most abundant ORFs in 12 wells. However, LHX9 was not particularly abundant in the pool of ENTR constructs, suggesting that the pooled gateway reaction introduces some factor of chance. LHX9 was the 4th most abundant ORF in the expression vector pool, which possibly explains its prevalence among the analyzed wells. Similarly, TCF23 was the 7th most abundant ORF in the pool of expression vectors and was among the top 2 most abundant ORFs in the analyzed wells 8 times. The average size of the enriched ORFs was 1170.14 base pairs, which is lower than the average size of the library (1473 base pairs). This might indicate that smaller size can favor the abundance of a particular ORF. The average rank in the expression vector pool of the enriched ORFs was 96.43, suggesting that abundance in the expression vector pool is a good indicator for enrichment in the Phenosudoku format. Interestingly, the TPM values of 16 of the 21 enriched ORFs were within a tenfold range of the most abundant ORF in the expression vector pool, further strengthening the link between abundance in the expression vector pool and the Phenosudoku library. The large variance in the number of constructs per well, the fact that the average number of constructs per well is much lower than estimated by counting colonies and the fact that certain constructs are abundant in a many wells suggest that the inherent skewness of the initial library lead to multiple bacteria containing the same constructs being transferred into the same well. This then might have caused those constructs to take over most of the wells.

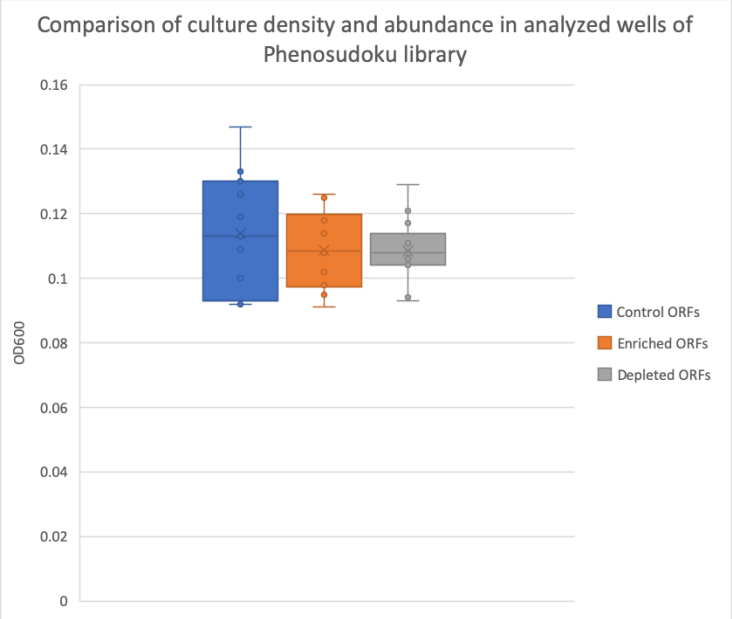


Figure 31 | Comparison of culture density and abundance in the analyzed wells of the Phenosudoku library

Since this prevalence of some plasmids might be explained by enhanced growth rates of the bacteria that replicate those dominant constructs, I tested this hypothesis by comparing optic density (OD) values of individual cultures to their abundance in the library. Therefore, I choose ORFs that were under the top 2 most abundant genes in a well (enriched ORFs), ORFs that had no aligned reads in any of the analyzed wells (depleted ORFs) and ORFs that had a TPM value of below 2 in their respective well (control ORFs). I then measured and compared their OD at 600 nm. Unexpectedly, there was no significant difference in the OD measurements between the 3 groups (Figure 31). This implies that bacterial density at the culture stage was not the root cause of the skewing of the library. Although no definite conclusion can be drawn from these results, they suggest that bacterial growth is probably not the major factor for abundance of a particular ORF in the library.

3.19 Generating a pooled lentiviral ORF library consisting of a sub-selection of transcription factors

Since the pooled gateway cloning step seemed to distort the library to an unacceptable extent, a smaller sub-library of the EMTF library was prepared. Therefore, 260 transcription factors were selected and each ORF was individually cloned into the lentiviral vector by Gateway cloning. Out of the 260 selected transcription factors, 241 yielded colonies after the LR clonase reaction. Cultures derived from those 241 colonies were pooled to generate a library of lentiviral vectors. The size distribution of this library – from this point onward referred to as TF241 – was similar to those of the EMTF library, with an average ORF size of 1,428.69 base pairs and ORF sizes ranging from 246 to 3,477 base pairs (Figure 32A).

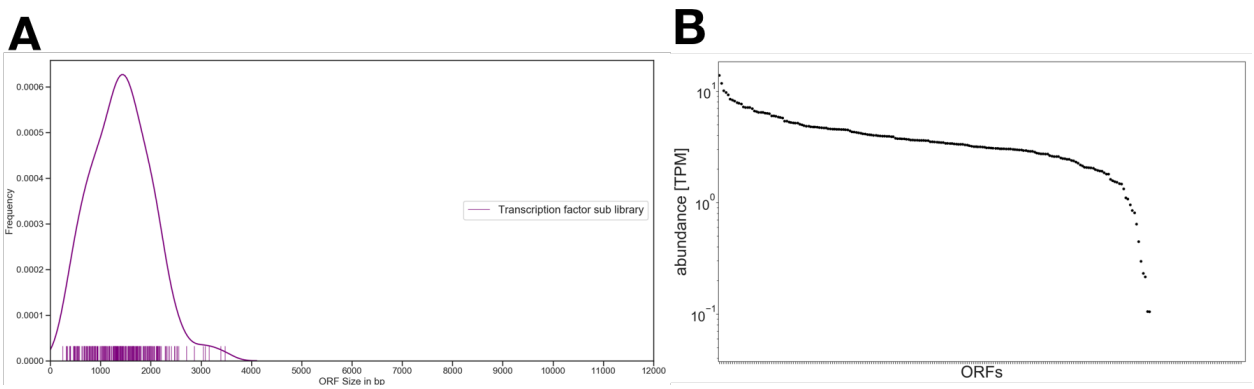


Figure 32 | A Size distribution of the 241 ORFs in the transcription factor sub library (=TF241) **| B** Quantitative composition of expression vector pool (VMS009) of TF241

The lentiviral library was quality controlled by next generation sequencing. Out of 241 ORFs, 26 (10.79%) had no reads aligned in the lentiviral library pool, indicating their absence from the final library (Figure 32B). The quantitative composition of the lentiviral library was highly homogeneous, with the TPM values of 186 ORFs (77.18%) being within tenfold range of the most abundant ORF (Figure 32B). This strengthens the assumption that most of the skewing in the previous libraries is caused by the pooled gateway reaction.

3.20 Converting a pooled lentiviral ORF library consisting of transcription factors to the Phenosudoku format

The library of expression vectors was converted into the Phenosudoku format the resulting library was quality controlled. Therefore, 24 wells were analyzed using next generation sequencing to determine the quantitative composition of their plasmid pools. The estimated number of colony-forming units (CFU) per well was around 30 (Table 77).

Dilution factor	Plated volume [µl]	colonies	CFU/ml (stock)	CFU/well
650000	1,200	26	1.69E+07	26
130000	130,000	146	1.90E+07	29.2
1300000	1,300,000	18	2.34E+07	36

Table 77 | Estimation of number of ORFs/well by estimating CFU

However, the sequencing data showed an average of 21.6 constructs per well, with a minimum of 15 and a maximum of 34 constructs per well (Figure 33). This discrepancy indicates that some CFUs might have contained the same plasmid and lead to higher abundances of those ‘duplicated’ plasmids. The relatively broad range of the number of constructs per well also supports this hypothesis.

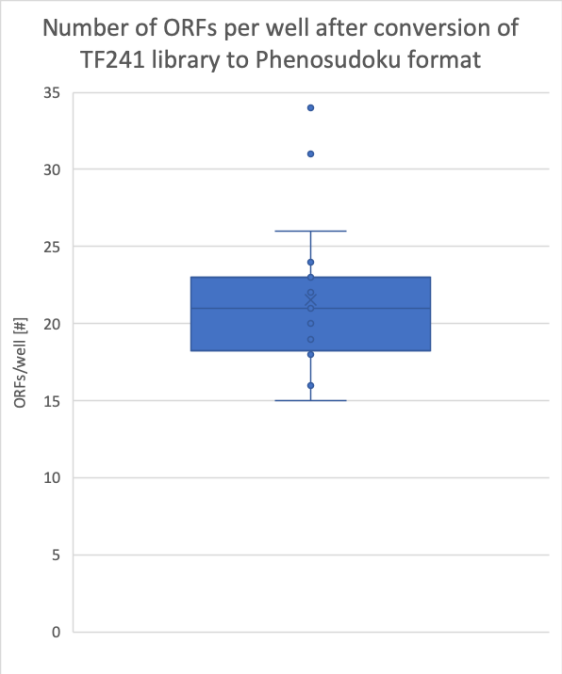


Figure 33 | Number of ORF constructs per well after conversion of TF241 library to Phenosudoku format. 24 wells were analyzed.

Moreover, the differences in abundance between constructs was still up to hundredfold in many wells, although overall the distribution of abundances was more uniform than for previous attempts (Figure 34).

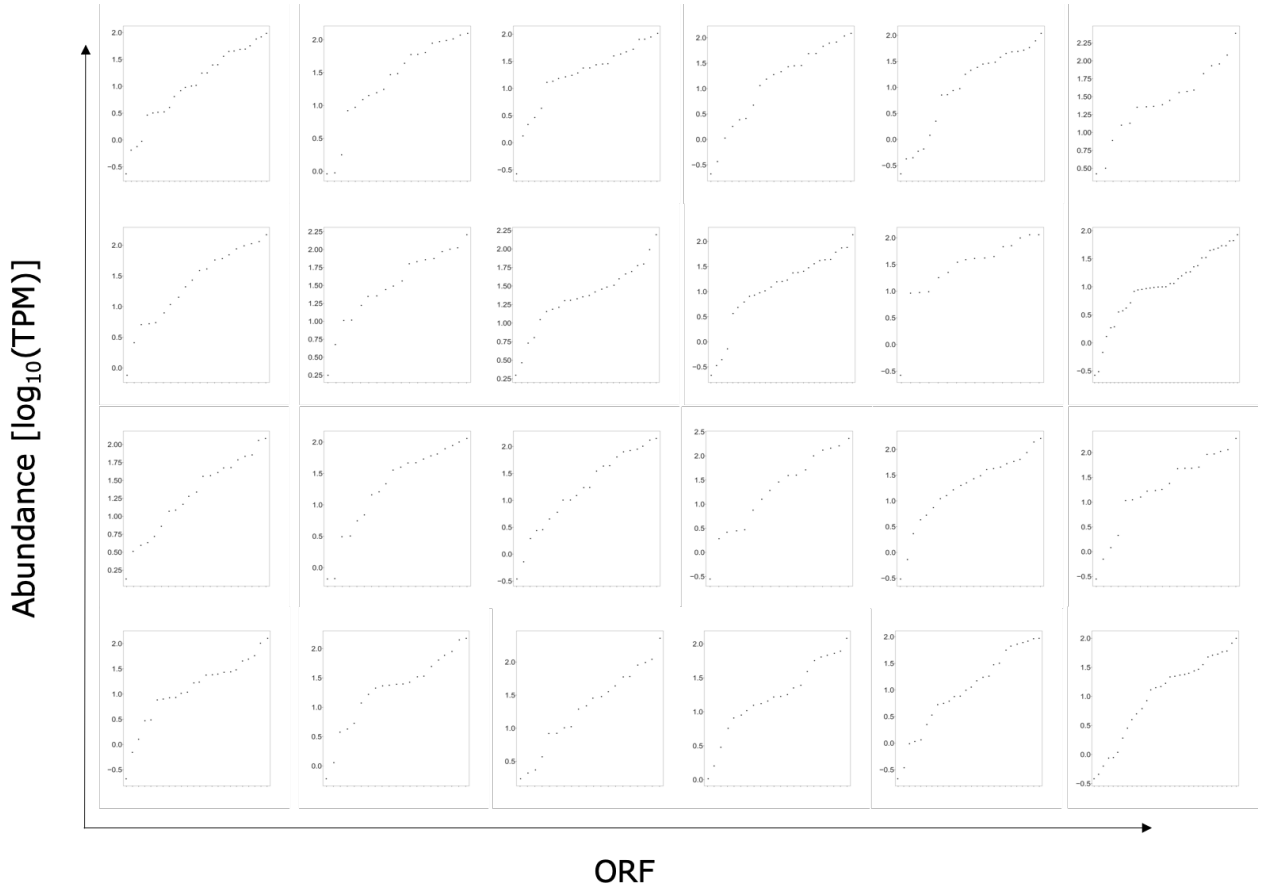


Figure 34 | Quantitative composition of the 24 analyzed wells of the TF241 library in the Phenosudoku format. Only ORFs with TPM > 0.2 are shown.

In one well, the distribution of abundances was approximately within a tenfold range (Figure 34, 2nd row, 5th column). There were 40 unique ORFs enriched in the top 2 most abundant constructs per well, with 7 of them occurring twice and one of them occurring three times (Table 78).

ORF	Size [bp]	Expr. Pool [TPM]	Expr. Pool [rank]	Occurrence in top 2
SOX10	1401	13.881718	1	1
HIST1H2AI//HIST1H2AK//HIST1H2AL//HIST1H2AM//HIST1H2AG	393	9.763073	4	1
TCF21	540	8.496584	6	2
ZIM3	1416	8.321138	7	1
MESP2	1179	8.150535	8	1
TWIST2	483	7.906428	9	2

FOXN4	1551	7.805132	10	3
SMAD3	1278	7.153199	13	1
ZNF169	1770	7.144511	15	1
PDE7B	1353	6.966877	16	1
MTHFD2	1035	6.456519	21	1
ZSCAN4	1302	6.025779	25	1
TAL2	324	5.966787	27	1
SOX13	1869	5.879026	28	1
ZNF404	1656	5.397805	32	2
ZNF41	2340	5.205366	35	1
NR0B2	774	4.78412	45	1
KCNIP3	771	4.731308	47	2
ZNF35	1581	4.707657	48	1
CCDC106	843	4.690899	50	1
SP7	1296	4.61462	51	2
HOXB6	672	4.572127	53	1
HOXA6	702	4.562906	55	2
MED30	537	4.535296	58	1
HLF	885	4.322464	62	2
CRX	900	4.273255	64	1
TAF9	792	4.239505	65	1
ASCL4	516	4.155972	67	1
NCBP2	471	4.090581	69	1
BCL6B	1443	4.071613	70	1
TAF12	486	3.75292	83	1
YEATS4	684	3.638715	91	1
TFAP2A	1296	3.408324	105	1
FHL5	855	3.4003	108	1
ZNF79	1494	3.366496	109	1
ZNF547	1209	3.308897	114	1
GSC	771	3.043687	132	1
ZNF329	1626	2.988561	138	1
RBL1	3045	1.576157	182	1

Table 78 | Summary of data on ORFs which are among the top 2 most abundant in the Phenosudoku wells.

This indicates that there was no clear bias for any of the constructs to reach more wells than others, which is in agreement with the relatively uniform distribution of abundances of the initial expression vector library. The average size of the enriched ORFs was 1116.38 base pairs, which is lower than the average size of the library (1428.69 bp). This could suggest that smaller ORFs are more likely to be enriched. The average rank of the enriched ORFs in the expression vector

library was 54.44, which is in agreement with previous results that more abundant ORFs are more likely to be enriched. Together, these results show that performing the gateway reactions individually can mitigate skewing of the library, although it significantly limits the size of the library. Furthermore, while the resulting library in the Phenosudoku format still was not ideal due to big differences in the abundance of constructs in some wells, its parameters (number of constructs per well, abundance spread) were improved compared to the previous Phenosudoku library obtained by a pooled gateway reaction.

4 Discussion

The static concept of cell states and types does not reflect the dynamic and plastic changes in gene expression programs that govern cell state or type transitions. Methods to understand the molecular mechanisms underlying those changes in a systematic and unbiased manner are lacking. Furthermore, there is a clear need for approaches that allow delineating subtle and reversible transitions between states such as between the cancer cell state and the persister cell state[30].

This research project aimed to address these knowledge gaps by laying the foundation for a novel tool to understand transitions between cell states in an unbiased manner. This tool combines a novel genetic screening paradigm (Phenosudoku), an ORF-based lentiviral library and a targeted bulk RNA-Sequencing approach. As a proof of concept, the cell state transition between human BJ fibroblasts and iPSCs was tackled, due to the strong differences between the transcriptomes of the two states and the opportunity to discover novel factors in the reprogramming process which is still far from being completely understood.

To develop an appropriate RNA-Seq approach to use as a readout for the screening tool, I designed two versions of a pool of targeted primers which were integrated into a modified CEL-Seq2 protocol. One modification to the protocol was that aRNA was sheared by sonication instead of chemical shearing. This led to more reproducible results and fragment size distributions with sharper peaks. Since CEL-Seq2 is essentially a single-cell RNA-Seq method, reagents such as the type of reverse transcriptase and their quantities were altered to suit the needs for bulk RNA-Seq. To benchmark the sequencing approach, I spiked in various amounts of hESCs into fibroblasts and generated libraries from the extracted RNA.

The first version of targeted primers did not enrich for targeted transcripts very well, probably also due to non-ideal choice of targeted genes and targeted section of genes. Those primers were designed using publicly available data obtained using whole-transcript methods, while the applied modified CEL-Seq2 protocol generated libraries from 3’ ends of mRNAs. The poor performance of the first version of targeted primers was also apparent by the relatively low quality of the generated library and the lacking resolution between spike-in conditions. The use of a reverse transcriptase with a reaction temperature at 42°C in combination with the targeted primers v1 which had annealing temperatures around 55°C might have also played a role.

The second version of targeted primers were designed using 3’ mRNA Sequencing data generated from samples containing only hESCs or only fibroblasts using the modified CEL-Seq2 approach. Moreover, the design was influenced by information on coverage across the gene body, to

exclude genes which could not be captured at the 3' poly(A) tail. The improved design and use of a reverse transcriptase with an incubation temperature of 55°C lead to increased enrichment of target genes in the resulting sequencing libraries. The overall quality of the libraries was also better and compared to an untargeted sequencing approach, the use of targeted primers slightly improved the separation of spike-in samples.

The targeted sequencing method could be further improved by extending the targeted gene set and by further optimizing of the conditions for reverse transcription. An alternative approach could be to increase specificity by amplifying the targeted transcripts in a two-step PCR reaction, with the first PCR being specific to the transcripts of interest, and the second one being specific to the products of the first PCR. However, such an alternative approach would require significant changes to the current protocol. A compromise could be to introduce more specificity by using primers specific to targeted transcript products from the reverse transcription step to amplify the library in the final PCR reaction of the modified CEL-Seq2 protocol.

The DASH approach[70] was leveraged to deplete irrelevant fibroblast sequences from the sequencing libraries and thus increase the coverage of relevant targets. Surprisingly, the use of an sgRNA library consisting of 1000 sgRNAs lead to significant depletion of unwanted reads, despite the fact that DASH has only been shown to work with few sgRNAs[70]. Furthermore, it is normally used in combination with full-length RNA-Sequencing protocols, which is why the compatibility with a 3' mRNA library preparation technique shown in this thesis is particularly encouraging. However, when comparing DASHed samples to not DASHed samples, the separation between samples with different numbers of hESCs spiked in does not improve. This could indicate that the necessary sequencing depth to separate those samples is already reached. These results suggest that the size of the sgRNA library could be further increased and establish DASH as a useful tool to cut sequencing costs by increasing sequencing depth of genes of interest. This could be especially valuable in combination with the Phenosudoku screening paradigm, since most wells will not contain a phenotype (in this case an iPSC). Therefore, DASHing out abundant fibroblast genes will reduce unwanted reads from each well, thus increasing the positive effect of DASH in an additive manner.

As an alternative approach, the Finding Lowly Abundant Sequences using Hybridization (FLASH) method[76] could be used to enrich for transcripts of interest. 836 sgRNAs directed against 164 genes that are specific to the hESC state were designed. However, the approach was never experimentally validated.

To prove the concept of the developed tool to decipher cell state transitions, fibroblasts were transduced with a vector to ectopically overexpress Sox2, Klf4 and c-Myc. A separate vector was used to overexpress the constituents of the ORF library. In a pilot experiment, Oct4 was cloned into the library vector, fibroblasts were transduced with both vectors and reprogrammed in wells of a 96-well plate and a 6-well plate. While the morphological changes were in agreement with expected results, most cells in the 96 well died and no iPSC colonies were observed. This could indicate suboptimal concentration of the used lentivirus. Alternatively, the reprogramming protocol might need further optimization, especially since the cell number and space in the well of a 96-well plate might be too small. It could also be that the used promoter – the short EF1a promoter – might not be ideal in terms of strength or ability to be silenced. Furthermore, the

separation of the Yamanaka factors onto two different vectors could have also impaired the reprogramming outcome.

ORFs were used as a means of perturbation for the tool. There were many difficulties in cloning a pooled library of ORFs, mainly due to their large and variable size. Initial attempts to develop a protocol to clone multiple ORFs into a destination vector in a pooled Gateway cloning reaction showed strong skewing of abundance of constituents of the resulting libraries. This skewing is probably caused by preference of the LR recombinase for both certain distance between the L1 and L2 recombination sites and for sequence context surrounding the recombination sites. Expectedly, the results indicate that the abundance of an ORF also seems to be an important factor in the pooled Gateway reaction. Potentially, stochastic effects also influence the reaction. Individual Gateway cloning of the ORFs prevented skewing. However, individual Gateway cloning is highly labor intensive and limits the size of the ORF libraries. That is why finding an alternative approach to individual Gateway cloning and pooled Gateway cloning – such as maybe Golden Gate cloning – should be identified in the future.

When converting the generated libraries to the Phenosudoku format, the skewing of abundances increased. As expected, abundant ORFs seemed to have had a higher chance of being allocated to more wells, thus inhibiting a random distribution of all ORFs among the wells when the abundances were too different. The pooled library obtained by individual Gateway cloning reactions was not distorted, and the resulting Phenodusudoku library seemed to be not skewed and ready for proof of concept screens. For the future, improved sequencing methods and pipelines to identify and quantify the ORF composition in the wells are needed.

The results of this thesis offer a foundation for a tool to decipher the causal factors behind cell state or type transitions in a systematic manner. The tool combines a novel genetic screening paradigm called Phenosudoku, targeted transcriptome analysis, and perturbagens in the form of a library consisting of human ORFs. Currently, no methods are available that allow delineating transitions of cell state or type in a systematic and unbiased manner while uncovering causality underlying the transition. Most factors or factor combinations that causally drive cell state or type transitions are discovered by small-scale trial and error endeavors. CRISPRa is a powerful genetic tool to screen gain-of-function phenotypes in a high-throughput manner[63,77,78]. However, the activation of genes using the CRISPRa system can be quite variable and relatively weak[79]. Furthermore, Cas9 binding has been shown to be affected by chromatin state[64], implicating that also dCas9 binding is affected.

While the concept of the method remains to be proven in a proof of concept experiment and certain parts of the tool could be further optimized, the method developed in this thesis project could overcome current limitations and offer a novel avenue to investigate and understand the causal factors behind cell state transitions in a systematic and unbiased manner.

5 Appendix

5.1 PCR primers

ID	alias	Sequence (5'-3')	bp	Usage	Vector
OMS004	VMS002_hOKMS_aP2A_rev	AAGGCTTGGCCATGGGTCCAGGGTTTTCTTCGACATCTCCAGCCTGCTTCAGCAGGCTGAAGTTAGTAGCtccagatcccatgtgtgagaggggcagtg	99	HiFi assembly	VMS002
OMS005	VMS002_aP2A_Blast_for	ctctcacacatgggatctggaGCTACTAACTTCAGCCTGCTGAAGCAGGCTGGAGATGTCGAAGAAAACCCTGGACCCATGGCCAAGCCTTTGTCTCA	98	HiFi assembly	VMS002
OMS006	VMS002_Blast_rev	tacctagtggaaccggaacccttaaattaGCCCTCCCACACATAACCAGAG	51	HiFi assembly	VMS002
OMS007	VMS002_hOKMS_for	GTTTGCCGCCAGAACAgtgagCTAGgccaccatggcgggacacctgg	47	HiFi assembly	VMS002
OMS019	VMS004_5_OKM_rev	AAAGGCTTGGCCATgggaccggggttactttca	33	HiFi assembly	VMS005
OMS020	VMS004_5_blast_for	taaccccggtcccATGGCCAAGCCTTTGTCTCAAG	35	HiFi assembly	VMS005
OMS021	VMS004_5_6_7_blast_rev	gaacccttactgccatcggcatgcatttaGCCCTCCCACACATAACCAGAG	51	HiFi assembly	VMS005
OMS022	VMS005_sox_for	gggtttgccgccagaacacagggttctagagccaccatgtacaacatgatggagacggagc	61	HiFi assembly	VMS005
OMS023	VMS005_Sox_rev	cctctccagatcccatgtgtgagaggggcagtg	33	HiFi assembly	VMS005
OMS024	VMS005_KM_for	ctctcacacatgggatctggagagggcagagga	33	HiFi assembly	VMS005
OMS029	VMS008_puro_f	tgaaaaacacgatgataaGGccGCCACCATGACCGAGTACAAG	43	HiFi assembly	VMS008_no_T7
OMS030	VMS008_puro_r	gaaccggaacccttaaaAGGttaGGCACCGGGCTTGCGGGT	41	HiFi assembly	VMS008_no_T7

OMS031	VMS008_T7_infusion_for	TAATACGACTCACTATCTAGCGctagcAtcACAAGTTTGT	40	T7 infusion	VMS008
OMS032	VMS008_T7_infusion_rev	CTATAGTGAGTCGTATTActcactgttctggcggcaaa	38	T7 infusion	VMS008
OMS033	VMS009_cmr_for	ACACAACATATCCAGTCACTATGGCGGCCGCACACAACATATCCAGTCACTATGGC	56	HiFi assembly	VMS009
OMS034	VMS009_cmr_rev	CCGGTTAGCGCTAGCTCATTACTAAACCACTTTGTACAAGAAAGCTGAA	49	HiFi assembly	VMS009
OMS035	VMS009_ires_for	TTAGTAATGAGCTAGCGCTAACCGGTGATCTAGAGGGCCAAAAAAAAAAAAAAAAAAAAA	60	HiFi assembly	VMS009
OMS036	VMS009_ires_rev	CGTTTTTTAACCTCGACTAAACACATGT	28	HiFi assembly	VMS009
OMS037	VMS009_Stop_for	ATGAGATCTAGAGGGCCAAAAAAAAAAAAAAAAAAAAAA	41	mutagenesis_pcr	VMS009
OMS038	VMS009_stop_rev	TACTAAACCACTTTGTACAAGAAAGCTGAAC	31	mutagenesis_pcr	VMS009
OMS039	VMS010_KLD_for	tcggtctcgattctacgtagtaatgaGATCTAGAGGGCCaaaaaaaaaaaaaaaaaaaa	60	KLD reaction	VMS010
OMS040	VMS010_KLD_rev	ggagagggttagggataggcttaccAACCACTTTGTACAAGAAAGCTGAACG	52	KLD reaction	VMS010
OMS045	pENTR11_linearize_for	CCAAACCCAGCTTTCTTGTACAAAGTTGG	29	linearize pentr11	VMS011, 012, 013
OMS046	pENTR11_linearize_rev	AGCCTGCTTTTTTGTACAAAGTTGG	25	linearize pentr11	VMS011, 012, 013
OMS047	Oct4_infusion_kozak_for	ACAAAAAAGCAGGCTgccaccatggcgggac	31	infusion	VMS012, 013

OMS048	Oct4_infusion_nostop_rev	ACAAGAAAGCTGGGTttgggtttgaatgcatggagagccc	41	infusion	VMS012
OMS049	Oct4_infusion_stop_rev	ACAAGAAAGCTGGGTttggtcagtttgaatgcatgggagagcc	43	infusion	VMS013
OMS050	GFP_infusion_for	ACAAAAAAGCAGGCTatggtgagcaagggcgag	33	infusion	VMS011
OMS051	GFP_infusion_nostop_rev	ACAAGAAAGCTGGGTttggcttgtacagctcgtccatgcc	40	infusion	VMS011
OMS054	Sox2_for_VMS015	gccagaacacagggttctagagccacca	28	HiFi assembly	VMS015
OMS055	E2A_rev_VMS015	ctcaccatgggaccggggttac	22	HiFi assembly	VMS015
OMS056	EGFP_for_VMS015	cggtcccatggtgagcaagggcg	23	HiFi assembly	VMS015
OMS057	EGFP_rev_VMS015	tgccatcggcatgcattacttgtacagct	29	HiFi assembly	VMS015
OMS058	VMS016_017_IRES_for	cggtgcacatgctttacatgtgtttagt	28	HiFi assembly	VMS017
OMS059	VMS016_017_WPRE_rev	tcattggtcttaaaggtacctgaggggtgtgactgga	37	HiFi assembly	VMS017

5.2 Vectors

5.2.1 Adopted vectors

Identifier	Description	Source	Catalog # or ID #
FUW-tetO-hOKMS	FUW-tetO-hOKMS	addgene	51543
MTM_672	pSicoR-sEF1a-mCherry	McManus lab	672
MTM_1073	pROSA26short_(sp)dCas9-VPR-t2a-blast	McManus lab	1073
MTM_277	pSicoR-Blasti-T2A-EGFP	McManus lab	277
V18034	pLN-sEF1a-SpCas9-mTagBFP2-Blast	Neil Tay	NA

V18033	pLN-EF1a-EGFP	Neil Tay	NA
pBID-Dest	pBID-attR1-Cmr-ccdb-attR2-pre-GFP-Puro	Anton Ogorodnikov	NA
pSuperInf-IRES	pSuperInf-attR1-Cmr-ccdb-attR2-IRES	Anton Ogorodnikov	NA
pENTR11	pENTR11-attL1-Cmr-ccdb-attL2	Thermo Fisher Scientific	A10467
FUW-tetO-hOCT4	FUW-tetO-hOCT4	addgene	20726
pSuperInf-IRES-mCherry	pSuperInf-IRES-mCherry	Anton Ogorodnikov	NA
pLEX307	pLEX307	addgene	41392

5.2.2 Cloned Vectors

Identifier	Description	Source
VMS002	pSicoR-sEF1a-O-K-M-S-aP2A-Blast	Moritz Schlapansky
VMS005	pLN-sEF1a-SKM-Blast	Moritz Schlapansky
VMS008	pL-sEF1a-T7-attR-IRES_Puro	Moritz Schlapansky
VMS009	pL-sEF1a-T7-attR-STOP-IRES_Puro	Moritz Schlapansky
VMS010	pL-sEF1a-T7-attR-V5-IRES_Puro	Moritz Schlapansky
VMS011	pENTR11-attL1-EGFP_nostop-attL2	Moritz Schlapansky
VMS012	pENTR11-attL1-Oct4_nostop-attL2	Moritz Schlapansky
VMS013	pENTR11-attL1-Oct4_nativestop-attL2	Moritz Schlapansky
VMS015	pLN-sEF1a-SKM-EGFP	Moritz Schlapansky
VMS017	pL-sEF1a-T7-attR-STOP-IRES-mCherry	Moritz Schlapansky
VMS026	pL-sEF1a-T7-GFP_nostop-STOP-IRES_Puro	Moritz Schlapansky
VMS027	pL-sEF1a-T7-GFP_nostop-V5-IRES_Puro	Moritz Schlapansky
VMS028	pL-sEF1a-T7-Oct4_nativestop-STOP-IRES_Puro	Moritz Schlapansky
VMS029	pL-sEF1a-T7-Oct4_nativestop-V5-IRES_Puro	Moritz Schlapansky
VMS030	pL-sEF1a-T7-Oct4_nostop-STOP-IRES_Puro	Moritz Schlapansky
VMS031	pL-sEF1a-T7-Oct4_nostop-V5-IRES_Puro	Moritz Schlapansky

References

1. Hooke, R. *Micrographia: or, some physiological descriptions of minute bodies made by magnifying glasses*. *Micrographia* (1665).

2. Mazzarello, P. A unifying concept: The history of cell theory. *Nat. Cell Biol.* (1999). doi:10.1038/8964

3. Murray, J. M. *et al.* Linkage relationship of a cloned DNA sequence on the short arm of the X chromosome to Duchenne muscular dystrophy. *Nature* (1982). doi:10.1038/300069a0

4. Clunes, M. T. & Boucher, R. C. Cystic fibrosis: the mechanisms of pathogenesis of an inherited lung disorder. *Drug Discovery Today: Disease Mechanisms* (2007). doi:10.1016/j.ddmec.2007.09.001

5. Noguchi, M. *et al.* Interleukin-2 receptor γ chain mutation results in X-linked severe combined immunodeficiency in humans. *Cell* (1993). doi:10.1016/0092-8674(93)90167-O

6. Penney, D. P. A brief history of the biological stain commission: Its founders, its mission and the first 75 years. *Biotech. Histochem.* (2000). doi:10.3109/10520290009066496

7. Arendt, D. *et al.* The origin and evolution of cell types. *Nature Reviews Genetics* (2016). doi:10.1038/nrg.2016.127

8. Köhler, G. & Milstein, C. Continuous cultures of fused cells secreting antibody of predefined specificity. *Nature* (1975). doi:10.1038/256495a0

9. Coons, A. H., Creech, H. J. & Jones, R. N. Immunological Properties of an Antibody Containing a Fluorescent Group. *Proc. Soc. Exp. Biol. Med.* (1941). doi:10.3181/00379727-47-13084P

10. Picot, J., Guerin, C. L., Le Van Kim, C. & Boulanger, C. M. Flow cytometry: Retrospective, fundamentals and recent instrumentation. *Cytotechnology* (2012). doi:10.1007/s10616-011-9415-0

11. Langer-Safer, P. R., Levine, M. & Ward, D. C. Immunological methods for mapping genes on Drosophila polytene chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* (1982). doi:10.1073/pnas.79.14.4381

12. Chao, M. P., Seita, J. & Weissman, I. L. Establishment of a normal hematopoietic and leukemia stem cell hierarchy. in *Cold Spring Harbor Symposia on Quantitative Biology* (2008). doi:10.1101/sqb.2008.73.031

13. Kim, C. C. & Lanier, L. L. Beyond the transcriptome: completion of act one of the Immunological Genome Project. *Current Opinion in Immunology* (2013). doi:10.1016/j.coi.2013.09.013

14. Uhlén, M. *et al.* A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics* (2005). doi:10.1074/mcp.M500279-MCP200

15. Regev, A. *et al.* The human cell atlas. *Elife* (2017). doi:10.7554/eLife.27041

16. Engel, P. *et al.* CD Nomenclature 2015: Human Leukocyte Differentiation Antigen Workshops as a Driving Force in Immunology. *J. Immunol.* (2015). doi:10.4049/jimmunol.1502033

17. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* (2019). doi:10.1038/s41467-019-10802-z

18. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* (2019). doi:10.7554/eLife.43803

19. Wynn, T. A., Chawla, A. & Pollard, J. W. Macrophage biology in development, homeostasis and disease. *Nature* (2013). doi:10.1038/nature12034

20. Rebhahn, J. A. *et al.* An animated landscape representation of CD4+ T-cell differentiation, variability, and plasticity: Insights into the behavior of populations versus cells. *European Journal of Immunology* (2014). doi:10.1002/eji.201444645

21. Waddington, C. H. The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. *Strateg. genes A Discuss. some …* (1957). doi:10.1007/3-540-32786-X_7

22. Chen, S. *et al.* Computing free energy landscapes: Application to Ni-based electrocatalysts with pendant amines for H2 production and oxidation. *ACS Catal.* (2014). doi:10.1021/cs401104w

23. Wales, D. J. & Bogdan, T. V. Potential energy and free energy landscapes. *J. Phys. Chem. B* (2006). doi:10.1021/jp0680544

24. Mallamace, F. *et al.* Energy landscape in protein folding and unfolding. *Proc. Natl. Acad. Sci. U. S. A.* (2016). doi:10.1073/pnas.1524864113

25. D'Alessio, A. C. *et al.* A systematic approach to identify candidate transcription factors that control cell identity. *Stem Cell Reports* (2015). doi:10.1016/j.stemcr.2015.09.016

26. Ostuni, R. & Natoli, G. Lineages, cell types and functional states: A genomic view. *Current Opinion in Cell Biology* (2013). doi:10.1016/j.ceb.2013.07.006

27. Wu, F., Su, R. Q., Lai, Y. C. & Wang, X. Engineering of a synthetic quadrastable gene network to approach Waddington landscape and cell fate determination. *Elife* (2017). doi:10.7554/eLife.23702

28. Kauffman, S. A. *The Origins of Order: Self-organization and Selection in Evolution*. (Oxford University Press, 1993).

29. Shaffer, S. M. *et al.* Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* (2017). doi:10.1038/nature22794

30. Sharma, S. V. *et al.* A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. *Cell* **141**, 69–80 (2010).

31. Ramirez, M. *et al.* Diverse drug-resistance mechanisms can emerge from drug-tolerant cancer persister cells. *Nat. Commun.* **7**, 10690 (2016).

32. Davis, R. L., Weintraub, H. & Lassar, A. B. Expression of a single transfected cDNA converts fibroblasts to myoblasts. *Cell* (1987). doi:10.1016/0092-8674(87)90585-X

33. Ieda, M. *et al.* Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. *Cell* **142**, 375–386 (2010).

34. Xiao, D. *et al.* Direct reprogramming of fibroblasts into neural stem cells by single non-neural progenitor transcription factor Ptf1a. *Nat. Commun.* **9**, 2865 (2018).

35. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* (2010). doi:10.1038/nature08797

36. Du, Y. *et al.* Human hepatocytes with drug metabolic function induced from fibroblasts by lineage reprogramming. *Cell Stem Cell* (2014). doi:10.1016/j.stem.2014.01.008

37. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* (2006). doi:10.1016/j.cell.2006.07.024

38. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861–72 (2007).

39. Shi, Y., Inoue, H., Wu, J. C. & Yamanaka, S. Induced pluripotent stem cell technology: a decade of progress. *Nat. Rev. Drug Discov.* **16**, 115–130 (2017).

40. Takeda, Y., Harada, Y., Yoshikawa, T. & Dai, P. Chemical compound-based direct reprogramming for future clinical applications. *Bioscience Reports* (2018). doi:10.1042/BSR20171650

41. Biswas, D. & Jiang, P. Chemically Induced Reprogramming of Somatic Cells to Pluripotent Stem Cells and Neural Cells. *Int. J. Mol. Sci.* **17**, 226 (2016).

42. Zhang, M. *et al.* Pharmacological Reprogramming of Fibroblasts into Neural Stem Cells by Signaling-Directed Transcriptional Activation. *Cell Stem Cell* **18**, 653–667 (2016).

43. Hou, P. *et al.* Pluripotent stem cells induced from mouse somatic cells by small-molecule compounds. *Science (80-.).* (2013). doi:10.1126/science.1239278

44. Zheng, J. *et al.* A combination of small molecules directly reprograms mouse fibroblasts into neural stem cells. *Biochem. Biophys. Res. Commun.* (2016). doi:10.1016/j.bbrc.2016.05.080

45. Hu, W. *et al.* Direct Conversion of Normal and Alzheimer's Disease Human Fibroblasts into Neuronal Cells by Small Molecules. *Cell Stem Cell* (2015). doi:10.1016/j.stem.2015.07.006

46. Sayed, N., Liu, C. & Wu, J. C. Translation of Human-Induced Pluripotent Stem Cells from Clinical Trial in a Dish to Precision Medicine. *Journal of the American College of Cardiology* (2016). doi:10.1016/j.jacc.2016.01.083

47. Farid, S. S. & Jenkins, M. J. Bioprocesses for Cell Therapies. in *Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes* (2018). doi:10.1016/B978-0-08-100623-8.00044-X

48. Hwang, B., Lee, J. H. & Bang, D. Single-cell RNA sequencing technologies and bioinformatics

pipelines. *Experimental and Molecular Medicine* (2018). doi:10.1038/s12276-018-0071-8

49. Simoni, Y., Chng, M. H. Y., Li, S., Fehlings, M. & Newell, E. W. Mass cytometry: a powerful tool for dissecting the immune landscape. *Current Opinion in Immunology* (2018). doi:10.1016/j.coi.2018.03.023

50. Gao, R. *et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.* (2016). doi:10.1038/ng.3641

51. Kelsey, G., Stegle, O. & Reik, W. Single-cell epigenomics: Recording the past and predicting the future. *Science* (2017). doi:10.1126/science.aan6826

52. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Research* (2015). doi:10.1101/gr.190595.115

53. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol. Cell* (2017). doi:10.1016/j.molcel.2017.01.023

54. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols* (2018). doi:10.1038/nprot.2017.149

55. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* (2017). doi:10.1186/s13073-017-0467-4

56. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* (2019). doi:10.1038/s41586-019-0969-x

57. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* (2014). doi:10.1038/nmeth.2967

58. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics* (2019). doi:10.3389/fgene.2019.00317

59. Cahan, P. *et al.* CellNet: Network biology applied to stem cell engineering. *Cell* (2014). doi:10.1016/j.cell.2014.07.020

60. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).* (2012). doi:10.1126/science.1225829

61. Qi LS *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* (2013).

62. Qin, H. *et al.* Systematic Identification of Barriers to Human iPSC Generation. *Cell* **158**, 449–461 (2014).

63. Liu, Y. *et al.* CRISPR Activation Screens Systematically Identify Factors that Drive Neuronal Fate and Reprogramming. *Cell Stem Cell* (2018). doi:10.1016/j.stem.2018.09.003

64. Verkuijl, S. A. & Rots, M. G. The influence of eukaryotic chromatin state on CRISPR–Cas9 editing efficiencies. *Current Opinion in Biotechnology* (2019).

doi:10.1016/j.copbio.2018.07.005

65. Chavez, A. *et al.* Comparative Analysis of Cas9 Activators Across Multiple Species HHS Public Access. *Nat Methods* (2016). doi:10.1038/nmeth.3871

66. Wiemann, S. *et al.* The ORFeome Collaboration: A genome-scale human ORF-clone resource. *Nature Methods* (2016). doi:10.1038/nmeth.3776

67. Prelich, G. Gene overexpression: Uses, mechanisms, and interpretation. *Genetics* (2012). doi:10.1534/genetics.111.136911

68. Sittampalam, G. *et al. Assay Guidance Manual*. *Assay Guidance Manual* (2016). doi:PMID:22553881

69. Hashimshony, T. *et al.* CEL-Seq2: Sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* (2016). doi:10.1186/s13059-016-0938-8

70. Gu, W. *et al.* Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* **17**, 41 (2016).

71. Johannesson, B. *et al.* Comparable frequencies of coding mutations and loss of imprinting in human pluripotent cells derived by nuclear transfer and defined factors. *Cell Stem Cell* **15**, 634–42 (2014).

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

73. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* (2002). doi:10.1093/nar/30.1.207

74. Orozco-Fuentes, S. *et al.* Quantification of the morphological characteristics of hESC colonies. *Sci. Rep.* **9**, 17569 (2019).

75. Courtot, A. M. *et al.* Morphological analysis of human induced pluripotent stem cells during induced differentiation and reverse programming. *Biores. Open Access* (2014). doi:10.1089/biores.2014.0028

76. Quan, J. *et al.* FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz418

77. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–661 (2014).

78. Chavez, A. *et al.* Highly efficient Cas9-mediated transcriptional programming. *Nat. Methods* (2015). doi:10.1038/nmeth.3312

79. Chavez, A. *et al.* Comparison of Cas9 activators in multiple species. *Nat. Methods* **13**, 563–567 (2016).

I have made an effort to appropriately reference all figures in this thesis. However, should a copyright infringement have occurred despite all precautions, I would ask to be notified.

Table of figures

Table of tables