# TECHNISCHE UNIVERSITÄT MÜNCHEN

Fakultät für Medizin

# Development of a Novel Target Enrichment and Barcoding Method for Next-Generation Sequencing, and Implementation for Single-Cell Transcriptomics

Fatma Uzbaş

Diese Dissertation wurde an der
Fakultät für Medizin der Technischen Universität München
zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften**

vorgelegt.

**TECHNISCHE UNIVERSITÄT MÜNCHEN**

FAKULTÄT FÜR MEDIZIN

# Development of a Novel Target Enrichment and Barcoding Method for Next-Generation Sequencing, and Implementation for Single-Cell Transcriptomics

Fatma Uzbaş

Vollständiger Abdruck der von der
Fakultät für Medizin der Technischen Universität München
zur Erlangung des akademischen Grades eines

**Doktors der Naturwissenschaften**

genehmigten Dissertation.

Vorsitzender: Prof. Dr. Radu Roland Rad

Prüfer der Dissertation:

1. Prof. Dr. Heiko Lickert
2. Prof. Dr. Wolfgang Wurst

Die Dissertation wurde am 11.11.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Medizin am 16.06.2020 angenommen.

# Abstract

The past ten years witnessed the exponential evolution of the single-cell sequencing technologies, nowadays reaching to analysis of over a million cells in a single experiment. They provide unprecedented insights into developmental trajectories in complex organisms, significant rare cells such as adult stem cells or circulating tumor cells, cancer microevolution, and environmental studies. Most popular single-cell analysis techniques have certain limitations though, such as shallow coverage, lack of quantitativeness for many genes, and high costs. In order to address these problems, I developed a method for enrichment of selected genomic/transcriptomic loci and barcoding for next-generation sequencing, named Barcode Assembly for Targeted Sequencing (BART-Seq).

In this study, I initially optimized a workflow and implemented it for targeted transcriptomics. After verifying dynamic range measurements of bulk samples, I used the method for analyzing transcripts in single cells. I explored the expression of selected pluripotency genes in self-renewing human embryonic stem cells (hESCs) and observed that they embrace different flavors of pluripotency depending on the maintenance media composition. Next, I analyzed the cell subpopulations that emerge from hESCs upon activation of the Wnt/ß-catenin pathway at different levels, and observed that they correspond to distinct regions of the gastrulating embryo based on the inducer. Moreover, I have contributed to two projects for targeted genotyping and compound screening of bulk gDNA/RNA samples. In parallel, I developed bioinformatics tools for analyzing the BART-Seq data; from raw count matrices to biological interpretations.

BART-Seq is the first targeted sequencing technology that is applicable for both transcriptomics of single cells and genomics/transcriptomics of bulk samples. It addresses drawbacks of existing methods by offering increased sequencing depth, quantitative measurements, and ability to analyze also the non-poly(A) transcripts. The simple and cost-effective workflow that can be performed with basic laboratory equipment and open-access bioinformatic tools makes BART-Seq accessible to any research group. I therefore expect that it will serve as an important companion to existing technologies for a wide spectrum of research fields.

This project was carried out in the Institute of Stem Cell Research, Helmholtz Center Munich - Germany, under the supervision of Dr. Micha Drukker and Prof. Dr. Heiko Lickert.

Key parts of my thesis were published in the following peer-reviewed article:

Uzbas F, Opperer F, Sönmezer C, Shaposhnikov D, Sass S, Krendl C, Angerer P, Theis FJ, Müller NS, Drukker M (2019) **BART-Seq: cost-effective massively parallelized targeted sequencing for genomics, transcriptomics, and single-cell analysis.** *Genome Biology* 20 (1):155. https://doi.org/10.1186/s13059-019-1748-6

# Zusammenfassung

In den letzten zehn Jahren hat sich die Technologie der Einzelzellsequenzierung exponentiell weiterentwickelt, sodass mittlerweile über eine Million Zellen in einem einzigen Experiment analysiert werden können. Dies liefert beispiellose Einblicke in Entwicklungsverläufe in komplexen Organismen, signifikante seltene Zellen wie adulte Stammzellen oder zirkulierende Tumorzellen, Krebsmikroevolution und Umweltstudien. Die meisten gängigen Einzelzellanalysetechniken weisen jedoch bestimmte Einschränkungen auf, z. B. geringe Abdeckung, mangelnder quantitativer Sinn für viele Gene und hohe Kosten. Um diese Probleme anzugehen, habe ich eine Methode zur Anreicherung ausgewählter genomischer/transkriptomischer Loci und zum Barcoding für Next-Generation Sequenzierung mit dem Namen Barcode Assembly for Targeted Sequencing (BART-Seq) entwickelt.

In dieser Studie habe ich zunächst einen Workflow optimiert und diesen für gezielte Transkriptomanalysen implementiert. Nachdem ich die dynamischen Bereichsmessungen von Gesamtproben verifiziert hatte, verwendete ich die Methode zur Analyse von Transkripten in einzelnen Zellen. Ich untersuchte die Expression ausgewählter Pluripotenzgene in sich selbst erneuernden humanen embryonalen Stammzellen (hESCs) und stellte fest, dass sie je nach Zusammensetzung des Erhaltungsmediums unterschiedliche Arten der Pluripotenz aufweisen. Als nächstes analysierte ich die Zell-Subpopulationen, die bei Aktivierung des Wnt/ß-Catenin-Weges auf verschiedenen Ebenen aus hESCs hervorgehen, und beobachtete, dass sie je nach Induktor unterschiedlichen Regionen des gastrulierenden Embryos entsprechen. Darüber hinaus habe ich an zwei Projekten zur gezielten Genotypisierung und zum Screening von gDNA/RNA-Gesamtproben mitgewirkt. Parallel dazu entwickelte ich Bioinformatik-Tools zur Analyse der BART-Seq-Daten; ausgehend von den Rohdaten bis hin zu deren biologischer Interpretation.

BART-Seq ist die erste zielgerichtete Sequenzierungstechnologie, die sowohl für die Transkriptomik einzelner Zellen als auch für die Genomik/Transkriptomik von Gesamtproben anwendbar ist. Es behebt die Nachteile bestehender Methoden, indem es eine erhöhte Sequenzierungstiefe, quantitative Messungen und die Möglichkeit bietet, auch Nicht-Poly(A)-Transkripte zu analysieren. Der einfache und kostengünstige Workflow, der mit grundlegenden Laborgeräten und frei zugänglichen Bioinformatik-Tools durchgeführt werden kann, macht BART-Seq für jede Forschungsgruppe zugänglich. Ich gehe daher davon aus, dass es für ein breites Spektrum an Forschungsbereichen ein wichtiger Begleiter bereits bestehender Technologien sein wird.

Dieses Projekt wurde im Institut für Stammzellforschung des Helmholtz-Zentrums München unter der Leitung von Dr. Micha Drukker und Prof. Dr. Heiko Lickert durchgeführt.

Wichtige Teile dieser Dissertation wurden in dem folgenden Peer-Review-Artikel veröffentlicht:

Uzbas F, Opperer F, Sönmezer C, Shaposhnikov D, Sass S, Krendl C, Angerer P, Theis FJ, Müller NS, Drukker M (2019) **BART-Seq: cost-effective massively parallelized targeted sequencing for genomics, transcriptomics, and single-cell analysis.** *Genome Biology* 20 (1):155. https://doi.org/10.1186/s13059-019-1748-6

*"Investigating the cascades that give rise to a whole organism commencing with fertilization is like witnessing the very beginning of a miracle, an alchemy combining an egg and a sperm to create an exquisite organism via a process similar to the formation of the Universe after the Big Bang; both start with a multi-potential but invariable structure and end in a vast multiplicity."*

*Fatma Uzbaş*

# Acknowledgements

*I am deeply grateful to DAAD for supporting my Ph.D. study, in particular to Laura Mendelssohn for her excellent support throughout the scholarship period; and to the HELENA Graduate School for the opportunities they provided for my professional and personal development as a researcher.*

*Thank you, Dr. Micha Drukker, for giving me the chance to carry out this project in your research group. For being not only my Ph.D. supervisor, but also a life coach and a friend at times.*

*My thesis committee members, Prof. Dr. Heiko Lickert and Prof. Dr. Wolfgang Wurst, I greatly appreciate your valuable feedback and advice.*

*Thank you, Dr. Florian Opperer, for teaching me all the basics of this project and offering your generous guidance whenever I needed; Can Sönmezer, for initiating this work; and, Dr. Dmitry Shaposhnikov, for sharing your knowledge and for the nice and insightful discussions.*

*All the other members of ISF-P; Dr. Friederike Matheus, Chaido Ori, Markus Grosch, Dr. Anna Pertek, Sebastian Ittermann, Ejona Rusha, Polyxeni Nteli, Dr. Miha Modic, and Dr. Christian Krendl, it was a great pleasure to work in the same group with you. Thank you, Karen Biniossek, for your assistance in the administrative matters. And, Valentyna Rishko, I am very glad that our friendship extended beyond the borders of the lab.*

*Bacolar, my family in Munich, the stressful Ph.D. life would not be tolerable without the getaways, lengthy chats, and the fun we had together…*

*Dr. Zeynep Altıntaş; best friend, life coach, and advisor… The hidden supervisor of this thesis with unfailing moral support and persistent encouragement. Thank you, for everything…*

*Mehmet Uzbaş and Ayşegül Uzbaş, thank you my siblings, for your eternal friendship, and emotional support throughout my life and the whole Ph.D. process.*

*Şerife Uzbaş and Mustafa Uzbaş, my parents and first teachers in life. My perseverance, determination, and enthusiasm to learn is your inheritance; thank you for supporting me in all the steps of my life… You deserve my deepest gratitude for not only providing me a good education but also for raising me as a kind human being that cares for the others and the whole earth, which I believe is the primary merit towards becoming a true scientist…*

# Table of Contents

## List of Figures

## List of Tables

## List of Appendices

## Abbreviations

| | |
|---|---|
| BART-Seq | Barcode Assembly for Targeted Sequencing |
| Bc | barcode |
| BMP | Bone Morphogenetic Protein |
| bp | base pair(s) |
| Cat. No | catalogue number |
| cDNA | complementary DNA |
| Ct | cycle threshold |
| DMEM | Dulbecco's Modified Eagle's Medium |
| dNTP | deoxyribonucleotide triphosphate |
| Dox | Doxycycline |
| DPBS | Dulbecco's Phosphate-Buffered Saline |
| EDTA | ethylenediaminetetraacetic acid |
| emPCR | emulsion PCR |
| EpiSC | epiblast stem cells |
| ERCC | External RNA Controls Consortium |
| EST | Expressed Sequence Tag |
| FACS | Fluorescence-Activated Cell Sorting |
| FBS/FCS | Fetal Bovine/Calf Serum |
| FGF | Fibroblast Growth Factor |
| g | gram(s) |
| gDNA | genomic DNA |
| GSK3 | Glycogen Synthase Kinase 3 |
| GWAS | genome-wide association study |
| h | hour(s) |
| hESC/mESC | human/mouse embryonic stem cell |
| hiPSC | human induced pluripotent stem cell |
| ID | identity |
| indel | insertion or deletion of bases |
| iPSC | induced pluripotent stem cell |
| IVT | *in vitro* transcription |
| KSR | Knockout Serum Replacement |
| l | liter |
| LCM | laser capture microdissection |
| LIF | Leukemia Inhibitory Factor |
| lncRNA | long non-coding RNA |
| M | molar |
| MEF | mouse embryonic fibroblast |
| MEK | Mitogen-Activated Protein Kinase |
| mfe/MFE | minimum free energy |
| min | minute(s) |
| MIP | Molecular Inversion Probe |
| mRNA | messenger RNA |

| | |
|---|---|
| nb-glm | negative binomial generalized linear model |
| NEAA | non-essential amino acids |
| NEB | New England Biolabs |
| NGS | Next-Generation Sequencing |
| nt | nucleotide(s) |
| PCA | principal component analysis |
| PCR | polymerase chain reaction |
| phos | Phosphate |
| PreAmp (PCR) | pre-amplification PCR |
| PSCs | pluripotent stem cells |
| qPCR | quantitative polymerase chain reaction |
| PTO | phosphorothioate |
| rc | reverse complementary |
| RCA | rolling circle amplification |
| RNA-Seq | RNA sequencing |
| rRNA | ribosomal RNA |
| RT | reverse transcription |
| rWnt3a | recombinant Wnt3a |
| SAGE | Serial Analysis of Gene Expression |
| scRNA-Seq | single-cell RNA sequencing |
| SNP | single nucleotide polymorphism |
| TGFß | transforming growth factor-ß |
| t-SNE | t-Distributed Stochastic Neighbor Embedding |
| U | Unit(s) |
| UMAP | Uniform Manifold Approximation and Projection |
| UMI | Unique Molecular Identifier |
| λ-exo | Lambda exonuclease |

# 1 INTRODUCTION

## 1.1 Transcriptomics

The transcriptome is defined as the set of all transcripts produced in a cell, which characterize a certain physiological or pathological state (Piétu et al., 1999). Although proteins are the incarnation of one-dimensional digital genetic code as "flesh and blood", it is often laborious to identify and quantify them. Therefore, measurement of mRNA molecules is traditionally used as a proxy (Svensson et al., 2017). The coding transcriptome transfers the genetic information from the genomic DNA to ribosomes in the form of messenger RNAs (mRNAs) for the synthesis of proteins, the ultimate functional products of the central dogma. The non-coding transcriptome comprises 98% of the transcripts, and includes ribosomal RNAs (rRNA), transfer RNAs (tRNA), long noncoding RNAs (lncRNAs), and small RNAs (miRNAs, promoter associated RNAs), which serve structural, epigenetic, and regulatory functions (Ozsolak and Milos, 2011).

Besides the genes that are constantly expressed based on a cell's identity, different sets of transcripts are produced during development, and in response to internal or external cues in homeostatic or disease conditions (Lowe et al., 2017). Gene expression is fine-tuned by mechanisms such as alternative promoter usage, allele-specific expression, or alternative splicing. Alternative splicing, for instance, is known to be important for stem cell differentiation and development (Salomonis et al., 2010). Disruption of these regulatory mechanisms for both coding and non-coding RNAs are implicated in many inherited and acquired diseases, including cancer, and cardiovascular and neurological disorders (Esteller, 2011; Lee and Young, 2013).

### 1.1.1 Analysis of gene expression

The discovery of the reverse transcriptase in 1970 marks a key milestone for transcriptomics (Baltimore, 1970; Temin and Mizutani, 1970), as it allowed the conversion of rather unstable RNA molecules into complementary DNAs (cDNA) with the same sequence, which are much easier to preserve and analyze. Since then, numerous techniques have been developed to determine the sequence and abundance of the transcripts in cells (**Figure 1**), which fall broadly under two categories; hybridization-based (indirect) and sequencing based (direct) methods.

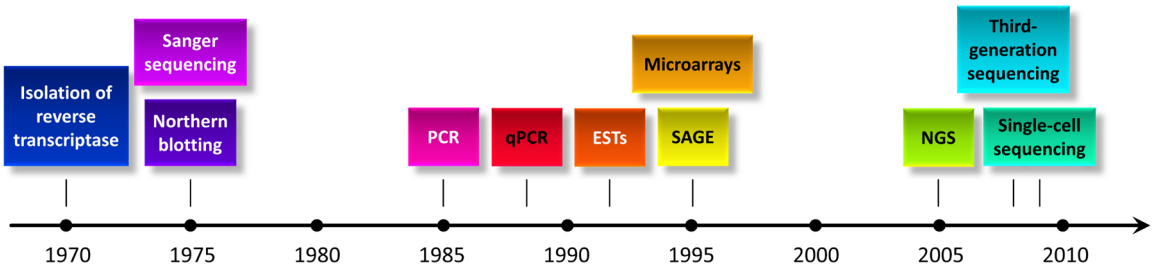



**Figure 1: The timeline of transcriptomics**

#### 1.1.1.1 Hybridization-based methods

**Hybridization-based** methods analyze transcripts indirectly by measuring secondary signals. Northern blotting is one of the earliest examples of this type, which involves gel electrophoresis followed by hybridization of the labeled complementary probe and detection of the transcript (Alwine et al., 1977). Later, nylon membrane arrays (macroarrays) allowed analysis of several transcripts at once, whereas their processing was laborious (Lowe et al., 2017). Invention of the oligonucleotide microarrays scaled the number of target sequences up to thousands, which today is still a powerful high-throughput method widely used in genomics and transcriptomics. It is based on printing numerous oligonucleotide spots on a solid surface, each of which is complementary to a fragment of their target (Schena et al., 1995). Test and control samples differentially labeled with fluorescence (e.g. red vs green) are simultaneously hybridized to the array for relative measurement of thousands of targets. On the downside, prior knowledge of the sequences is required to manufacture the arrays, and sensitivity and dynamic range are rather limited (Saliba et al., 2014).

**Real-time PCR** (also known as **qPCR**) is another indirect method widely used for quantification of gene expression (Higuchi et al., 1993) that is based on measuring the signals emitted from samples in each cycle of PCR during the exponential phase, using DNA intercalating dyes or fluorophore-tagged probes. It allows simultaneous analysis of template concentrations that differ by several orders of magnitude, with sub-picogram sensitivity. Although qPCR can analyze hundreds of samples in parallel, it is not truly high-throughput in terms of the number of targets (Lowe et al., 2017; Marín de Evsikova et al., 2019).

#### 1.1.1.2 Sequencing-based methods

A direct approach to analyze gene expression is sequencing that determines the order of nucleotides in transcripts without prior knowledge. Sequencing-based methods can be classified into three main categories; first-generation (since 1975), next-generation (since 2005), and third-generation (since 2008) (**Figure 2**).

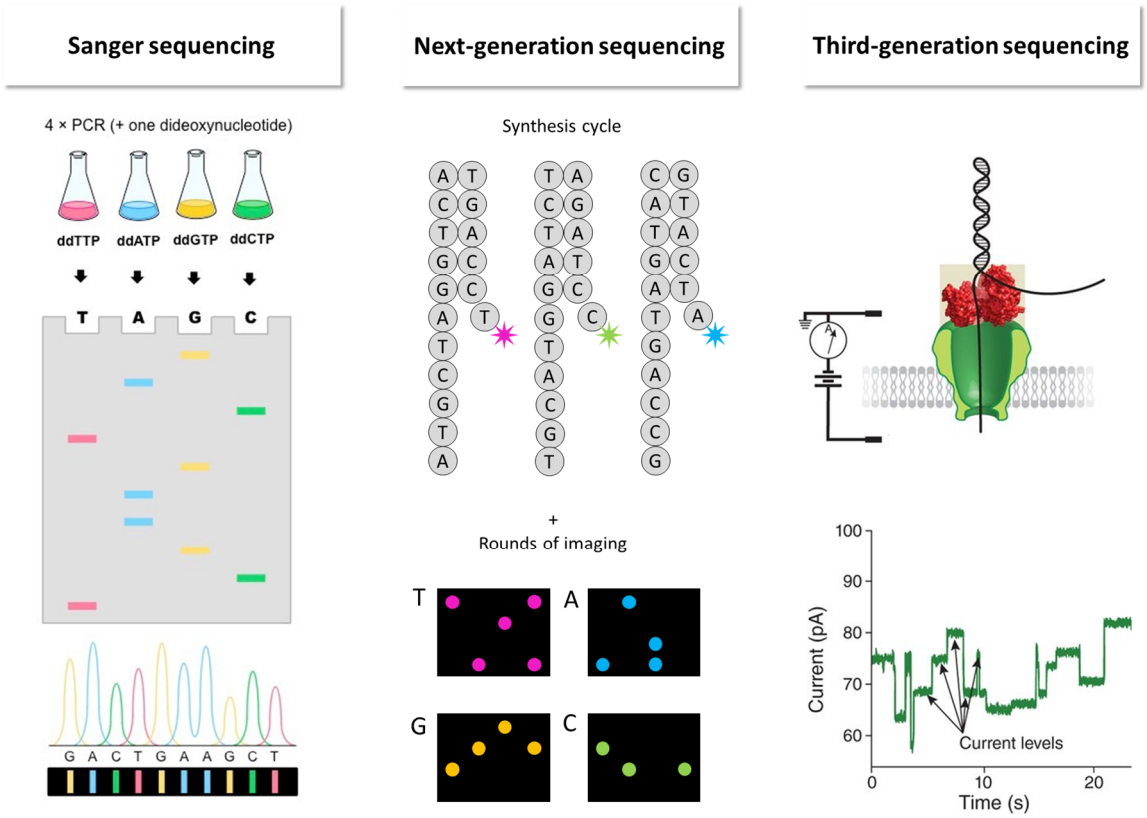



**Figure 2: Three generations of sequencing-based transcriptomics** (left and right panels are adapted from Online Biology Notes[1] and Deamer et al. (2016), respectively)

#### 1.1.1.2.1 First-generation (Sanger) sequencing

Introduction of the Sanger method in 1975 launched the **first generation of sequencing**, which utilized the primer extension and chain-termination idea (Sanger and Coulson, 1975). Including small amounts of a modified di-deoxynucleotidetriphosphate (ddNTP) in the dNTP mixture causes the DNA polymerase to stop at semi-random positions during replication, resulting in a mixture of DNA strands with varying lengths terminated in theoretically all possible positions where the specific nucleotide is found. Gel electrophoresis of the reactions each performed with one of the four ddNTPs reveals position of the bases in the template DNA fragment (**Figure 2**, left).

Adoption of Sanger sequencing, in combination with the routine production of cDNA libraries from different individuals and species enabled massive endeavors such as the Human Genome Project, a major leap forward for discovery of genes, understanding their regulation, and disease. For example, Expressed Sequence Tags (ESTs), random fragments (100-800 nt) cloned from cDNA libraries, allowed *de novo* discovery of genes from various species (Adams et al., 1991). While ESTs were useful for the analysis of individual genes, they did not allow comparative quantitative

[1] https://www.onlinebiologynotes.com/sangers-method-gene-sequencing/

analyses and were low-throughput. Subsequently introduced Serial Analysis of Gene Expression (SAGE) covered a larger portion of the transcriptome by concatenating small (11 nt) random fragments from each mRNA, and enabled quantification of transcript frequencies by counting fragments (Velculescu et al., 1995). Because entire transcripts were not sequenced, it was not optimal for homologous loci or repeats though (Marín de Evsikova et al., 2019).

#### 1.1.1.2.2 Next-generation sequencing

First introduced in 2005 (Margulies et al., 2005), **next-generation (second generation) sequencing** (NGS) revolutionized the transcriptomics and genomics fields by enabling parallel analysis of millions or billions of sequences. The key principle of NGS is determining the order of nucleotides by imaging during the synthesis of a complementary strand using fluorescently labeled nucleotides (**Figure 2**, middle), with the following basic steps: Preparation of an RNA sample for NGS begins with conversion to cDNA. It is possible to enrich the target RNA molecules, for example via rRNA depletion, reverse transcription with oligo(dT) primers (e.g. poly(A) mRNAs), or size selection (e.g. micro RNAs) (Lowe et al., 2017). Genomic DNA (gDNA) or cDNA samples are then fragmented to a size range compatible with the sequencing instrument and the kit; via chemical hydrolysis, nebulization, sonication, or tagmentation. Adapters are attached to both ends of the fragments, which are typically used for (optional) PCR enrichment. Finally, single-stranded fragments are captured on the sequencing surface via adapters, and clonal amplification takes place (e.g. Illumina) to create tight clusters of several hundred copies of the initial oligonucleotides, to ensure the visibility of the fluorescent signal to the imaging system during sequencing. In some systems, clonal amplification is performed on the surface of the beads in emulsion (emPCR), which are subsequently captured on a surface (e.g. Roche/454). Next, sequencing synthesis takes place using fluorescently labelled nucleotides, and interpretation of multiple images taken at each fluorescent channel in each sequencing cycle determines the order of nucleotides per cluster (Metzker, 2010). There are various techniques that differ in one or more of these steps such as pyrosequencing (Roche), sequencing by synthesis (Illumina and Life Technologies), or sequencing by ligation (SOLiD) (Kulski, 2016). Ion Torrent from Life Technologies differs from these, in that it measures the voltage changes caused by the released $H^+$ ions during the synthesis reaction instead of imaging fluorescent signals (Rothberg et al., 2011).

NGS technology quickly replaced numerous applications of microarrays and other transcriptomic techniques, since it provides higher throughput, speed, resolution, and dynamic ranges up to five orders of magnitude, while lowering costs and input materials. In contrast to SAGE or microarrays, the entire length of the transcripts can be determined by NGS. Since the NGS technology requires no prior knowledge of the target sequences, it enables discovery of new forms of gene regulation including transcription initiation sites, 5' and 3' untranslated regions (UTRs),

alternative splicing events, sense-/anti-sense transcripts, modifications (e.g. indels, SNPs)[2], gene fusion events, and more (Ozsolak and Milos, 2011; Saliba et al., 2014). For example, it revealed pervasiveness of the transcription, i.e. although only 2% of the human genome is protein-coding genes, more than 80% of it is transcribed, adding to the overall complexity (Hangauer et al., 2013).

Next-generation sequencing boosted many research fields. Many disease mechanisms are known to relate to transcription, such as alternative promoter usage, allele usage, modifications of the regulatory elements (i.e. control of transcription), or SNPs (Lowe et al., 2017). Genome-wide association studies (GWAS) aim to identify the complete genome or exome sequences of complex organisms to gain insights into the full range of variation, and to learn how they influence the phenotypic traits and disease progression (Mamanova et al., 2010a). As opposed to Mendelian diseases that can be explained with a single gene, many of the prevalent diseases such as autism, obesity, diabetes, and schizophrenia are multifactorial, caused by multiple genetic and environmental factors, and their exact molecular pathophysiologies remain to be explained (Karczewski and Snyder, 2018). Next-generation sequencing aided the GWAS projects greatly by allowing routine and cheap resequencing of individual genomes and transcriptomes. In addition to disease discovery, high-throughput sequencing has already started to benefit personalized therapies as well. Since disease progression has a multifactorial nature in complex organisms, resequencing information can be readily used to determine the optimum treatment regime based on the patient-specific targets, rather than relying on the ultimate symptoms (Karczewski and Snyder, 2018). Intermediate screening can enable timely modification or fine-tuning the treatments based on the patient's response, for example in cancer (Marín de Evsikova et al., 2019).

Beyond NGS, which requires the conversion of RNA molecules to cDNA first (Saliba et al., 2014), there are also a few methods that can infer the sequence of RNA molecules directly, without reverse transcription or amplification, which preserves strand specificity that is often lost with the methods that involve amplification. For example, FRT-Seq analyzes the poly(A)+ RNAs directly on the flow cell during cDNA synthesis by reverse transcriptase. The fragmented RNA is ligated to two DNA-RNA hybrid adapters homologous to Illumina's P5 and P7 primers, captured by the flow cell, and sequenced (Mamanova et al., 2010b). Direct RNA-Seq (DRS) is a similar method where poly(A)+ RNA molecules are captured on a flow cell with oligo(dT) probes on the surface, and sequencing takes place during cDNA synthesis using a special polymerase with reverse transcription function (Helicos BioScience; Ozsolak and Milos, 2011). Nevertheless, these methods hinge on the synthesis of a complementary strand, which does not preserve base modifications, and the reads are too short for capturing alternative splicing events in eukaryotes. Although sequencing full-length cDNA molecules was made possible via strand switching and

[2] https://www.illumina.com/science/technology/next-generation-sequencing/microarray-rna-seq-comparison.html

using a long-read sequencer, it can still suffer from the problems related to reverse transcription (Thomas et al., 2014).

#### 1.1.1.2.3 Third-generation sequencing

The **third generation of sequencing** is the direct determination of template sequences, including modified nucleotides (e.g. epigenetic modifications), without any amplification or reverse transcription step, which is not possible with the NGS workflows. An example is single-molecule real-time (SMRT) sequencing that uses zero-mode waveguide (ZMW) nanostructures to significantly reduce the observation area per single reaction in order to detect the signals over the background (Pacific Biosciences, Eid et al., 2009). A template strand is attached to the DNA polymerase that is immobilized at the bottom of each structure, and nucleotide-specific fluorophores released during synthesis of the complementary strand are imaged. While templates of >10 kb can be sequenced with this method, the error rate of 10-13% is still higher than NGS (Picelli, 2017). Nanopore is another technology (**Figure 2**, right), which passes the DNA/RNA molecules through protein nanopores that are embedded on a synthetic hydrophobic membrane (MinION/PromethION from Oxford Nanopore Technologies[3], Garalde et al., 2016). In this system, the current passing through individual nanopores is continuously recorded, which is modified by the nucleic acid translocating through the nanopore in a sequence-specific manner. To enable recording, the strands are attached to a motor protein that slows down the process. Nanopore systems can sequence full length of DNA/RNA molecules (>2 million bp) directionally, including non-poly(A) transcripts, however their throughput is currently lower than NGS (Picelli, 2017). Although still in its infancy, the third generation of sequencing is a promising tool of the future.

#### 1.1.1.2.4 Targeted sequencing

More than 90% of the transcripts have less than 50 copies per cell despite having central roles in biological processes, including critical genes such as signaling proteins and transcription factors. Due to the fact that the global (unbiased) RNA-sequencing (RNA-Seq) approaches randomly sample the transcripts, sensitive detection and in-depth analysis of lowly expressed genes are hindered as the most abundant (e.g. housekeeping) genes consume majority of the sequencing reads (Eberwine et al., 2014; Mercer et al., 2014). Therefore, targeted analysis of a small number of loci of interest can be advantageous in particular cases over exhausting the sequencing resources for the uninteresting information (Hodges et al., 2007).

With a fixed number of total reads consumed for a specific experiment, enriching the transcripts of interest over others (targeted sequencing) can maximize the coverage, e.g. of the lowly expressed genes (**Figure 3**). Targeted sequencing can be used for a broad range of applications such as detecting mutations, RNA editing events, and fusion transcripts; studying the dynamics of specific processes (e.g. stem cell

[3] https://nanoporetech.com/products

differentiation) or screening the response of selected genes to a compound library. A large number of samples can be analyzed in parallel while minimizing sequencing costs and time, which would benefit both clinical and research applications (Li et al., 2012). There are different methods for enrichment of selected sets of transcripts, which can be classified under the following categories:

- Amplification based:
  - PCR based
  - MIP-based
- Hybridization-based:
  - On-array capture
  - In-solution capture

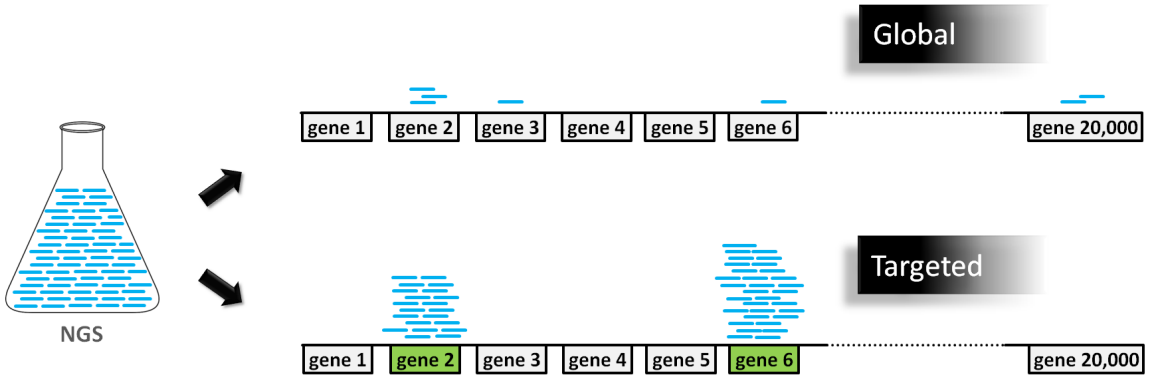



**Figure 3: Global (unbiased) and targeted sequencing approaches** trade the number of genes with the sequencing depth reciprocally (when the output is fixed)

One strategy to enrich the selected set of transcripts is **amplification**, which can be done either per target via individual reactions, or in a multiplexed manner. For example, Craig et al. (Craig et al., 2008) resequenced 46 individuals in parallel, to identify genetic variants. Multiple 5 kb regions from each individual were co-amplified, samples were fragmented, barcoded and mixed in equimolar concentrations for sequencing. Commercial platforms such as Ion AmpliSeq gene panels (Life Technologies) also use the multiplex PCR technology. Multiplex PCR can have certain disadvantages such as uneven efficiency of different primer pairs, non-specific amplification, or cross-hybridization of the pooled primers. This can be avoided if each target can be amplified in isolation, such as the RainStorm platform[4], which is based on parallel individual PCR reactions that take place in microdroplets (Mamanova et al., 2010a; Tewhey et al., 2009). The fragmented sample is captured together with the PCR reagents within multiple droplets, each of which receives a single primer pair targeting a specific locus. The droplets per target are combined equimolarly and PCR is run within each droplet, which are subsequently combined and prepared for sequencing. Isolation of the reactions circumvents the problems inherent to multiplex PCR, and thousands of loci can be

[4] http://raindancetech.com/

co-analyzed this way; however, cost and processing time limits the length of the targets (<2-3 Mb) and high-throughput analyses.

**Capturing by circularization** is based on the principle of placing inversely oriented target-specific probe pairs at the 3' and 5' of an oligonucleotide stretch with a spacer, to increase the specificity of multiplex amplification, such as padlock or molecular inversion probes (MIPs) (Mercer et al., 2014). Following hybridization, the gap in between the probes are filled-in, and the resulting circular molecule contains the target fragment and the spacer, which contains loci for global primers to amplify the captured DNA stretch with PCR or rolling circle amplification. The probes on the original padlock systems hybridized to the entire target without any space, and the gap is closed via ligation, which required full complementarity. The MIP system derived thereof contains a space between the probes, which is filled in by a polymerase using the target as the template, allowing to capture also the variable sequences in between, e.g. SNPs or indels (Porreca et al., 2007; Turner et al., 2009). Nonetheless, target-specific sequence of each probe lowers the uniformity of the padlock/MIP systems compared to hybridization-based methods (Mamanova et al., 2010a).

Another strategy to select the sequences of interest is **hybridization** to pre-designed probes, either on arrays or in solution. The first adaptation of the **array-based target capturing** to NGS was established by NimbleGen (Hodges et al., 2007). It is much faster and easier to perform in comparison to PCR, yet requires expensive instruments and relatively large amount of starting material (10-15 μg of DNA), and is not suitable for parallelization of many samples. **In-solution capturing**, on the other hand, requires lower amount of starting material compared to array-based versions and does not need special instruments (Mamanova et al., 2010a). An example is the RNA CaptureSeq, in which a pool of custom oligonucleotides attached to beads are used to capture the targets of interest from fragmented samples, and then are pulled down. The use of multiple probes for the same target can help normalizing the variations (like in microarrays) that might arise from individual probes, which can ensure a higher uniformity (Mercer et al., 2014). Nevertheless, in comparison to amplification-based methods, hybridization-based target enrichment demands higher amount of starting material (Ozsolak and Milos, 2011), which does not suit single-cell applications.

To my knowledge, there are currently two methods for targeted transcriptomics of single cells. CytoSeq combines oligo(dT) capturing and gene-specific primers to analyze up to 111 genes (3' ends) in tens of thousands of cells using nanowell plates (Fan et al., 2015). The recently introduced RAGE-Seq combines targeted nanopore sequencing of full-length transcripts with the short-read transcriptome sequencing in single cells (Singh et al., 2019).

### 1.1.2 Single-cell analysis

Human body is made up of tens of trillions of cells that can be categorized merely into about 200 cell types. Even if located in the same tissue, the cells of the same type often have heterogeneous gene activity though, which plays important roles during development, homeostasis, and disease (**Figure 4**). Fluctuations of gene expression during embryonic development allow the cells to explore alternative lineages. In adult tissues, it ensures the continuous presence of a small population of cells that are ready to rapidly respond to physiological or external cues, thereby warrant adaptiveness. Besides, whereas all the cells of an organism hypothetically contain the same genome, exceptions are common, such as the immune system, germline cells, tumor cells, as well as replication-related somatic mutations.

Traditionally, biological mechanisms are studied using materials obtained from pools of thousands or millions of cells in order to have enough material to study, which, as a result provides averaged information on the sampled cell population. However, it is impossible to know whether these values reflected an underlying uniform profile or are the average of bimodal or multimodal subpopulations. It is now known that gene expression levels varies significantly (up to 1000-fold) even within the presumably homogenous cell populations (Raj et al., 2006).

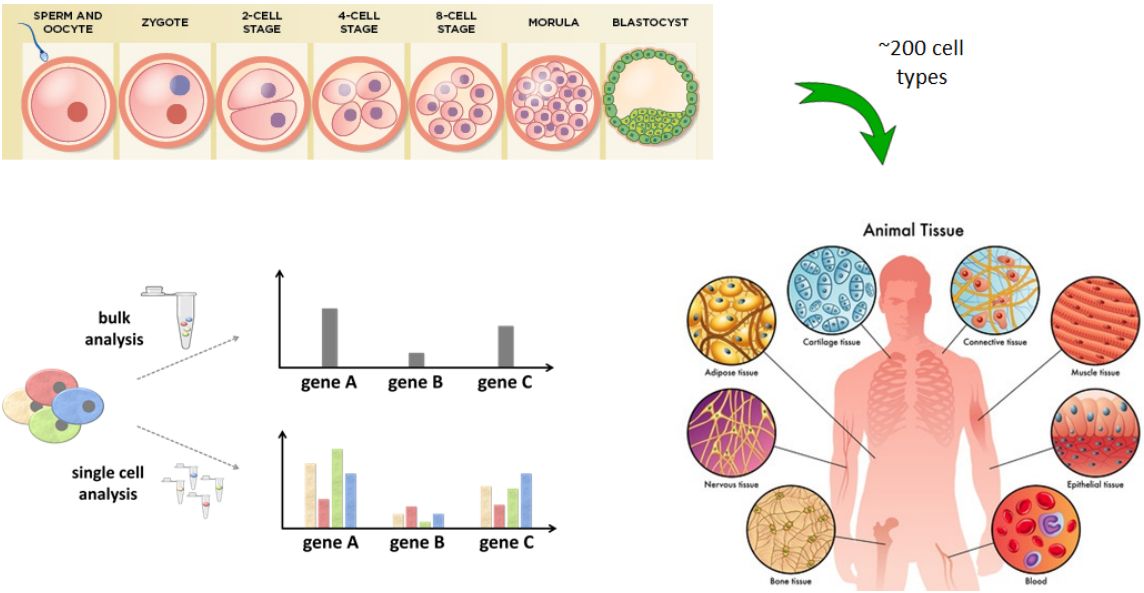



**Figure 4: Analysis of cells in bulk masks the underlying heterogeneity**, which plays crucial roles in development, homeostasis, and disease. Single-cell analysis can provide higher-resolution and more accurate information. Top and right figures were adapted from online resources[5,6]

[5] https://cdn.the-scientist.com/assets/articleNo/30175/iImg/1098/-kst1.jpg
[6] http://www.marketreportgazette.com/wp-content/uploads/2019/07/Organ-Transplant-Immunosuppreant.jpg

Being able to analyze single cells can provide unprecedented insights into biological mechanisms, and address questions that were impossible to answer previously. How to create a colossally complex organism step-by-step starting from a single zygote? Is stem cell differentiation stochastic or deterministic, is it reversible, and when does it become irreversible? Reconstructing developmental lineage trees, for example by examining accumulated somatic mutations in single cells can partially answer these questions. Significant rare cells can be investigated, such as adult stem cells in tissues, circulating tumor cells in the blood, or cells that cause resistance to antibiotics (Picelli, 2017). Single-cell analysis of biopsies can provide insights into the tumor microevolution and cancer relapse and advance the strategies to target them using precision medicine (Shapiro et al., 2013). Environmental studies can benefit from single-cell analysis, too, e.g. for novel discoveries of microorganisms that cannot be cultured in the lab (Saliba et al., 2014).

#### 1.1.2.1 Single-cell sequencing techniques

Following the first single-cell RNA-sequencing experiment that reported the analysis of "seven" cells (Tang et al., 2009), the technology grew tremendously over the past ten years that simultaneous sequencing of over a million cells is possible today (**Figure 5**). This is made possible by co-development of different techniques for isolating single cells, capturing and amplification of the transcripts, introducing the barcodes, and analyzing the data. A technical summary of the techniques is given in **Table 1**.

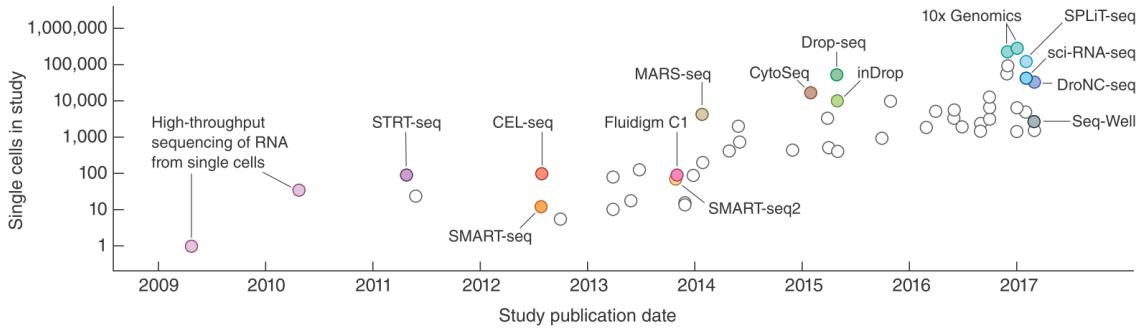



**Figure 5: The exponential growth of single-cell sequencing technologies over the past ten years** (Svensson et al., 2017)

##### 1.1.2.1.1 Isolating single cells

The very first step of single-cell analysis is their proper isolation from the primary tissue or the culture plate, and transfer to the reaction site (**Figure 6**). Following is an overview of the existing techniques (based on Hwang et al., 2018; Marín de Evsikova et al., 2019; Picelli, 2017; Saliba et al., 2014):

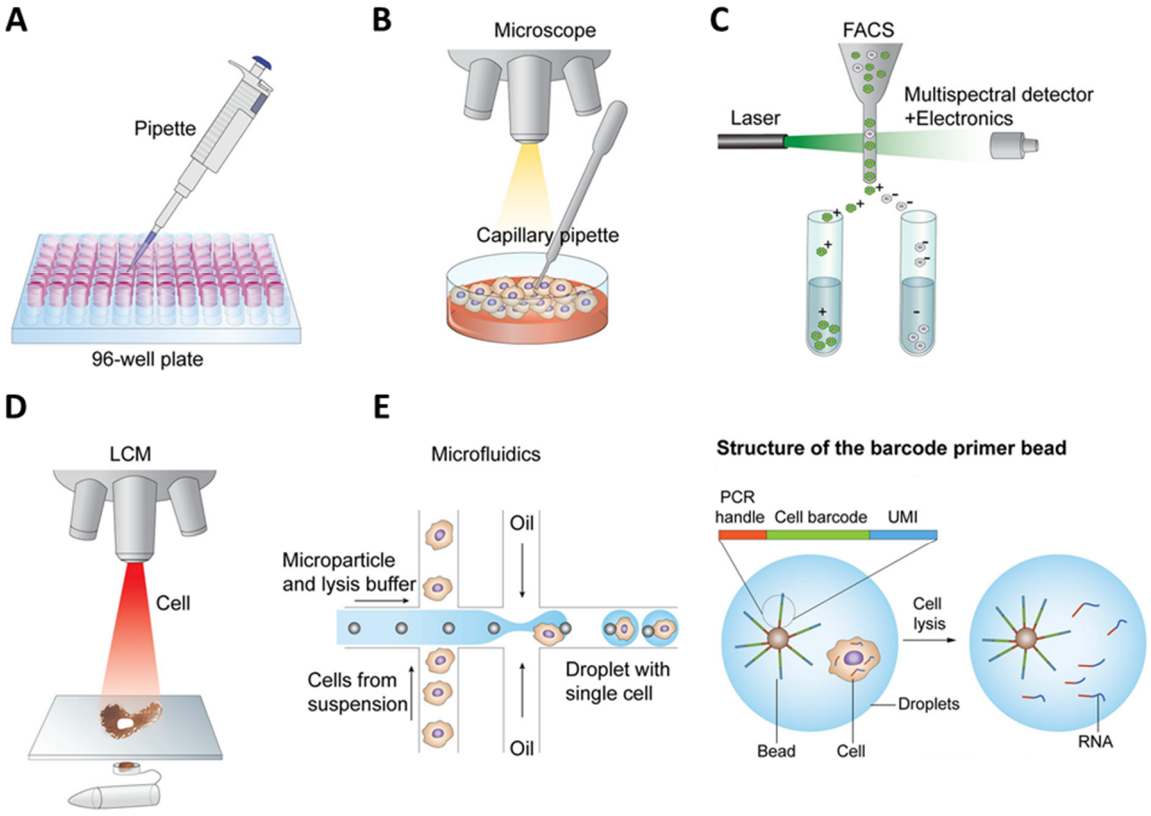



**Figure 6: Single cell isolation techniques. (A)** Limiting dilution. **(B)** Micromanipulation. **(C)** Fluorescence-activated cell sorting (FACS). **(D)** Laser capture microdissection (LCM). **(E)** Droplet emulsion. Figure modified from Hwang et al. (2018)

- **Micromanipulation** is the isolation of cells from a suspension e.g. using mouth pipette under microscope. Pros: cells can be observed and handling also the fragile cells is possible. Cons: cells have to be in suspension, low-throughput, labor-intensive.
- **Optical tweezers** use laser beams to hold and move the cells. Pros: cells can be observed. Cons: cells have to be in suspension.
- **Laser capture microdissection (LCM)** uses laser beams to dissect the cells from solid tissues. Pros: spatial information is preserved. Cons: low throughput, might not recover the whole cytoplasm, thus sub-optimal for transcriptomics.
- **Limiting dilution** of the cells to a certain concentration allows sampling single cells based on Poisson distribution. Cons: majority of the wells will contain zero cells rather than one, leading to unnecessary reagent consumption.
- **Fluorescence-activated cell sorting (FACS)** isolates highly purified single cells, as well as single nuclei using an electric field (e.g. MARS-Seq, snRNA-Seq). Pros: can be either an unbiased (all live cells) or biased technique (specific size/morphology or marker expression), high-throughput, economical, easy, accessible by many research groups. Cons: requires large number of cells, suboptimal for mixture of cells with different sizes.
- **Microfluidic chips** offer compartmentalized nanoliter-sized units to capture cells, into which additional components can be transferred in a controlled manner (e.g. Fluidigm C1). Pros: low sample consumption, reduced risk of

contamination due to closed circuitry, high throughput. Cons: limited cell size range, expensive chips.

- **Nanowells/microwells** are microfabricated surfaces containing hundreds of thousands of wells that capture cells via limiting dilution (e.g. ICELL8, CytoSeq). Pros: works by gravity, possible to visually inspect the wells. Cons: only a small percentage of wells contain cells.
- **Droplet emulsion** captures the cells and reagents in tiny aqueous droplets enclosed by an oil phase (e.g. InDrop, Drop-seq, 10x Chromium). Pros: low cost per cell, high throughput. Cons: Large number of starting cells are required, majority of the droplets are empty.

#### 1.1.2.1.2 RNA capturing and amplification

The next step to isolating the cells is recovery and amplification of their transcripts. The workflows roughly consist of the following steps: reverse transcription, second strand synthesis (optional), amplification, fragmentation, and library preparation (Shapiro et al., 2013) (**Figure 7**):

**Reverse transcription** is often performed using oligo(dT) primers, either dissolved in reaction solution (e.g. plate-based or microfluidic platforms) or attached to beads (e.g. droplet or nanowell formats). Oligo(dT) stretch is flanked with a universal PCR primer locus for the subsequent amplification. Later methods additionally include a stretch of cell barcode and a molecular barcode -unique molecular identifier (UMI)- in between the oligo(dT) and the universal primer (**Figure 6E**). The cell barcodes -unique per cell- are used to sort the reads to the cell of origin, while the UMIs -unique per molecule- are used to correct for the amplification bias (Kivioja et al., 2012). Most of the techniques capture the transcripts at the poly(A) tail and count the sequences in the 3' or 5' end (tag-based) at the expense of full-length coverage (**Table 1**). There are a few techniques that can provide full-length coverage, such as Smart-seq that uses template switching and RamDA-seq that combines random displacement amplification with not-so-random primers to analyze full length (>10 kb) RNA including non-poly(A) transcripts (Hayashi et al., 2018; Picelli et al., 2013). Notwithstanding the in-depth information they offer about transcripts, e.g. isoforms or SNPs, these methods are often limited to a smaller number of cells due to labor and costs, and they might not accommodate cell barcodes.

Many of the methods include a **second strand synthesis** step. One strategy is **homopolymer tailing**, which is the addition of ~30 nt poly(A) tail to the 3' end of the first strand cDNA using a terminal deoxynucleotidyl transferase (Quartz-Seq: Sasagawa et al., 2013; Tang et al., 2009). Subsequently, oligo(dT) flanked by a second primer locus is used to synthesize the second strand. The drawbacks include loss of strand information and 3' bias, thus uneven coverage of the transcripts, because reverse transcription may terminate prematurely resulting in incomplete sequences (Picelli, 2017; Saliba et al., 2014). Another strategy is **template switching** (SMART: Switching Mechanism at the 5′ end of the RNA Transcript) that uses Moloney Murine leukemia virus reverse transcriptase (*M-MuLV RT*) to ensure full transcript coverage since only complete mRNA molecules are processed

(Zhu et al., 2001). It is based on an intrinsic property of the *M-MuLV RT* to add 3-4 cytosines to the 3' end of the first strand cDNA. When a universal primer ending with a short poly(G) motif is added to the reaction, it is anchored by the newly synthesized poly(C) stretch, and the reverse transcriptase proceeds to synthesize the second strand with its DNA dependent DNA polymerase activity. While majority of the methods target the poly(A) tails of mRNAs, there are a few exceptions, for example, MATQ-Seq is able to analyze non-poly(A) transcripts (Sheng et al., 2017), and Hayashi et al. (2018) designed not-so-random (NSR) primers, which are bioinformatically optimized to target all the RNA molecules except rRNAs.

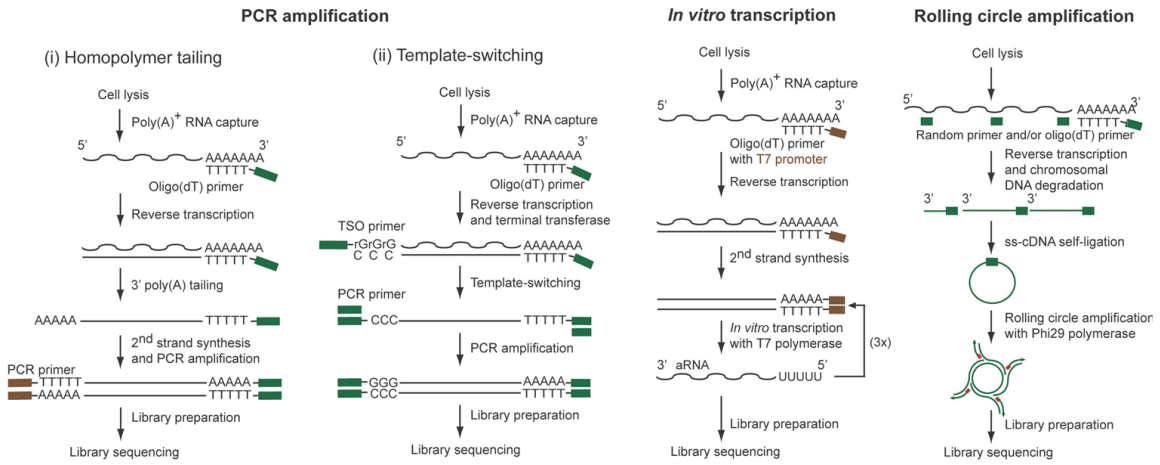



**Figure 7: Methods used for capturing and amplification of transcripts from single cells** (Saliba et al., 2014)

**Amplification** is the next step of the workflows due to the limited RNA content of single cells (1-50 pg) (Livesey, 2003) (**Figure 7**). For the methods that flank the transcripts with universal primer loci on both ends, **PCR** is the choice for initial amplification of the full-length transcripts. The following tagmentation simultaneously adds the sequencing adapters while fragmenting the amplicons. ***In vitro* transcription (IVT)** is used by the methods that attach a T7 promoter to the oligo(dT) primers during the first strand synthesis. Anti-sense RNAs (aRNA) are produced via IVT, which are then fragmented, reverse transcribed, and prepared for sequencing. Although linear amplification with IVT prevents potential PCR artifacts, it tends to cause 3' bias, and second reverse transcription might decrease the overall efficiency (Ziegenhain et al., 2018). A third strategy is **rolling circle amplification (RCA)**, in which Phi29 DNA polymerase is used to amplify the circularized cDNA (Pan et al., 2013).

Single-cell Combinatorial Indexing RNA-seq (sci-RNA-seq) is a recently introduced method that couples the cell isolation with barcoding and amplification, using the fixed cells or nuclei as *in situ* reaction chambers (Cao et al., 2017, 2019). Combinatorial barcoding is achieved via splitting the cells as pools into 96/384-well plates for the first round of barcoding, then collecting and mixing them, and

splitting again for the second round of barcoding. SPLiT-Seq is a similar method (Rosenberg et al., 2018) that includes multiple split-pool rounds, during which barcodes are added by ligation. These methods achieve analysis of very large number of cells, scaling up to millions.

**Table 1:** Technical summary of single-cell sequencing methods

| Method | Strategy | Throughput | Transcriptome | Coverage | Barcoding | Cell isolation | Amplification | Reference |
|---|---|---|---|---|---|---|---|---|
| Tang method | Homopolymer tailing | 7 | Whole | Full | - | Manual | PCR | Tang et al. 2009 |
| Smart-seq2 | Template switching | $10^2 – 10^3$ | Whole | Full | - | FACS | PCR | Picelli et al. 2013 |
| STRT/C1-Seq | Template switching | $10^2$ | Whole | 5′ | Cell barcode + UMIs | Microfluidics | PCR | Islam et al. 2014 |
| MARS-Seq | IVT | $10^2 – 10^3$ | Whole | 3′ | Cell barcode + UMIs | FACS | IVT | Jaitin et al. 2014 |
| DropSeq | Template switching | $10^3 – 10^4$ | Whole | 3′ | Cell barcode + UMIs | Droplet | PCR | Macosko et al. 2015 |
| inDrop | IVT | $10^3 – 10^4$ | Whole | 3′ | Cell barcode + UMIs | Droplet | IVT | Klein et al. 2015 |
| CytoSeq | Multiplex PCR | $10^3 – 10^5$ | Selected genes | 3′ | UMIs | Microwells | PCR | Fan et al. 2015 |
| CEL-Seq2 | IVT | $10^2 – 10^3$ | Whole | 3′ | Cell barcode + UMIs | Microfluidics | IVT | Hashimshony et al. 2016 |
| 10X Chromium | Template switching | $10^3 – 10^4$ | Whole | 3′ | Cell barcode + UMIs | Droplet | PCR | Zheng et al. 2016 |
| MATQ-seq | Homopolymer tailing | $10^2 – 10^3$ | Whole | Full | UMIs | Manual | PCR | Sheng et al. 2017 |
| Seq-Well | Template switching | $10^3 – 10^4$ | Whole | 3′ | Cell barcode + UMIs | Microwells | PCR | Gierahn et al. 2017 |
| Fluidigm HT IFC | | | Whole | 3′ | Cell barcode | Microfluidics | PCR | Fluidigm |
| ICELL8 | Template switching | $10^3$ | Whole | 3′ | Cell barcode + UMIs | Nanowells | PCR | Goldstein et al. 2017 |
| SPLiT-seq | Template switching | $10^3 – 10^5$ | Whole | 3′ | Cell barcode + UMIs | Split-pool, manual | PCR | Rosenberg et al. 2017 |
| Quartz-Seq2 | Homopolymer tailing | $10^3$ | Whole | 3′ | Cell barcode + UMIs | FACS | PCR | Sasagawa et al. 2018 |
| mcSCRB-Seq | Template switching | $10^2 – 10^3$ | Whole | 3′ | Cell barcode + UMIs | FACS | PCR | Bagnoli et al. 2018 |
| RamDA-seq | Not-so-random primers | | Whole + non-poly(A) | Full | - | FACS | Strand displacement | Hayashi et al. 2018 |
| sci-RNA-seq3 | Template switching | $2 \times 10^6$ | Whole | 3′ | Cell barcode + UMIs | Split-pool, FACS | PCR | Cao et al. 2019 |

#### 1.1.2.1.3 Analysis of single-cell sequencing data

Many tools were developed for analyzing the single-cell sequencing data, mostly based on Python or R; including Monocle (Trapnell et al., 2014), SEURAT (Satija et al., 2015), and Scanpy (Wolf et al., 2018). Some sequencing platforms offer accompanying software for complete analysis starting from raw reads, such as the Cell Ranger of 10x Chromium[7]. The main analysis steps include filtering out the low-quality cells, normalizing the data, dimensionality reduction and clustering the cells, and identifying differentially expressed genes or gene trajectories.

Processing the raw sequencing data results in a read count matrix that consists of cells and genes as columns and rows. The first step is **filtering** out the cells with sub-optimal biological or computational quality, for example when the transcripts could not be recovered due to failed capturing, incomplete lysis, degradation, or subsequent reactions. Low number of UMIs or detected genes, or high percentage of spike-in reads (if used) might indicate incomplete recovery of transcripts or captured ambient RNA molecules instead of cells. High number of UMIs or detected genes might indicate doublets. Increased percentage of mitochondrial transcripts may

[7] https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger

imply stress or cell death. Such samples are omitted from subsequent steps. Optionally, genes expressed in lower than a certain number of cells are also excluded (Luecken and Theis, 2019). Nevertheless, the cell types being analyzed, their metabolic demands and potential size heterogeneity within the population should be taken into account on an experiment-to-experiment basis when deciding on the filtering thresholds.

Since the sequencing depth per cell can vary due to biological or technical confounders such as dropouts or random sampling, **normalization** aims to scale the read counts per cell and per gene to comparable levels. Some methods simply presume equal number of transcripts in each cell; hence, they use the total UMIs detected in a cell as the scaling factor. Despite some limitations, use of ERCC RNA spike-ins (rather for plate-based methods) is another strategy to deconvolute the technical and biological variations, and to estimate the total mRNA count per cell (Ziegenhain et al., 2018). Nevertheless, the zero-inflated nature of single-cell data usually necessitates more complex approaches. To address this, various techniques were developed, such as using pooled read counts from multiple cells as reference, omitting the highest expressed genes, building non-linear models of the data, quantile regression, and so on, as discussed extensively in multiple reviews (Hwang et al., 2018; Luecken and Theis, 2019; Ziegenhain et al., 2018). An additional **correction** step might be necessary to adjust the variations coming from cell cycle stage or batch effects, for example when combining samples from different experiments, namely **data integration**, which is important for the projects accommodating data produced using various platforms, such as the Human Cell Atlas Project (Regev et al., 2017).

Although the number of dimensions of a count matrix equals to the number of genes detected, the underlying biological variability can indeed be explained with a much smaller set of dimensions. **Dimensionality reduction** aims to discover the inherent biological variance while reducing the computational burden (Luecken and Theis, 2019). Principal component analysis (PCA) is a linear and the simplest method, which also serves as a basis for the commonly used non-linear approaches such as t-distributed Stochastic Neighbor Embedding (t-SNE: Maaten and Hinton, 2008) or Uniform Manifold Approximation and Projection (UMAP: McInnes et al., 2018) that allows visualization of the multi-dimensional data in two dimensional space. There are further methods that visualize the movements and bifurcations of a cell population (e.g. diffusion maps: Haghverdi et al., 2015), and the connectivity among the clusters (e.g. PAGA: Wolf et al., 2019). Finally, **differential expression** analysis enables discovery of new pathways, gene regulatory networks, and the molecular mechanisms behind them.

#### 1.1.2.2 Alternative techniques for single-cell analysis

There are also non-NGS transcriptomics techniques to analyze single cells. For example, Fluidigm Biomark is built upon multiplexed pre-amplification of selected genes from single cells in a microfluidic device and subsequent analysis by qPCR

(Sanchez-Freire et al., 2012). The advantages include the ability to design and optimize primers, high sensitivity, specificity, and wide dynamic range. On the other hand, cells should be homogenous in size to comply with the microfluidic device (Hwang et al., 2018), the genes should be known in advance, and the data is based on secondary signals which might lead to false positive signals (Kalisky et al., 2018).

Another approach is fluorescent *in situ* sequencing that is based on fixing the cells on a surface, and sequencing via hybridization of fluorescent probes and imaging. Transcripts are either sequenced at single nucleotide resolution (e.g. FISSEQ: Lee et al., 2015), or estimated by the combinatorial information obtained using multiple oligonucleotide probes (e.g. MERFISH: Chen et al., 2015). These methods have complex workflows, high costs, and low throughput; nevertheless, they can be useful for analyzing a smaller number of cells with high resolution since they preserve spatial information of the RNA molecules.

## 1.2 Human Pluripotent Stem Cells

### 1.2.1 Pluripotent stem cells, *in vivo* and *in vitro*

Early embryonic development is an exceptional stage of a human's life, where a perfectly fine-tuned cascade starting from the **zygote**, a single cell with a single genome, flows towards a colossally complex structure. In a very short time, a small number of seemingly homogenous cells undergo major remodeling steps to initiate a process that will give rise to hundreds of different cell types in the body. It starts with the formation of the zygote which undergoes multiple cell divisions to create a ball of cells resembling a mulberry, named **morula** (*morus*: mulberry in Latin). Then, the cells divide further to create a fluid-filled ball named **blastocyst**, consisting of two cell lineages; the **trophoblast** that surrounds the structure and contributes to placenta, and the **inner cell mass** that gives rise to the embryo proper and some extraembryonic tissues (**Figure 4**).

**Pluripotency** is defined as the ability of cells to differentiate into all three germ layers (ectoderm, endoderm, mesoderm) and to the germline, but not to extraembryonic tissues (Weinberger et al., 2016). As the embryo develops from morula towards gastrula stage, pluripotent stem cells (PSCs) come into existence in a very narrow window. However, they could be locked in an indefinitely self-renewing state *in vitro* by controlling the culture components. The first **human embryonic stem cells (hESCs**) were derived from inner cell mass outgrowths of the donated IVF embryos by Thomson et al. (1998).

Naturally, development is a one directional flow of cells from less specialized to more specialized states, as exemplified first by Waddington's epigenetic landscape (Waddington, 1957). Once established, these states are very stable throughout the organism's life, lasting over multiple division cycles (Smith et al., 2016). However, Takahashi and Yamanaka (2006) showed that the cells could be driven back along

the Waddington's landscape to a pluripotent state by ectopically expressing as little as four factors: *OCT4*, *SOX2*, *KLF4*, and *c-MYC* (known as OSKM or Yamanaka factors) in mouse fibroblasts, which are termed **induced pluripotent stem cells (iPSCs)**. Human somatic cells, too, were reprogrammed to iPSCs a year later by two different research groups (Takahashi et al., 2007; Yu et al., 2007).

Since their first derivation, hESCs and hiPSCs became valuable tools for studying the pluripotency and early lineage commitment events, because the cells of the human embryo at the equivalent stage *in vivo* are inaccessible. Besides basic research, understanding the stepwise events giving rise to hundreds of different cell types from pluripotent stem cells enables us to understand the developmental roots of diseases, and would allow us in the future to mimic them to regenerate tissues and organs using the hiPSCs that are routinely generated today. Since diseased tissues are often not accessible and appropriate for experimentation, iPSCs will serve as standardized tools for high-throughput compound screening in pharmacology, and for precision medicine.

hESCs were traditionally maintained on mitotically inactivated mouse embryonic fibroblasts (MEFs) using fetal bovine/calf serum (FBS/FCS) and growth factors (Dakhore et al., 2018). Later, the factors secreted by MEFs were identified, which included fibroblast growth factors (FGFs), transforming growth factor-ß (TGFß), BMPs, and extracellular matrix components. Accordingly, different culture media formulations were defined that can maintain the hPSCs without MEFs (**Table 2**).

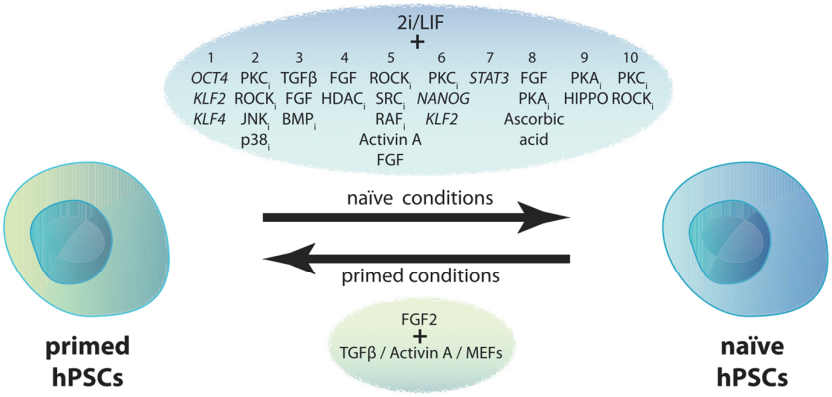
**Table 2:** Compositions of the media used in this study (adapted from Dakhore et al. (2018))

| Medium | Components | Extracellular matrix | Reference |
|---|---|---|---|
| mTeSR™1 | DMEM/F12, bFGF, insulin, transferrin, selenium, l-ascorbic acid, TGFß, BSA, cholesterol, lipids, pipecolic acid, GABA, ß-mercaptoethanol, Glutathione, trace elements, Lithium Chloride, $NaHCO_3$, x | Corning® Matrigel® | Ludwig et al. 2006, Navara et al. 2018 |
| E8 | DMEM/F12, bFGF, insulin, transferrin, selenium, l-ascorbic acid, TGFß (or Nodal), $NaHCO_3$ | Corning® Matrigel® | Chen et al. 2011 |
| KSR-bFGF | DMEM/F12, bFGF, 20% KSR, Glutamax, nonessential amino acids, ß-mercaptoethanol, and 1% penicillin–streptomycin | irradiated mouse embryonic fibroblasts (MEFs) | Krendl et al. 2017 |

### 1.2.2 Ground-state (naïve) and primed pluripotency

ESCs were initially derived from the inner cell mass of pre-implantation blastocysts in mouse. Subsequently, pluripotent stem cells were derived also from early post-implantation epiblast, which are termed epiblast stem cells (EpiSC) (Tesar et al., 2007). Like ESCs, EpiSCs also express pluripotency markers, have both activating and repressing (bivalent) histone marks on the developmental genes, and can differentiate into all three germ layers and to the germline (Theunissen et al., 2014).

Nevertheless, ESCs and EpiSCs differ in certain aspects; hence, their states are described as ground-state (naïve) and primed pluripotency, the latter corresponding to a later stage in development (Nichols and Smith, 2009). It was noted that cultured hESCs were indeed more similar to mouse EpiSCs (mEpiSCs) than mESCs, for example both depend on FGF2/Activin signaling (Weinberger et al., 2016). Primed pluripotent stem cells have an inactive X chromosome, use the proximal *OCT4* enhancer, have lower germline contribution and higher multilineage differentiation potential, express primed pluripotency markers (e.g. *DNMT3A*, *DNMT3B*) and early differentiation genes, and have a glycolytic metabolism (**Figure 8**). On the other hand, presence of two active X chromosomes, open chromatin (H3K27 hypomethylation), usage of distal enhancer of *OCT4*, higher clonogenicity, expression of naïve markers (e.g. *NANOG*, *REX1*/*ZFP42*), and a metabolism dependent on oxidative phosphorylation (instead of glycolytic) are some characteristics of naïve pluripotency (Lee et al., 2017; Ware, 2017; Warrier et al., 2017).




| | | |
|---|---|---|
| *DNMT3B, CD24, OTX2/ZIC2* | Marker genes | *TBX3, DNMT3L, NANOG, REX1/ZFP42* |
| CD24, CD57, CD90 | Cell surface markers | CD7, CD75, CD77, CD130 |
| proximal | *OCT4* enhancer usage | distal |
| high | DNA methylation | low |
| high | H3K27me3 on polycomb-associated genes | low |
| $X_aX_i$ | X chromosome inactivation | $X_aX_a$ |
| Glycolysis | Energy metabolism | Oxidative phosphorylation |
| yes | ERK pathway dependence | no |
| low | Clonogenicity | high |

**Figure 8: Primed vs ground-state (naïve) pluripotency** (adapted from Ware (2017) & Yilmaz and Benvenisty (2019))

Naïve and primed states of pluripotency can be interconverted by tuning the culture conditions. Traditionally, mouse naïve cells were maintained by including leukemia inhibitory factor (LIF), and inhibitors of mitogen-activated protein kinase (MEK) and glycogen synthase kinase 3 beta (GSK3β) in the medium (i.e. 2i) (Silva and Smith, 2008; Ying et al., 2008). The concentration of the GSK3β inhibitor

CHIR99021 should be fine-tuned, because it maintains ground-state pluripotency at low concentrations, while it causes differentiation at higher concentrations (Takashima et al., 2014). Likewise, several groups published protocols for conversion of hPSCs to ground-state pluripotency (**Figure 8**), most of which share the basic 2i/LIF condition, as reviewed by Ware (2017) and Yilmaz and Benvenisty (2019) in detail, though a universal recipe is not established yet.

### 1.2.3 Early lineage commitment

In the early embryo, the development of the blastocyst proceeds with the alignment of the cells of the inner cell mass in the form of two adherent discs; epiblast and hypoblast, which are the precursors of embryonic and extraembryonic tissues, respectively. Subsequent fates of PSCs are determined by different combinations of Activin/Nodal, BMP, FGF, and Wnt signaling gradients (Vallier et al., 2009a) (**Figure 9**).

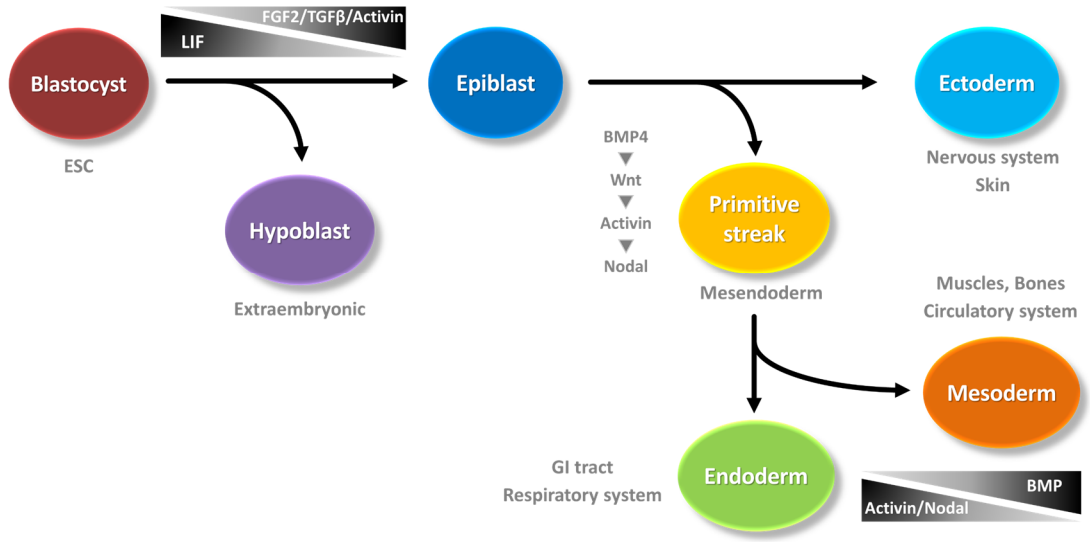



**Figure 9: Lineage bifurcations during early embryonic development**. Important signaling pathways are indicated

Initially, BMP4 signaling activates the Wnt pathway, which in turn induces the formation of the primitive streak, a structure extending from the middle of the embryo towards the posterior end. Subsequently, the cells along the streak start to migrate inwards, which is called gastrulation (**Figure 10**). Wnt-induced activation of the Activin/Nodal signaling results in the emergence of mesendoderm cells at this stage (Ben-Haim et al., 2006). Positive regulation of the Activin/Nodal signaling gives rise to the endoderm (Yiangou et al., 2018), a second layer formed by the migrating cells that align along the hypoblast, which gives rise to the organs or tissues that are related to the inner body cavity, such as gastrointestinal tract and respiratory system. On the other hand, negative regulation of the Activin/Nodal

signaling in the migrating cells that remain in between the epiblast and endoderm results in the mesoderm layer, which is the source of the tissues like muscles, bones, and the circulatory system. Formation of only one primitive streak in the embryo, and its exclusion from the anterior part is ensured by the Nodal-antagonizing signals secreted from the hypoblast (Perea-Gomez et al., 2002). The cells that remain in the epiblast layer form the neuroectoderm, which gives rise to tissues such as skin or the nervous system. Neuroectoderm is the default layer formed upon the clearance of anti-differentiation signals in the absence of other lineage-specifying signals, while endoderm and mesoderm are actively induced (Yiangou et al., 2018). There are several genes that are used as markers (some of which are master lineage regulators) of specific stages during gastrulation; such as *T* (*BRACHYURY*) and *MIXL1* for primitive streak; *EOMES* and *GSC* for mesendoderm; *T* for mesoderm; *CER1*, *FOXA2*, *GSC* and *SOX17* for endoderm (Faial et al., 2015); and *SOX2* for neuroectoderm (Vallier et al., 2009b).

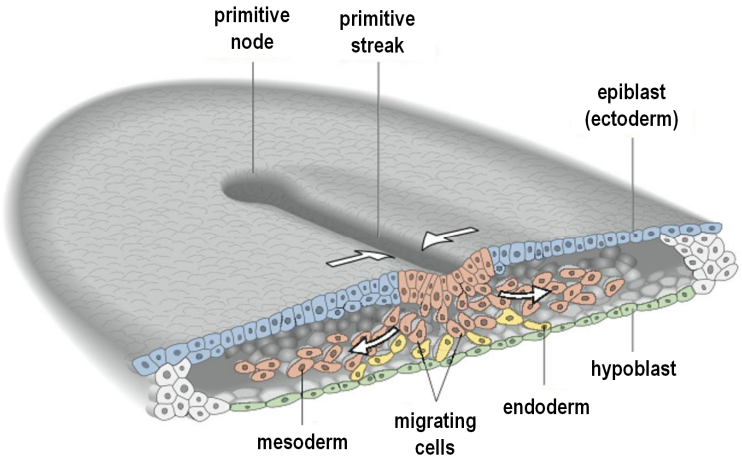



**Figure 10: Gastrulation**[8]

The canonical Wnt pathway (Wnt/ß-catenin pathway) has a crucial function in gastrulation, which is thought to exert its effect mainly via stabilization of ß-catenin (Doble and Woodgett, 2003). In the absence of Wnt ligand, a destruction complex in the cytoplasm constantly phosphorylates and ubiquitinates ß-catenin for degradation (**Figure 11**, left). In the presence of the Wnt ligand the ß-catenin destruction complex is sequestered to the cell membrane, and ubiquitination of ß-catenin is inhibited, leading to the saturation of the destruction complex and translocation of the excess ß-catenin to the nucleus (Li et al., 2012) (**Figure 11**, right). In the nucleus, ß-catenin transactivates the TCF/LEF family of transcription factors, which are bound to Wnt target genes (Doble and Woodgett, 2003). The Wnt pathway can be stimulated using isolated Wnt3a protein, chemically inhibiting

[8] http://www.mun.ca/biology/desmid/brian/BIOL3530/DEVO_03/devo_03.html

GSK3ß (e.g. using CHIR99021) which blocks the degradation of ß-catenin (Blauwkamp et al., 2012), or overexpressing ß-catenin (**Figure 46**).

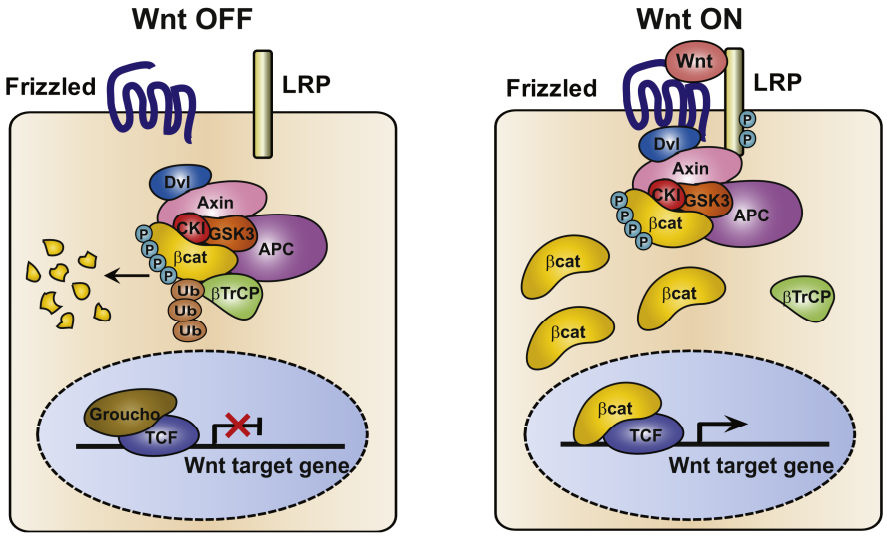



**Figure 11: Wnt/β-catenin pathway in the presence and absence of the Wnt ligand** (Li et al., 2012)

## 1.3 Barcode Assembly for Targeted Sequencing (BART-Seq)

The previous sections portrayed the history of transcriptomics and the advances in the field of single-cell sequencing. As a summary, the techniques for analyzing bulk samples that provide secondary signals as output (e.g. microarrays or qPCR) are not sequence-sensitive, and they do not allow high degree of sample multiplexing. Targeted approaches often aim to analyze a large number of loci in a small number of samples, require a high amount of starting material and provide poor dynamic readouts due to the intermittent purification steps, rendering them incompatible for single-cell analysis. The major challenge for the current single-cell transcriptomic approaches is the poor detection of the genes expressed in moderate to low levels, which are in fact majority of the genes in a cell. Since they attempt to sample all the transcripts, a few highly expressed (mostly housekeeping) genes consume most of the reads, hindering any mechanistic understanding of the expression patterns of the genes of special interest. Second, commercial methods usually require expensive instrumentation and consumables, which make them unaffordable for many research laboratories. Third, many of them are tag-based, meaning that they count only the 3' (or 5') of the transcripts, making the analysis of different isoforms or non-poly(A) transcripts impossible. The methods that can analyze full-length transcripts, such as Smart-Seq2 (Picelli et al., 2013), and the non-NGS methods such as FISSEQ (Lee et al., 2015) or MERFISH (Chen et al., 2015) are costly and often not high-throughput.

### 1.3.1 A novel target enrichment and barcoding workflow

With the purpose of addressing the aforementioned problems, I established a workflow for enriching the selected sets of loci from gDNA or cDNA samples (**Figure 12**). Preliminary work in the Drukker Lab had established the general concept for synthesizing differentially barcoded forward and reverse primers for target enrichment by multiplex PCR and sample indexing for NGS. The novel method is named Barcode Assembly for Targeted Sequencing (BART-Seq). It was envisioned that the PCR enrichment would enable focusing the sequencing capacity to the selected targets, thus provide high-resolution information. The simple workflow that circumvents intermittent steps of fragmentation, hybridization, or ligation would make quantitative analysis possible for both bulk samples and sorted single cells. The simple synthesis reaction and combinatorial indexing would allow massive multiplexing using only a small number of barcodes; therefore, reducing the costs substantially.

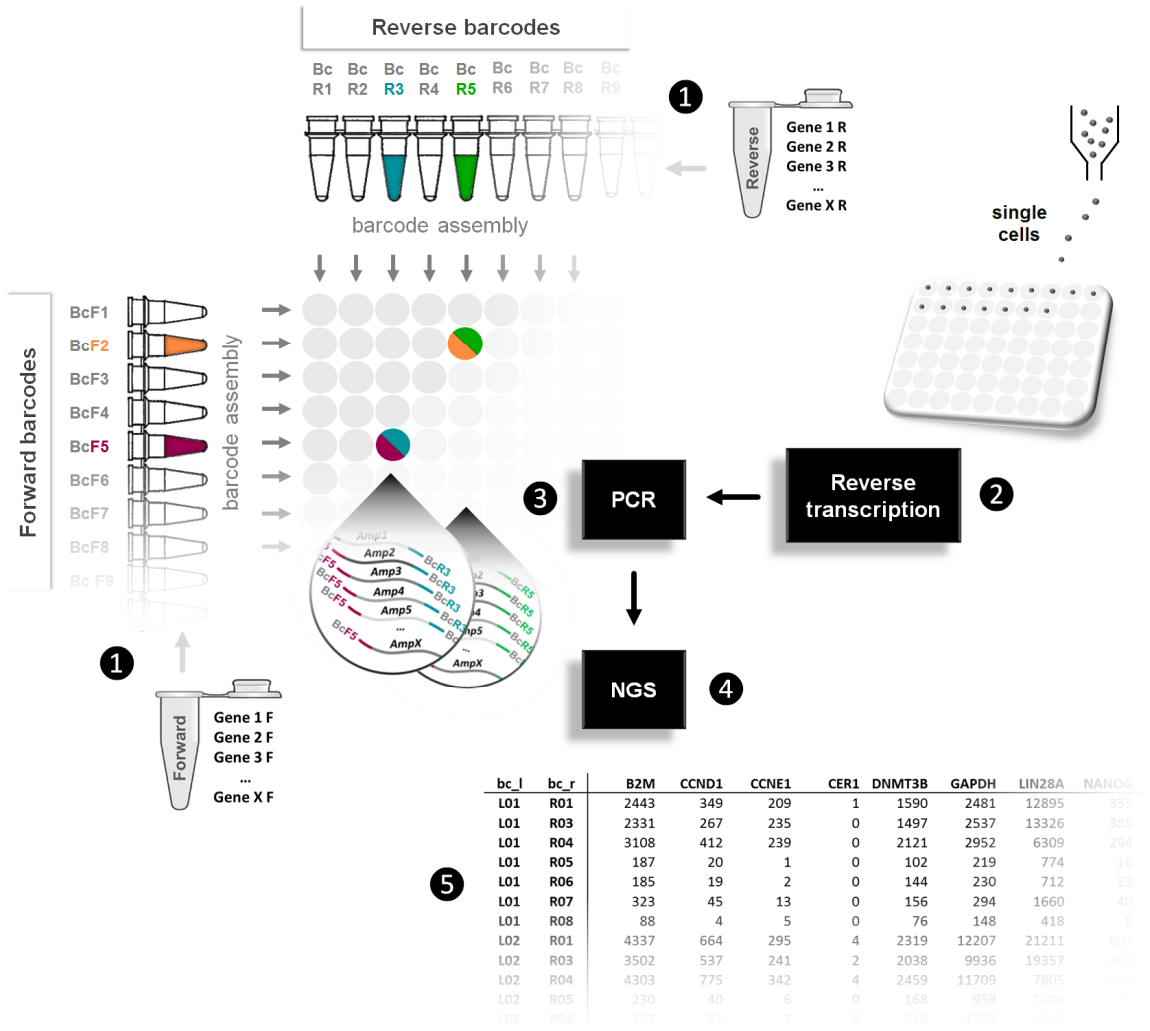



**Figure 12: Barcode Assembly for Targeted Sequencing (BART-Seq) workflow.** ❶ Invariant sets of multiplexed primers (*Gene1-GeneX*) are differentially indexed using panels of forward (BcF) and reverse (BcR) DNA barcodes. ❷ (c)DNA templates from bulk samples or single cells are prepared. ❸ Amplicons with dual barcodes are generated using combinations of barcoded-primer sets during the PreAmplification PCR. ❹ PCR products are pooled and sequenced in a paired-end NGS run. ❺ Sequencing reads are demultiplexed to count matrices for further analyses

### 1.3.2 Focus of the thesis

The global goal of my project was to establish a complete workflow based on the BART-Seq method, and then apply it to biological questions. I aimed to implement it for efficient analysis of selected genomic or transcriptomic loci in a large number of samples, including single cells, cost-effectively, and develop the accompanying computational methods to analyze the data. The specific questions of my project were the following:

- Can we increase the efficiency of individual steps of the workflow, and decrease the overall cost?
- Is BART-Seq suitable for sensitive quantification of template mRNAs?
- Can we detect gene expression changes in single cells?
- Does the method suit high-throughput analysis of bulk samples?

To address these, I initially carried out optimization experiments on individual steps of the workflow to increase the efficiency of reactions, decrease the cost per sample, or reduce the total number of steps. Next, I demonstrated that the method is suitable for dynamic range measurements in isolated bulk RNA samples or directly in sorted cells. Subsequently, I adapted it for high-throughput transcriptomics and applied it in a series of experiments including thousands of single cells. The specific questions were: whether different maintenance media influence the pluripotency state of hESCs, and whether activating the Wnt pathway using different stimuli yields the same transcriptional outcomes. I also contributed to two projects where we used the method to screen gDNA samples from patients for mutations or the transcriptional response of hepatocytes to a compound library. Along with the experiments, I developed bioinformatics scripts for processing the data starting from the raw count matrices and ending with biological interpretations, besides assisting the development of bioinformatics tools for de-multiplexing the raw sequencing data to read count matrices. I discuss these endeavors and their outcomes in this thesis.

# 2 MATERIALS & METHODS

## 2.1 Materials

**Table 3:** Reagents and kits

| Reagent | Supplier | Cat. No |
|---|---|---|
| Agencourt AMPure XP beads | Beckman Coulter | A63881 |
| Agilent High Sensitivity DNA Kit | Agilent | 5067-4626 |
| Ambion® ArrayControl™ RNA Spikes | Invitrogen™ | AM1780 |
| DNA Polymerase I large (Klenow) fragment | Invitrogen™ | 18012021 |
| dNTP Set 100 mM Solutions | Thermo Scientific™ | R0181 |
| Ethanol ≥99,8 % | Carl Roth | 9065.2 |
| GeneRuler™ 100bp DNA Ladder | Thermo Scientific™ | SM0241 |
| Lambda Exonuclease | New England Biolabs | M0262L |
| $MgCl_2$ (magnesium chloride) (25 mM) | Thermo Scientific™ | R0971 |
| MiSeq® Reagent Kit v2 (300 cycles) | Illumina | MS-102-2002 |
| NaOAc (pH 5.5) | Ambion | AM9740 |
| NEBNext® ChIP-seq Library Prep Reagent Set for Illumina® | New England Biolabs | E6200S |
| NEBNext® dA-Tailing Module | New England Biolabs | E6053S |
| NEBNext® dA-Tailing Reaction Buffer | New England Biolabs | B6059S |
| NEBNext® End Repair Module | New England Biolabs | E6050S |
| NEBNext® High-Fidelity 2X PCR Master Mix | New England Biolabs | M0541S |
| NEBNext® Multiplex Oligos for Illumina® (Index Primers Set 1) | New England Biolabs | E7335L |
| NextSeq® 500/550 Mid Output Kit v2 (300 cycles) | Illumina | FC-404-2003 |
| Nuclease-Free Water (not DEPC-Treated) | Invitrogen™ | AM9932 |
| $Oligo(dT)_{18}$ Primer | Thermo Scientific™ | SO132 |
| PhiX Control v3 | Illumina | FC-110-3001 |
| Platinum™ Multiplex PCR Master Mix | Applied Biosystems™ | 4464268 |
| Power SYBR™ Green PCR Master Mix | Applied Biosystems™ | 4367659 |
| Primers | Sigma Aldrich | |
| QIAGEN Multiplex PCR Plus Kit | QiaGen | 206152 |
| Quant-iT™ PicoGreen™ dsDNA Assay Kit | Invitrogen™ | P7589 |
| Qubit™ dsDNA HS Assay Kit | Invitrogen™ | Q32854 |
| React®2 Buffer (10X) | Invitrogen™ | 16302-010 |
| RNeasy Mini Kit | QiaGen | 74106 |
| SuperScript ™ III First-Strand Synthesis System | Invitrogen™ | 18080051 |
| SuperScript™ IV First-Strand Synthesis System | Invitrogen™ | 18091200 |
| SYBR™ Safe DNA Gel Stain | Invitrogen™ | S33102 |
| T7 Exonuclease | New England Biolabs | M0263S |
| TE Buffer (20X), RNase-free | Invitrogen™ | T11493 |

**Table 4:** Cell culture media, supplements, and cell lines

| Reagent | Supplier | Cat. No |
|---|---|---|
| 2-Mercaptoethanol (ß-Mercaptoethanol) (50 mM) | Gibco™ | 31350010 |
| Accutase® solution | Sigma-Aldrich | A6964-100ML |
| B-27™ Supplement, minus insulin | Gibco™ | A1895601 |
| BD FACSFlow™ Sheath Fluid | BD Biosciences | 342003 |
| CHIR 99021 trihydrochloride | Tocris | 4953 |
| Collagenase Type IV, powder | Gibco™ | 17104019 |
| DMEM, high glucose, pyruvate, no glutamine | Gibco™ | 21969035 |
| DMEM/F-12 | Gibco™ | 11320074 |
| Doxycycline hyclate | Sigma-Aldrich | D9891-1G |
| DPBS, no calcium, no magnesium | Gibco™ | 14190094 |
| EDTA disodium salt dihydrate | Carl Roth | X986.1 |
| Fetal Bovine Serum (FBS) | HyClone™ | SH30071.03 |
| FGF2 (Recombinant Human FGF-basic) (154 a.a.) | Peprotech | 100-18B |
| GlutaMAX™ Supplement | Gibco™ | 35050061 |
| Hygromycin B (50 mg/mL) | Gibco™ | 10687010 |
| Insulin-Transferrin-Selenium Supplement (100X) | Gibco™ | 41400045 |
| KnockOut™ Serum Replacement (KSR) | Gibco™ | 10828028 |
| L-Ascorbic Acid 2-Phosphate Magnesium | Sigma-Aldrich | A8960-5G |
| Matrigel® Growth Factor Reduced (GFR) | Corning® | 354230 |
| MEM Non-Essential Amino Acids Solution (100X) | Gibco™ | 11140050 |
| mTeSR™1 | Stemcell Technologies | 5850 |
| Penicillin-Streptomycin (10,000 U/mL) | Gibco™ | 15140122 |
| Propidium iodide | Sigma-Aldrich | P4170-10MG |
| Recombinant Human TGF-ß1 (HEK293 derived) | Peprotech | 100-21 |
| RPMI Medium 1640 - L-Glutamine | Gibco™ | 21875034 |
| Sodium bicarbonate ($NaHCO_3$) | Sigma-Aldrich | S5761-1KG |
| Trypsin-EDTA (0.25%), phenol red | Gibco | 25200056 |
| Y-27632 dihydrochloride | Tocris | 1254/10 |
| CD1-irradiated mouse embryonic fibroblasts | Drukker Lab | |
| H9 (WA09) hESCs line | WiCell | |
| H9 hESC line modified with dox-inducible ß-cateninΔN90 | Drukker Lab | |
| HMGU#1 human iPSC line (Kunze et al., 2018) | Drukker Lab | |
| Newborn human BJ fibroblasts (ATCC® CRL-2522™) | ATCC® | |
| rWnt3a (gift from Derk ten Berge) | Erasmus MC, Rotterdam | |

**Table 5:** Consumables

| Consumable | Supplier | Cat. No |
|---|---|---|
| 384-well PCR plates | Kisker Biotech | G034-ABI |
| PCR Plate, 96-well, non-skirted | Thermo Scientific™ | AB0600 |
| Sapphire PCR 8-tube sptrips, 0,2 mL, PP, natural | Grenier BioOne | 673210 |
| Conical Tubes (50 mL) | Invitrogen™ | AM12502 |
| DNA LoBind Tubes, 5 mL | Eppendorf | 30122348 |
| DNA LoBind Tubes, 2.0 / 1.5 mL | Eppendorf | 301080** |
| Sealing Tape Aluminium Foil | Starlab | E2796-9792 |
| MicroAmp™ Optical Adhesive Film | Applied Biosystems™ | 4311971 |
| Falcon® 5 mL Tubes with Cell Strainer Cap | Corning® | 352235 |
| Nunc™ Cell-Culture 6-well plates | Thermo Scientific™ | 140685 |
| 10/20 µl XL TipOne® filter tips | Starlab | S1120-3810-C |
| 200 µl TipOne® filter tips | Starlab | S1120-8810 |
| 1000 µl TipOne® filter tips | Starlab | S1126-7810-C |

## 2.2 Instruments

**Table 6:** Instruments and equipment

| Instrument | Company | Cat. No |
|---|---|---|
| +4 freezer, -20 freezer | Liebherr | |
| 10 / 20 / 200 / 1000 µl pipettes, single-channel | Eppendorf | 31210000** |
| 12-channel pipette, 10 µl | Sartorius | 725220 |
| 1-channel electronic pipette, 5-120 µl | Sartorius | 735041 |
| 8-channel electronic pipette, 10-300 µl | Sartorius | 735361 |
| Agilent 2100 Bioanalyzer | Agilent | G2939BA |
| BD FACSAria™ III | BD Biosciences | |
| Heracell™ 240i $CO_2$ Incubator | Thermo Scientific™ | |
| Heraeus™ Megafuge™ 40 Centrifuge Series | Thermo Scientific™ | |
| Heraeus™ Pico™ 21 Microcentrifuge | Thermo Scientific™ | |
| Mastercycler® nexus X2 / nexus eco / nexus gradient | Eppendorf | |
| Microcentrifuge, MiniStar silverline | VWR | 521-2844P |
| MiSeq® System | Illumina | SY-410-1003 |
| NextSeq® 500 System | Illumina | SY-415-1001 |
| Power Supply peqPOWER 300V | Peqlab | |
| QuantStudio 12K Flex Real-Time PCR System | Thermo Fisher | |
| Qubit® 2.0 Fluorometer | Thermo Fisher | |
| Revco™ ExF -86 °C Upright Ultra-Low Temperature Freezer | Thermo Scientific™ | |
| Safe 2020 Class II Biological Safety Cabinet | Thermo Scientific™ | |
| Safire II Microplate Reader | Tecan | |
| Thermomixer | Eppendorf | 5355 000.011 |

## 2.3 Computational Tools

**Table 7:** Software

| Software | Address |
|---|---|
| R (v3.5.2) | https://www.r-project.org/ |
| CLC Genomics Workbench (v8.5.1) | https://www.qiagenbioinformatics.com/products/clc-genomics-workbench/ |
| FastQC tool (v0.11.8) | https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ |

**Table 8:** Websites

| Website | Address |
|---|---|
| Ensembl Genome Browser | https://www.ensembl.org/index.html |
| Illumina BaseSpace | https://basespace.illumina.com |
| NCBI-GEO | https://www.ncbi.nlm.nih.gov/geo/ |
| Predict a 2° Structure Web Server | https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html |
| PrimerSelect | http://icb-bar.helmholtz-muenchen.de/primerselect |
| UCSC *In-Silico* PCR | https://genome.ucsc.edu/cgi-bin/hgPcr |

## 2.4 Methods

### 2.4.1 Design of barcode panels

All possible 8-mer (barcode) and 10-mer (linker) oligonucleotides of 50-60% GC content were computed omitting sequences with one, two, or three nucleotide repeats. All pairwise global alignment scores were computed separately for barcodes and linkers using pariwise2 from Biopython package. Whenever comparing two barcodes in all forward and reverse combinations, the maximal alignment scores were used for further analysis. Next, a global optimization heuristic (simulated annealing) was implemented to efficiently identify a set of highly unique sequences in terms of likelihood that mutations (exchange, deletion, insertion) might lead to a conversion into another sequence within the set. A random initial set of sequences was either shrunk (with 10% probability), altered by randomly exchanging sequences (36% probability), or randomly increased (54% probability). Changes were accepted if the new sum of alignment scores was lower or by change whenever exp(-Δsumscore÷T) was lower than another random number. This simulated annealing algorithm scanned temperatures T from 10,000 to 0 along 300 cooling iterations to reach a global optimum. The resulting sets were randomly divided into forward and reverse barcodes and linkers. Next, the 3' of the forward and reverse linkers were ligated *in silico* to the sequences of the forward and reverse barcode sets, respectively. Finally, BLAST was used to accept 18 nt sequences without any identified hit in the human genome (**APPENDIX G**) and transcriptome (**APPENDIX F**). (Description adapted from Dr. Nikola Müller)

### 2.4.2 Primer design and optimization

Primers were designed to have an adenine (A) base at the 3' position of the final primer sequence after barcode assembly. This was done due to the fact that the DNA Polymerase I large (Klenow) fragment frequently adds a template-independent A base to the 3' of the newly synthesized strand. Primer3 was used with default settings, but with modified internal primer predictions such that it enforces the primer's 3' to end with a T nucleotide. For each template, up to five forward and reverse primer pairs were predicted. Certain regions can be excluded or included just like the Primer3, e.g. using "< >" or "{ }". Each primer pair set was compared against the human genome using the blastn command from the blast+ package with the parameters -reward 1 -gapopen 5 -gapextend 5. Using our web-based software, the user can set the number of hits allowed for further processing. Next, given the predefined linkers, and 1-5 predicted primer pairs per loci, an *in silico* ligation step was performed to generate all possible primer-linker combinations. Hereby, matching forward and reverse primers defined one amplicon. To minimize the probability of forming stable dimers, we calculated the all-against-all minimal free energy (including all reverse complements) using the RNAcofold command from the ViennaRNA package version 2.1.8 with the parameters --noPS --noLP -P dna_mathews2004.par. Low predicted minimum free energy correlates to a high

probability of forming a stable dimer. A simulated annealing was implemented to identify optimal combinations of each primer pair per locus, thereby taking barcode and linker sequences into account. During optimization, the minimal value of free energy of the forward or reverse complement sequence was used for determining the probability of forming stable primer dimers. Per amplicon and gene, we started with a random initial set of primers. We proceeded to either randomly alter it (with 80% probability) or randomly exchanged amplicons if there were several amplicons available for a gene. In each step, the random change was accepted if the new sum of minimum free energies (mfe) is lower than in the last or randomly if exp(-Δmfe÷T) was lower than a uniformly drawn random number. We scanned over temperatures T from 15,000 to 0 along 500 cooling iterations. The primer prediction implementation is a Python-based web front end that is available online at: http://icb-bar.helmholtz-muenchen.de, of which we made the code freely available (see **2.5 Availability of Data and Materials**). (Description quoted from Dr. Nikola Müller and Philipp Angerer)

### 2.4.3 Design of primer sets

An amplicon size range of 75-248 nt was aimed to ensure detection by 2×150 bp paired-end sequencing. Primer sets targeting 10 specific mutations in *BRCA1* and *BRCA2* genes (Kaufman et al., 2006; Laitman et al., 2012; Lerer et al., 1998) were designed based on the human genome reference hg19 (**APPENDIX L**). Pluripotency primer set was designed based on the analysis of publicly available RNA-Seq datasets of hESCs via NCBI-GEO from H9, H7, and HD291 cells (GSM602289, GSM1163070, GSM1163071, GSM1163072, GSM1704789, GSM1273672, GSM1327339), and own datasets (**APPENDIX J**). The target regions were selected for Wnt stimulation (mesoderm) primer set using bulk RNA-Seq data produced by stimulation of hESCs by rWnt3a or CHIR99021 for 72 h (**APPENDIX K**). RNA-Seq reads were mapped to the genome reference hg38 using CLC Genomics Workbench using mismatch cost: 2, insertion cost: 3, and deletion cost: 3. To find the highest expressed loci of the genes regardless of transcripts, initially the ratio of reads per transcript variant to its length was calculated and summed up to obtain an average for the gene. This sum was multiplied by ~20 (empirical), which is used as the lower threshold to create coverage map of the gene in a particular sample. The regions higher than this threshold were marked. After repeating this for each sample, the loci overlapping in the majority of the samples were selected for primer design. This was repeated for each gene. The complete sequences of RNA spike-ins EC2 (RNA1), EC12 (RNA2), EC13 (RNA6), and EC5 (RNA8) were used.

**Table 9:** Derivation of reverse complementary (rc) primers from nested primers

| Process | Sequence |
|---|---|
| Sequencing (nested) primer | 5'CTGCCGTGTGAACCATGTG**A** 3' |
| Add linker to 5' & remove 3'A | 5'ATGCGCATTCCTGCCGTGTGAACCATGTG 3' |
| Reverse complement & add 5'[phos] | 5'[phos]CACATGGTTCACACGGCAGGAATGCGCAT 3' |

### 2.4.4 Cell culture

Undifferentiated hESCs (H9 line) were routinely maintained on Matrigel™-coated plates in mTeSR™1 medium in 5% (v/v) $O_2$. Cells were passaged as clumps using 2 mg/ml solution of Collagenase Type IV prepared in DMEM/F-12.

#### 2.4.4.1 Growth media comparison

Cells were split and maintained for five passages in mTeSR™1, E8 (on Matrigel™), and KSR-bFGF media (on CD1-irradiated mouse embryonic fibroblasts) in parallel. E8 medium consisted of DMEM/F12 supplemented with 64 mg/l l-ascorbic acid-2-phosphate magnesium, 14 µg/l sodium selenium, 100 µg/l FGF2, 19.4 mg/l insulin, 543 mg/l $NaHCO_3$ and 10.7 mg/l transferrin, 2 µg/l TGFß1, and osmolarity was adjusted to 340 mOsm at pH 7.4, as described by Chen et al. (2011). KSR-bFGF media consisted of DMEM/F12 supplemented with 20% KSR, Glutamax, nonessential amino acids, ß-mercaptoethanol, 10 ng/mL FGF2, and 1% penicillin-streptomycin as described by Krendl et al. (2017). Newborn human BJ fibroblasts were cultured in DMEM high glucose, supplemented with 1% GlutaMAX, NEAA, and 10% HyClone™ Fetal Bovine Serum.

#### 2.4.4.2 Wnt/β-catenin pathway activation

hESCs and hESC line modified with doxycycline-inducible ß-catenin (constitutively active form ΔN90) were maintained on Matrigel™-coated plates in mTeSR™1 medium, with 25 µg/ml Hygromycin B in the case of ß-cateninΔN90 line. For time course stimulations, the cells were dissociated to single-cell suspension with Accutase and seeded into 12-well plates at $2.5\times10^5$ cells per well in the presence of 10 µM Y-27632. The next day, the medium was changed to RPMI-1640 with L-glutamine supplemented with 1x non-essential amino acids and 1x B27 supplement without insulin. Ligands were as follows: 10 µM CHIR99021 and 240 ng/ml recombinant Wnt3a. ß-catenin expression was induced by adding 1 µg/ml doxycycline. The medium and ligands were freshly re-added every 24 h.

### 2.4.5 Single-cell sorting and cDNA synthesis

#### 2.4.5.1 Sorting

hESCs were dissociated using Accutase, and cells maintained in KSR-bFGF on MEFs were collected as clumps using Collagenase Type IV prior to Accutase treatment. Newborn human BJ fibroblasts were dissociated using Trypsin-EDTA 0.25%. For sorting, the cells were resuspended in 1 ml of FACS buffer (4% FBS and 5 µM EDTA in PBS), filtered through a 0.2 µm nylon mesh, and single live cells (propidium iodide negative) were sorted into the 384-well plates (1-32 cells for medium comparison, and single cells for Wnt pathway activation) pre-filled with 2 µl reverse transcription mixture, using Aria III sorter.

#### 2.4.5.2 cDNA synthesis

Reverse transcription mixture (RT mix) was prepared using SuperScript ™ III First-Strand Synthesis System with reverse transcriptase at a final concentration of 2.5 U/µl (nuclease-free water) and oligo-dT primers (2.5 µM). The percentage of reads each spike-in would possibly receive was calculated based on the previous NGS and qPCR data, and the corresponding amounts were combined with the RT mix, which was then aliquoted into individual wells of 384-well plates (2 µl/well). The cells were sorted (often with FACS) directly into the RT mix, plates were sealed with adhesive foils, placed immediately on dry ice for 2 min, and stored at 20 °C. Plates were thawed at room temperature, and the reverse transcription was performed using the thermocycler program: 50 °C for 50 min and 85 °C for 5 min; RNaseH was not used.

#### 2.4.5.3 Bulk RNA isolation

Total RNA was extracted using RNeasy Mini Kit according to manufacturer's instructions.

#### 2.4.5.4 RNA spike-ins

Molecular counts of four RNA spike-ins EC2 (RNA1), EC12 (RNA2), EC13 (RNA6), and EC5 (RNA8) in the stock solutions were calculated based on molecular weights and known concentrations (100 ng/µl). They were serially diluted initially to obtain 5 million/µl (1 million/0.2 µl for conventional reasons), out of which 10-fold serial dilutions down to 100 molecule/0.2 µl were prepared. They were aliquoted into PCR stripes and kept at -80 °C freezer, and thawed on ice before use.

### 2.4.6 Barcode assembly

#### 2.4.6.1 Klenow fill-in reaction

Unit reaction mixture was prepared in nuclease-free water by combining 1x React®2 Buffer, 0.267 mM dNTPs, 2.5 µM multiplexed rc-primer mix, 2.5 µM barcode, and 0.0167 U/µl DNA Polymerase I large (Klenow) fragment. The reaction was incubated at 25 °C for 1 h. Individual rc-primers were used at a 0.25 µM final concentration, and barcode concentrations were matched to the total concentration of rc-primers (incubation time of 2 h was also applicable). The enzyme was heat inactivated at 80 °C for 10 min.

#### 2.4.6.2 Reverse complementary strand removal by Lambda exonuclease

Products of the fill-in reaction were directly diluted as 2/3 volume ratio in the Lambda reaction mixture containing 1x reaction buffer and 0.33 U/µl Lambda

exonuclease and incubated at 37 °C for 30 min (incubation time of 1 h was also applicable). The enzyme was heat inactivated at 80 °C for 10 min.

#### 2.4.6.3 Pre-amplification PCR

PCR reactions (10 µl total) consisted of 2.5 µl Platinum® Multiplex PCR Master Mix (0.5x final), 1.8 µl 25 mM $MgCl_2$ (4.5 mM final), 1.5 µl forward Lambda reaction product (non-purified), 1.5 µl reverse Lambda reaction product (non-purified), 2 µl cDNA, and 0.7 µl nuclease-free water (not DEPC-treated). The reaction cycle profile was as follows: initial denaturation at 95 °C for 5 min; 22 cycles of 95 °C for 30 s, 60 °C for 3 min, 72 °C for 60 s; and final extension at 68 °C for 10 min. Unit PCR reaction of genotyping assays was 20 µl, with the same concentration of reagents, and 18 cycles of PCR. Unit PCR reaction of transcriptomics experiments was 10 µl, with cycle numbers between 16 and 22.

### 2.4.7 qPCR and melting curve analysis

qPCR analyses were carried out using nested primers, which were homologous to the barcode-assembled primers (**Figure 18**), excluding the barcode and the linker regions (**Appendices I-L**). Unit reaction (10 µl total) consisted of 5 µl (1x final) Power SYBR™ Green PCR Master Mix, 1 µl pre-amplification PCR product, 1 µl mixture of forward and reverse nested primers (each 0.2 µM final), and 3 µl nuclease-free water (not DEPC-treated). The reaction cycle profile was as follows: initial denaturation at 95 °C for 10 min followed by 35-40 cycles of 95 °C for 15 s and 60 °C for 1 min. Melting curve analysis was done by heating the amplicons from 60 to 95 °C, incrementing 0.05 °C/s. All the reactions were run as two or three replicates. For statistics, two tailed and paired Student's t-test was used. P values were indicated on the graphs according to following: $P>0.5$ ns; $P\leq0.05$ *; $P<0.01$ **; $P<0.001$***; $P<0.0001$****.

### 2.4.8 Next-generation sequencing

#### 2.4.8.1 Sample pooling and purification

PCR products were pooled in nuclease-free falcon tubes, mixed with 0.1 volume 3 M NaOAc (pH 5.5) and 2.5 volume 100% ethanol (molecular biology grade), and kept at 20 °C overnight for precipitation. Samples were centrifuged at 4000 g for 30 min in a centrifuge pre-cooled to 4 °C. The supernatant was discarded, and the samples were washed once with 500 µl ice-cold 70% ethanol. Tubes were centrifuged at 4000 g for 2 min (4 °C), and the remaining supernatant was pipetted out. The pellet was air dried for 2-3 min and re-suspended in 200-500 µl nuclease-free water. Prior to library preparation, double-sided size selection was performed using Agencourt AMPure XP beads. 0.5x and 1.5x bead to DNA ratio was used for upper and lower size limits, respectively.

### 2.4.8.2 RNA-Seq library preparation and sequencing

For library preparation 50% to 25% of the ethanol precipitated libraries were used, and the rest was kept as backup at -20 °C. Libraries were prepared using NEBNext® Multiplex Oligos for Illumina®, and the protocol was based on NEBNext® ChIP-Seq Library Prep Master Mix Set for Illumina® with the following modifications: end repair was performed using 1 µl NEBNext End Repair Enzyme Mix in 50 µl final reaction. PCR enrichment included 1 µl index and 1 µl universal primers in 50 µl final reaction. The enrichment PCR cycle profile was as follows: initial denaturation at 98 °C for 30 s; 10-15 cycles of 98 °C for 10 s, 65 °C for 30 s, 72 °C for 30 s; and final extension at 72 °C for 5 min. Fifteen, 12, 10 and 15 cycles of PCR enrichment was applied for RNA quantification, media comparison, Wnt pathway stimulation, and *BRCA* genotyping experiments, respectively. Beads to DNA ratios for purification steps using AMPure XP beads were adjusted based on the expected size range of each library after each step of the library preparation (**Figure 13**). The original size range was increased by 1, 65, and 122 after end repair, adapter ligation, and PCR enrichment, respectively.

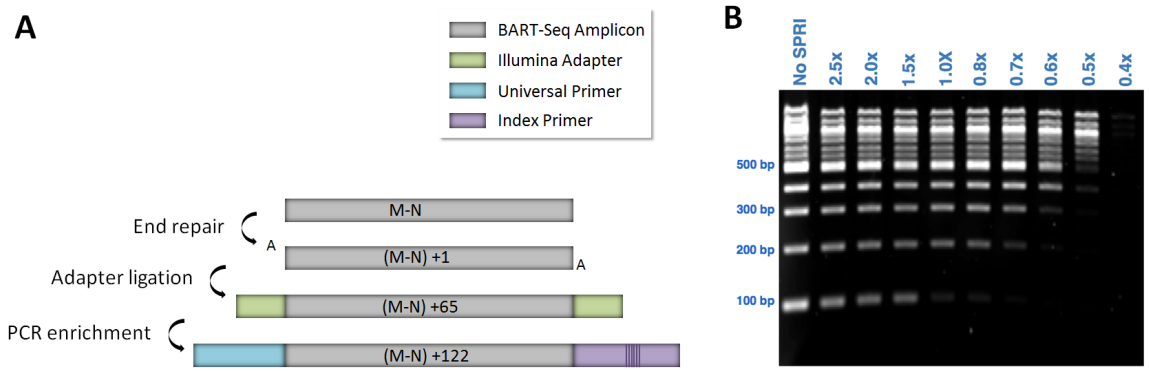



**Figure 13: Determination of the size selection thresholds during library preparation. (A)** Library sizes after each reaction was calculated relative to the original range (M-N). **(B)** Beads:DNA ratio was adjusted based on the library size at each step, referring to example analyses[9]

Final libraries were evaluated using Agilent 2100 Bioanalyzer by High Sensitivity DNA Kit and quantified using Qubit® 2.0 Fluorometer by Qubit® dsDNA HS Assay Kit, and by Safire II Microplate Reader using Quant-iT™ PicoGreen™ dsDNA Assay Kit. Libraries were sequenced (paired-end) on Illumina MiSeq using MiSeq® Reagent Kit v2 (300 cycles) or Illumina NextSeq 500 using NSQ® 500/550 Mid Output Kit v2 (300 cycles). Approximately ten percent PhiX control was included in the sequencing runs as a measure against index switching (Sinha et al., 2017).

[9] http://core-genomics.blogspot.com/2012/04/how-do-spri-beads-work.html

### 2.4.9 Demultiplexing of RNA-Seq reads to count matrices

To trace the origins of reads back to the samples, a pipeline that demultiplexed the reads and counted them while accounting for sequencing errors was implemented. FastQC software was used to create quality reports for manual inspection (Andrews, 2010). Given the acceptable quality, Snakemake workflow engine (Köster and Rahmann, 2012) was used for automatic or step-by-step analysis of raw reads, sets of primers, linkers, barcodes, and expected amplicons. This started by trimming the read ends according to quality using Sickle (Joshi and Fass, 2011), then a list of possible single nucleotide-mutated variants per barcode, excluding the ones shared with other barcodes, was created. Using the algorithm of Aho and Corasick (1975), this list efficiently assigned barcodes to all reads while allowing at most one unambiguous mismatch. We also annotate the reads with several boolean criteria for statistical analysis of libraries. This included the information if the read contained only a primer, multiple (or no) barcodes, if the barcode contained a mismatch or if the read contained bases before the protection group. We aligned the longer amplicons to the reads using HISAT2 (Kim et al., 2015). The final step of the pipeline is to summarize the results. Heatmaps for each library were created per amplicon using the forward and reverse barcodes as a coordinate system, and a spreadsheet file containing the aforementioned read statistics as well as count matrices was generated (**Figure 32**). The pipeline was also made available as described in **2.5 Availability of Data and Materials.** (Description quoted from Philipp Angerer)

### 2.4.10 Classification of *BRCA* mutations

To classify the amplicons corresponding to mutations 1-10, we generated read count per patient for both wild-type and mutation alleles (identified by top blast hit per read) and assigned the genotypes to patients based on the ratio of mutation to wild-type reads. We accepted the ratios >20% to call a mutation (due to the high background).

### 2.4.11 Analysis of protection groups

For the analysis of 5' protection groups, we identified barcodes using BLAT (Kent, 2002), a BLAST-like alignment tool, with options -minScore=0 -minIdentity=95 allowing for one base mismatch at most. This was necessary to screen all possible protection groups. For each detected wild-type or mutant allele, we calculated the frequency of 64 trinucleotides for each forward and reverse barcode. Then, summing the frequencies up across all the alleles, we obtained the total frequency of each trinucleotide per barcode.

### 2.4.12 Data correction and normalization

#### 2.4.12.1 Correction of RNA spike-in reads

Two alternative approaches were used for correcting the spike-in reads, first of which was the median method that is simpler and deals with only the severely inefficient combinations. Second method addresses all the combinations and is based on the correction of spike-ins using negative binomial generalized linear modeling. The initial step for both methods was removal of the wells with extreme outlier spike-in reads after inspecting the heatmaps of raw read counts (i.e., if exhibiting hundreds of folds higher/lower reads than the average). Next, samples exhibiting extremely low barcode-gene combinations were removed.

*Median Method:* Per spike-in, two-sided t-test (default parameters, R version 3.5.2) was performed for each barcode against the rest of the barcodes of the same type (i.e., forward or reverse), using the data between the 5th and 95th percentiles for both groups. Barcode-spike-in combinations with P values lower than an empirically set threshold were replaced with the median of the rest of the barcodes.

*GLM Method:* Initially, spike-in values were modeled (glm.nb from MASS package, R version 3.5.2) using the formula below (**Figure 35B, C**), and the samples that deviate from the predictions more than two-fold were flagged as outliers:

$$full\_model: \quad read\ count \sim variable + cells + well.location + forward + reverse + forward{:}variable + reverse{:}variable \tag{1}$$

The explanatory variables of the model were as follows: spike-in ID (*variable*), number of sorted cells to the well (*cells*), location of the well on the plate (side/corner/middle) (*location*), global efficiency of the barcodes (*forward/reverse*), and combination of barcodes and primers (*forward:variable*, *reverse:variable*). Then, the model was re-calculated excluding the flagged values. Wells were completely removed from the original dataset if only one spike-in was left non-flagged. The remaining outliers were replaced by the predictions of the re-calculated model using the same formula.

Next, two basic models were built for the correction:

$$\text{model\#1:} \quad read\ count \sim variable + forward + reverse + forward{:}variable + reverse{:}variable \tag{2}$$

$$\text{model\#0:} \quad read\ count \sim variable + forward + reverse \tag{3}$$

*Correction factors* were calculated by dividing the predictions of the *model#1* to the predictions of the *model#0*. Raw reads were divided by the *correction factors*. A shortened version of the R script containing the basic steps is provided in **APPENDIX E**.

### 2.4.12.2 Normalization of the data

Using the corrected dataset, scaling factors (RNAx) were calculated using spike-ins (left) or spike-ins and genes together (right) as follows:

$$\mathrm{RNA_x} = (2^{\left(\frac{1}{N}\sum_1^N \log_2(\mathrm{spike_n}+1)\right)} - 1)/median \quad (4)$$

or

$$\mathrm{RNA_x} = (2^{\left(\frac{1}{N}\sum_1^N \log_2(\mathrm{gene_n}+1)\right)} - 1)/median \quad (5)$$

Wells were removed if their scaling factor was ten-fold lower or higher than the median of all the factors, to prevent overcorrection. Then, the factors were median-centered via division to preserve the read count magnitudes. Finally, raw read counts of the transcripts were divided by the scaling factors (**Figure 36**). The corresponding script is available at the Github (see **2.5 Availability of Data and Materials**).

### 2.4.12.3 Well filtering in single-cell experiments

Wells sorted with single cells were operationally defined as "empty" if the ratio of the sum of the spike-in reads to the total reads per sample (normalized and log2-transformed) was same or higher than the negative controls (**Figure 14A**). No cells were sorted into negative control wells, yet they also received some reads due to index switching; therefore, they were used as the background (Sinha et al., 2017). Samples representing the wells sorted with multiple cells were filtered based on the calculated one-cell values of the genes. Filtering the samples sorted with two cells or more, i.e. "doublets", was done by placing a threshold estimated based on the two-fold of the median values and bimodal distribution of the sum of the genes (log2-transformed) (**Figure 14B**). Only housekeeping genes were used for filtering fibroblasts.

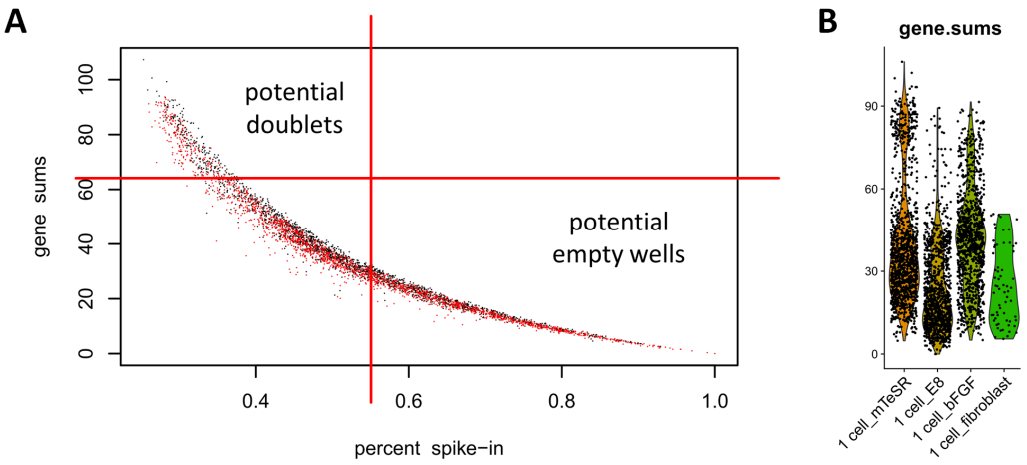



**Figure 14: Plots exemplifying filtering of the samples** that potentially contain **(A)** no cells or **(B)** more than one cell (calculations were made using log2 transformed read counts)

#### 2.4.12.4 Analysis of gene expression

Gene expression analyses were done using custom scripts or Seurat package in R (version 2.3.4), based on normalized and log2-transformed read counts. Linear regression models were calculated using lm function (default parameters, R version 3.5.2).

## 2.5 Availability of Data and Materials

**Data:** The raw and processed BART-Seq data discussed in this thesis is deposited in NCBI's Gene Expression Omnibus (NCBI-GEO) and is accessible under SuperSeries: GSE107723 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107723). Bulk RNA sequencing data used for comparison to 72 h samples (bCat: GSM3737181, GSM3737182; CHIR99021: GSM3737193, GSM3737194; rWnt3a: GSM3737203, GSM3737204) is available under: GSE130381 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130381).

**Codes:** The scripts for designing barcodes and primers and normalizing the read counts are available at https://github.com/theislab/bartSeq, licensed under GNU General Public License. The versions used in this thesis are permanently available under https://doi.org/10.5281/zenodo.3252205. The pipeline for demultiplexing the sequencing reads are available at https://github.com/theislab/bartseq-pipeline, licensed under GNU General Public License v3.0. The version used in this thesis is permanently available under https://doi.org/10.5281/zenodo.3251773. The website for designing the primers is available at http://icb-bar.helmholtz-muenchen.de.

# 3 RESULTS

## 3.1 Development and Optimization of the BART-Seq Workflow

A critical specific aim of my project was to create ways for the BART-Seq method to work effectively with very large cohorts of RNA and gDNA samples, including of single cells, with optimum efficiency and cost. This section provides a summary of the experiments I carried out to optimize the method towards these goals.

### 3.1.1 The principle of barcode-primer assembly

The BART-Seq workflow is built upon serial synthesis of a large number of differentially barcoded forward and reverse primers for target enrichment by multiplex PCR and sample indexing for next-generation sequencing (**Figure 12**, **Figure 15**). The synthesis of primers -named barcode-primer assembly- requires oligonucleotides as the building blocks, DNA Polymerase I large (Klenow) fragment, and Lambda exonuclease (λ-exo). The building blocks are eight-mer DNA barcodes coupled to ten-mer linker sequences, and reverse complementary (rc) primer sets coupled to rc-linkers (**Figure 16**). Different forward and reverse barcode panels and linker sequences are used for the forward and reverse primer sets.

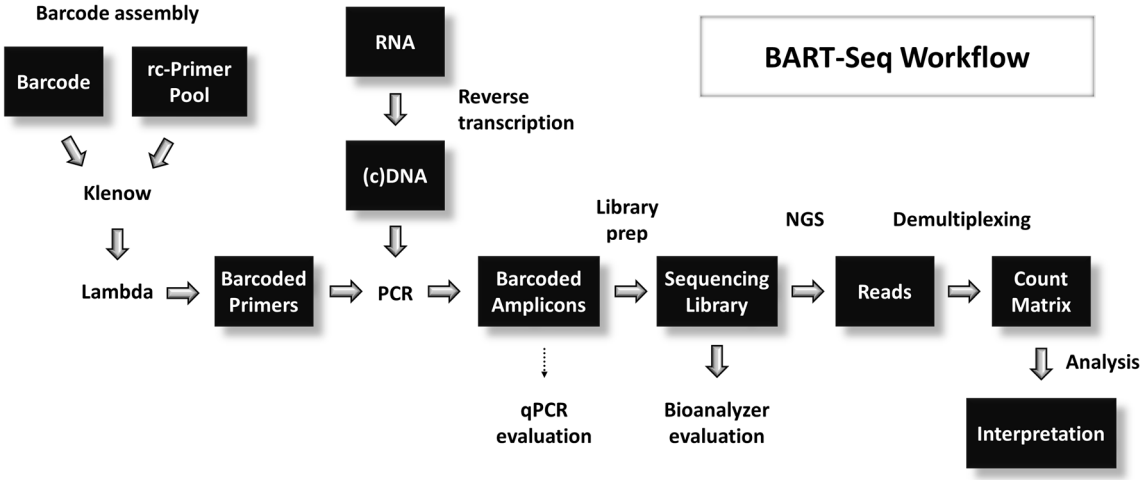



**Figure 15: The complete BART-Seq workflow.** Barcoded primers are synthesized in two steps (Klenow & Lambda). In parallel, cDNA or gDNA samples are prepared. Samples and primers are combined in the PreAmp PCR to generate amplicons with sample-specific dual barcodes, which are pooled and prepared for sequencing. Using a custom-made algorithm, sequencing reads are demultiplexed to count matrices, which are analyzed further to draw biological interpretations.

The steps of the assembly method are as follows: Barcodes and multiplexed rc-primers are hybridized via ten-mer complementary linkers, and double-stranded barcoded primers are synthesized through a bi-directional fill-in by a DNA

polymerase (Klenow fragment). In a second step, anti-sense primer strands are removed by an exonuclease, generating barcoded single-stranded gene-specific primers (**Figure 16**). Intermediate and end products of the barcode assembly were visually confirmed by Agarose gel electrophoresis (**Figure 17**).

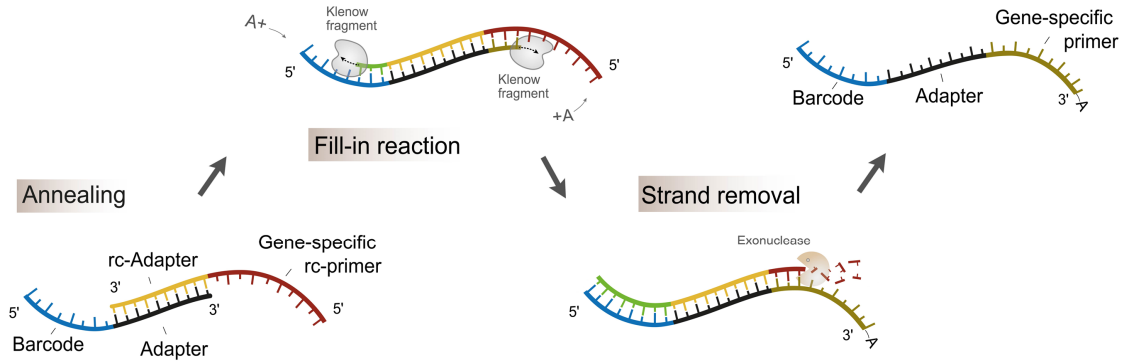



**Figure 16: The basic principle of barcode-primer assembly.** Barcodes and reverse complementary primers are hybridized via complementary linkers, and a fill-in DNA synthesis complements both strands (an A base is frequently added to the 3′ ends by Klenow fragment). Then, an exonuclease is used to remove the anti-sense strand, resulting in single-stranded barcoded primers

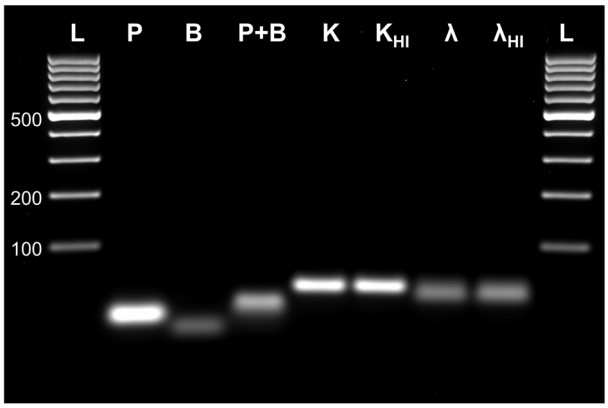



**Figure 17: Intermediate products of barcode assembly visualized by Agarose gel electrophoresis.** Rc-primer (**P**) and barcode (**B**) following hybridization (**P+B**) forms a molecule with higher molecular weight, which is increased further with Klenow fill-in reaction (**K**). λ-exo treatment (**λ**) yields single-stranded barcoded primers with reduced molecular weight. Heat inactivation steps (**K$_{HI}$**, **λ$_{HI}$**) do not influence the products. Samples were a single barcode with a linker and a single rc-primer with an rc-linker, ran on 2.5% Agarose gel with GeneRuler™ 100 bp DNA Ladder. To ensure co-visibility of single and double stranded products, reactions were loaded in different volumes

### 3.1.2 A concept to analyze the efficiency of intermediate reactions by qPCR

It was not feasible in terms of labor and costs to use NGS to evaluate the outcome of each optimization experiment during the development of BART-Seq. With the assumption that the changes in the amount of amplicons generated in the PreAmp PCR should essentially reflect the changes in the efficiency of primer synthesis or

PreAmp PCR, I employed qPCR with nested primers (without barcodes) to compare the amplicon yields of the tested conditions (**Figure 18A**). For this, I compared the control and test conditions side by side for the same target. For statistical significance, I aliquoted the PCR products to multiple reactions and tested several targets in parallel. Ct values were shown as bar plots with standard deviation of two or three technical replicates as error bars. I interpreted the reduced Ct values relative to the control as increased efficiency of the tested condition, and vice versa. Statistical significance of the difference between the control and the tested condition is shown next to the legend (described in the **Materials and Methods**, section **2.4.7**). Unique and well-defined melt curves were interpreted as higher efficiency compared to distorted melt curves with multiple peaks, since they might be indicative of byproducts. Presence of "+" signs on top of each bar reflects the distortion of melt curves, severity of which correlates with the number of signs (one to four) (**Figure 18B**).

For the initial optimization experiments I used a multiplex primer set targeting four and six genomic loci within the human *BRCA1* and *BRCA2* genes, respectively (**APPENDIX I**, **APPENDIX L**). I assembled barcodes with these primers through different reaction conditions and used them to pre-amplify the 10 loci from the bulk gDNA derived from human MCF-7 cell line. I conducted the experiments often using two alternative barcode combinations, one known to be efficient (e.g. A×1 or L14×R05) and one known to be inefficient (e.g. D×9 or L07×R10).

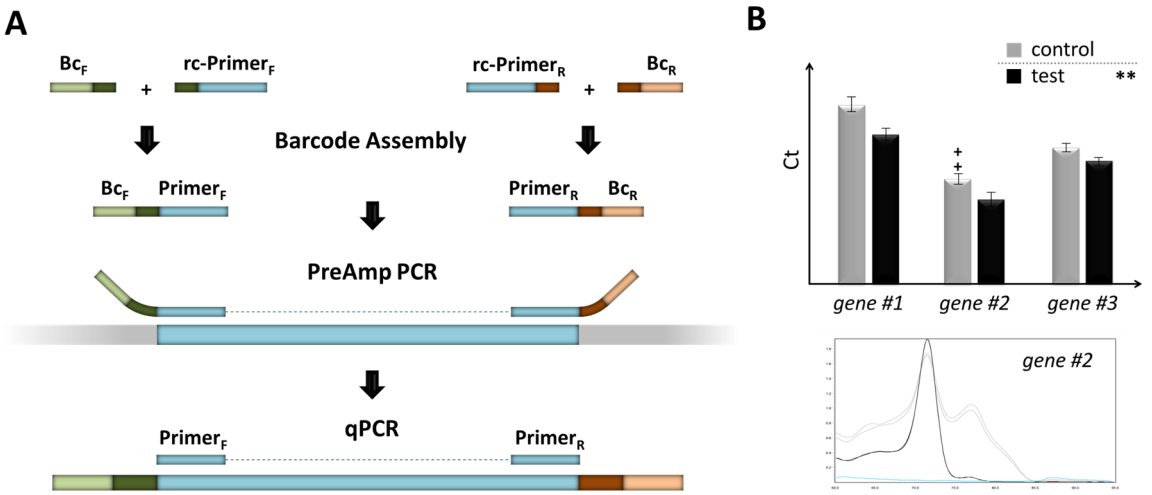



**Figure 18: A concept to assess the intermediate reactions with qPCR. (A)** Amplicons generated using barcoded primers were quantified with qPCR using nested primers (without barcodes), to estimate the efficiency of barcode assembly or PreAmp PCR. **(B)** Ct values of the nested qPCR were plotted with error bars indicating the standard deviation of two/three replicates. Degree of melt curve distortions were indicated with the "+" signs (one to four). Statistical significance of the comparisons was shown next to the legends

### 3.1.3 Barcode assembly

Given that barcode assembly is essentially the step where the primers are synthesized, the success of the whole BART-Seq workflow profoundly depends on the efficient conversion of the oligonucleotide building blocks into barcoded primers (**Figure 15**). Therefore, I first aimed to optimize the assembly steps by comparing reaction components, concentrations, and conditions.

#### 3.1.3.1 Klenow reaction

The first reaction of the barcode assembly is the fill-in synthesis of the hybridized barcodes and rc-primers by a DNA polymerase (**Figure 16**). For this, we used the Large Fragment of DNA Polymerase I (Klenow Fragment) of *E. coli* (Klenow and Henningsen, 1970) because its 3'→5' exonuclease activity allows proofreading, while its lack of 5'→3' exonuclease activity ensures the intactness of the newly synthesized primer duplex, which is crucial for the integrity of the barcode located to the 5' of the sense strand.

##### 3.1.3.1.1 Concentration of oligonucleotides

Initially, I explored the optimum oligonucleotide concentrations for the barcode assembly and PCR, because multiplex PCR has different dynamics than the singleplex one (Henegariu et al., 1997). For this, I screened individual primers in a range of 0.01-1 μM final in the Klenow reaction containing 10 multiplexed primers. Concentration of barcodes was matched to the total primer concentration. Dilutions below 0.25 μM resulted in reduced efficiency of PreAmp PCR, and additional melt curve peaks (**Figure 19A**). Concentrations over 0.25 μM, too, resulted in increased Ct values, and totally abolished specific amplification at 1 μM, possibly resulting from either incomplete primer synthesis or excess primers in the PCR (**Figure 19B**). As a result, I decided to adopt 0.25 μM of each primer as the standard for Klenow reaction, which translates to 0.025 μM in the PreAmp PCR.

##### 3.1.3.1.2 Klenow reaction duration

I tested the influence of the Klenow reaction length on the complete conversion of the oligonucleotide components to barcoded primers. Gradually increasing the duration of Klenow reaction brought the efficiency to a maximum at 60 min (**Figure 20**), possibly because the standard 30 min incubation was not sufficient for full conversion of the oligonucleotides to barcoded primers. Extending the reactions to 120, and 240 minutes did not reduce the Ct values further (not shown), and were not favorable since longer incubation times might result in recessed 3' primer ends due to 3'→5' exonuclease activity of Klenow fragment[10]. Therefore, I decided to adopt 60 minutes as the standard duration for the Klenow step.

[10] https://www.neb.com/~/media/Catalog/All-Products/8A59478F7C464A55A999D8FC33C0AFAE/Datacards%20or%20Manuals/M0210Datasheet-Lot0881209.pdf

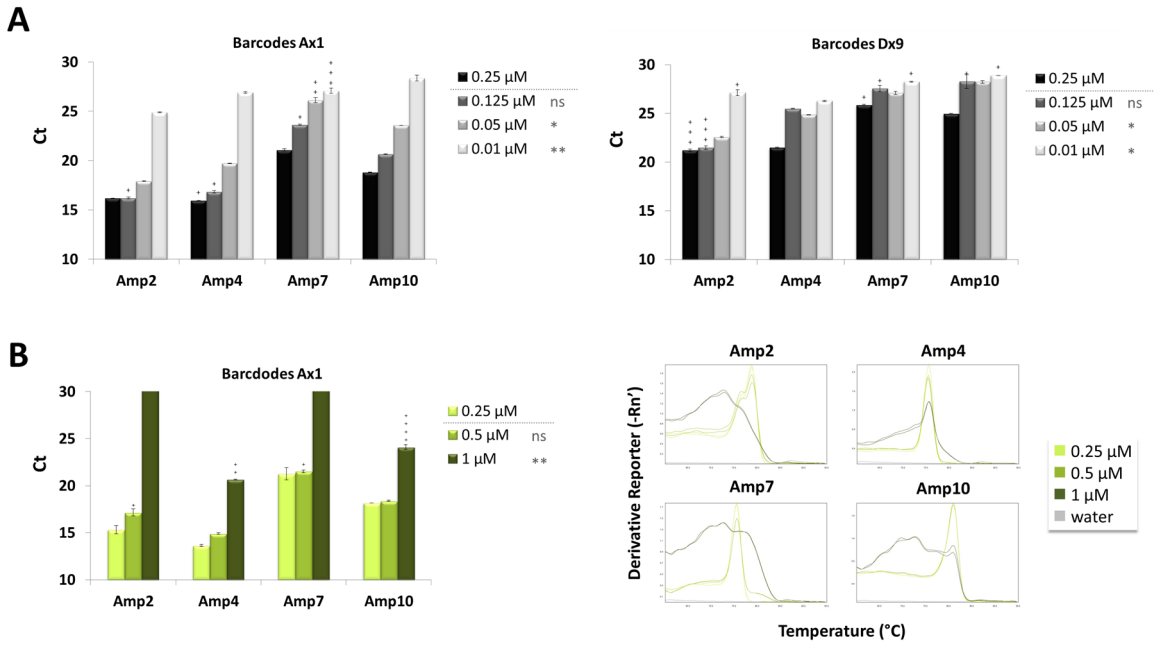



**Figure 19: Optimum oligonucleotide concentrations for the workflow.** Barcode assembly of primers in a range of 0.01 µM to 1 µM (each) final in the Klenow reaction were compared (10 primers were multiplexed). **(A)** Concentrations lower than 0.25 µM resulted in reduced PCR efficiencies, for both barcode combinations tested. **(B)** Concentrations over 0.25 µM resulted in reduced PCR efficiencies, too, and complete loss of specific amplification with 1 µM. Barcode combination D×9 had very similar results (not shown)

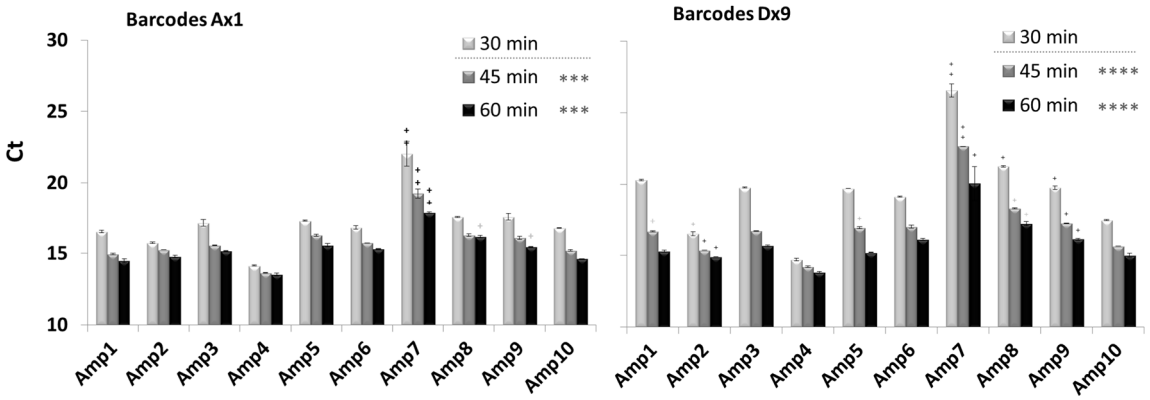



**Figure 20: Duration of the Klenow reaction.** Increasing the Klenow reaction from 30 min to 45 and 60 min improved the PCR efficiency significantly for both barcode combinations tested

### 3.1.3.2 Exonuclease reaction

Following synthesis of the primer duplex by Klenow fragment, the assembly protocol includes exonuclease removal of the anti-sense primer to prevent it from annealing to the sense primer during PCR cycles, which could inhibit amplification of the targets by reducing the availability of free primers in the reaction. The following sections summarize the experiments I performed for optimizing this step on multiple aspects.

#### 3.1.3.2.1 Comparison of T7 and Lambda exonucleases

I compared two 5'→3' exonucleases, T7 and Lambda (λ), to find out which one would perform better within our assembly protocol (**Figure 21A, B**). T7 preferentially hydrolyses the bases at the 5' end of a DNA duplex (Kerr and Sadowski, 1972), which can be inhibited by the presence of at least four phosphorothioate (PTO) bonds at the 5' (Nikiforov et al., 1994). Therefore, I coupled this enzyme with the barcodes containing PTO bonds between first six nucleotides, in order to protect the sense primers from degradation. Lambda exonuclease (λ-exo) has a much higher affinity towards 5'P ends compared to 5'OH ends (Little, 1967). Accordingly, I coupled this enzyme with the rc-primers ending with 5'P (purchased oligonucleotides contain 5'OH by default), so that the anti-sense strand would preferentially be degraded by the λ-exo.

Using λ exonuclease resulted in Ct values lower than or equal to T7 for the majority of the targets, and T7-treated samples exhibited distorted melting curves (**Figure 21C**). Furthermore, T7 exonuclease cannot be heat inactivated according to the manufacturer, which can result in undesired residual activity during the next steps of the workflow. For these reasons, I decided to include the λ exonuclease in the workflow in combination with rc-primers with 5'P ends.

#### 3.1.3.2.2 Exonuclease +/-

With an attempt to minimize the steps in the barcode assembly protocol, I asked whether the exonuclease removal of the anti-sense primers was indispensable for the PCR efficiency. I therefore compared using λ exo-treated primers with the non-treated counterparts for PreAmp PCR. Non-treated primers resulted in higher Ct values and/or distorted melt curves in comparison to the λ-treated ones. Running the same experiment with T7 exonuclease exhibited parallel patterns (not shown here because we discontinued its usage in the project). Consequently, I decided to keep the exonuclease removal of the anti-sense primers in the workflow.

#### 3.1.3.2.3 Lambda reaction duration

Having established that removing the anti-sense primers following Klenow reaction is crucial, the next logical step was to test whether increasing the duration of Lambda reaction can enhance the PreAmp PCR efficiency, possibly by ensuring complete hydrolysis of the anti-sense strands. Increasing the λ-exo treatment from 30 min to 60 and 120 min did not decrease, but slightly increased the Ct values (**Figure 22A**), possibly due to partial degradation of sense primers, too, due to residual activity of Lambda exonuclease towards 5'OH ends (Little, 1967). As a result, I decided to keep the 30 minutes λ-exo treatment, which seems to be adequate for hydrolyzing the anti-sense primers.

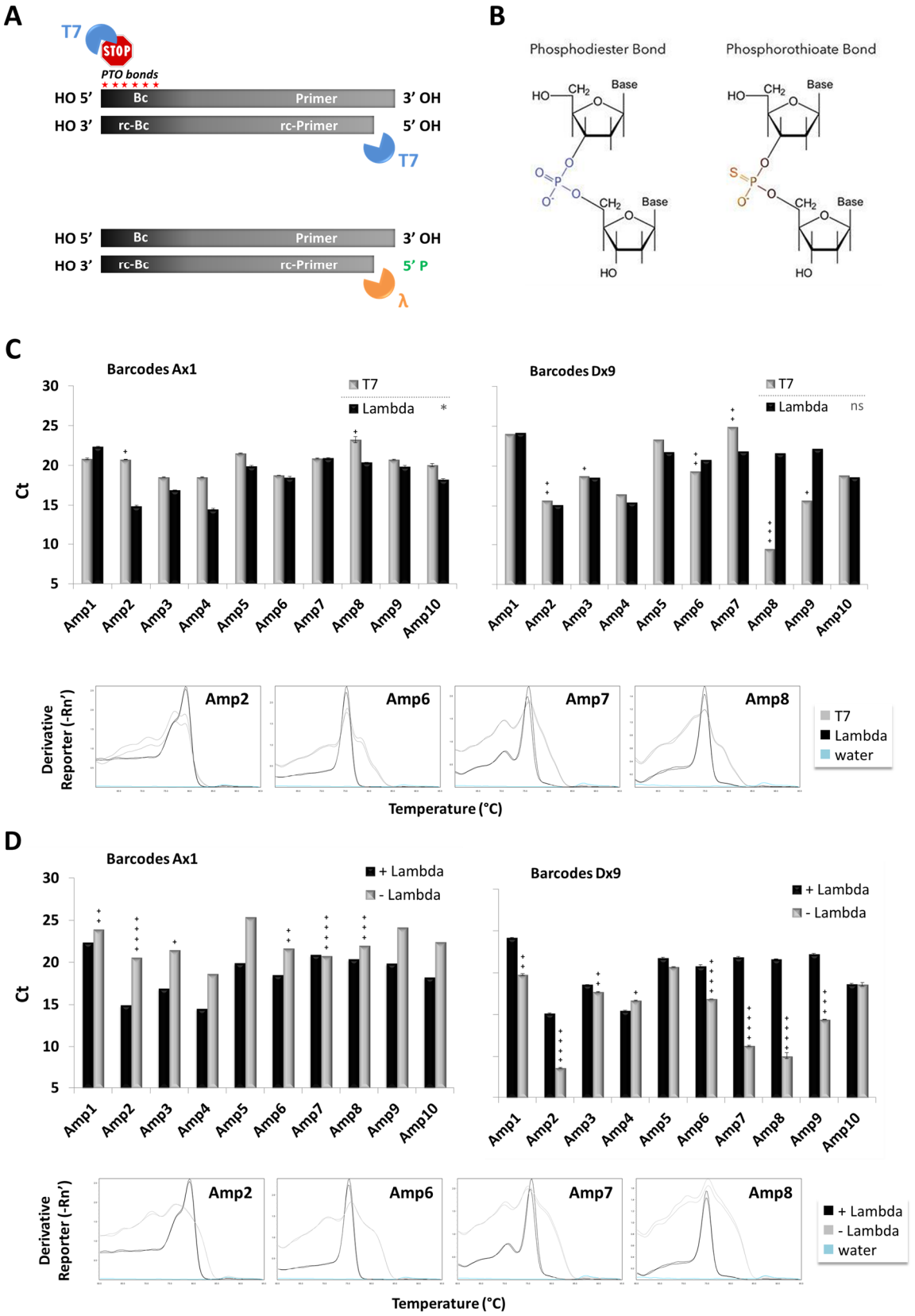



**Figure 21: Exonuclease treatment to remove anti-sense primers. (A)** T7-exo was combined with barcodes protected by 6 consecutive PTO bonds at the 5' end. λ-exo was combined with anti-sense primers harboring 5'P, to facilitate their preferential hydrolysis by the enzyme. **(B)** Comparison of

phosphorothioate (PTO) and phosphodiester bonds[11]. **(C)** Removing the anti-sense primers using λ-exo enhanced PCR efficiency compared to T7 with barcodes A×1. No difference in Ct values was observed with barcodes D×9, but melt curves were distorted with T7 in comparison to λ. **(D)** Skipping the exonuclease removal of anti-sense primers (- Lambda) led to reduced PCR efficiencies and almost complete loss of specific amplification of the targets. Reduced Ct values with barcodes D×9 is the result of non-specific products as implied by the melt curves

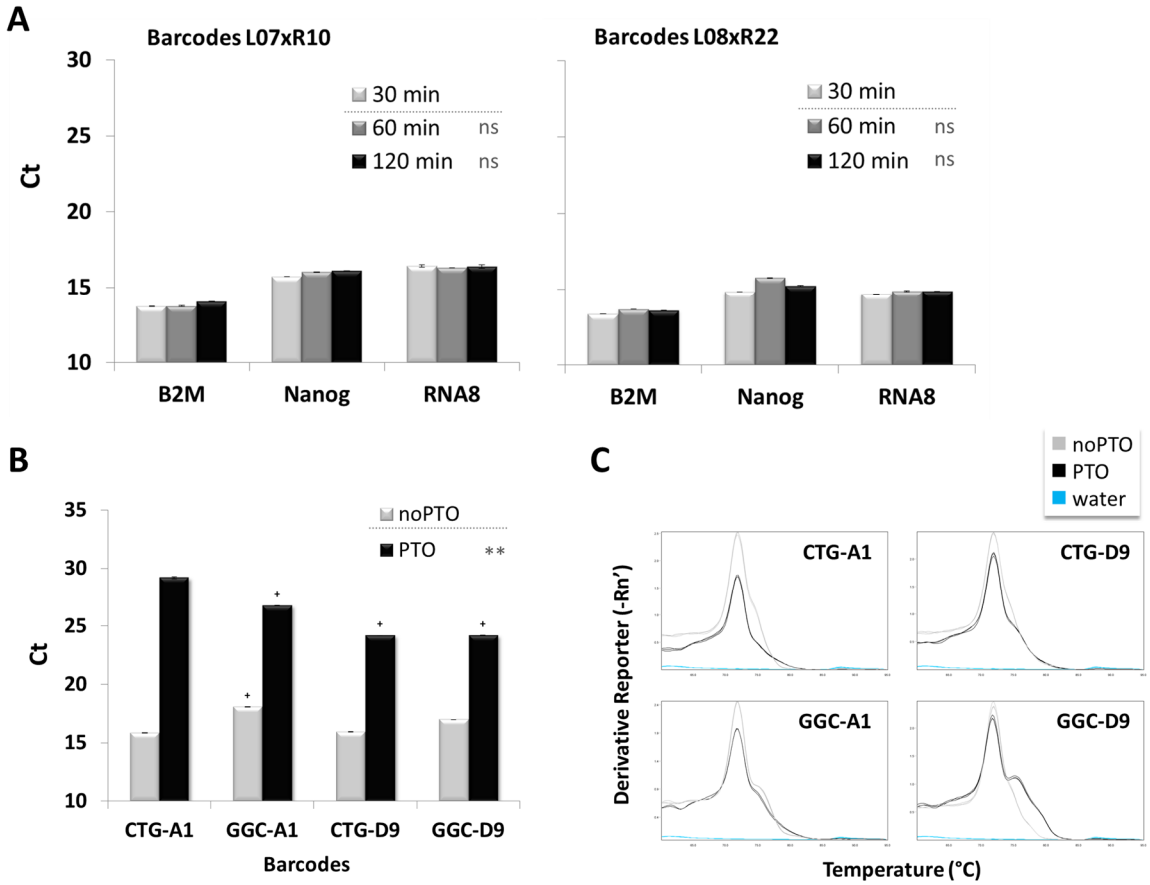



**Figure 22: Duration of λ exonuclease treatment. (A)** Increasing the duration of the Lambda reaction from 30 min to 60 min or 120 min did not reduce Ct values, while insignificantly increasing them. Melt curves did not display any secondary peaks in any of the conditions tested (not shown). **(B)** When the primers assembled with 5'PTO barcodes were treated with λ-exo, PCR efficiency decreased significantly in comparison to the same barcodes with regular phosphodiester bonds (noPTO). **(C)** Melt curves of the reactions in **B**. (target: Amp9)

#### 3.1.3.2.4 Protecting the barcode ends

Integrity of barcode sequences is imperative for the fidelity of BART-Seq workflow, because missing nucleotides pose the risk of barcode misidentification, thus read cross-contamination across the samples. Though with much lower processivity compared to 5'P ends, λ-exo can also degrade 5'OH barcode ends (Little, 1967). Because preliminary experiments that involved sequencing amplicon libraries displayed shortening of a small proportion of barcodes by a few bases at the 5' end,

[11] http://blog.biosearchtech.com/know-your-oligo-mod-phosphorothioate-bonds

we wanted to shield the barcodes from trimming by exonuclease. I initially tested whether using barcodes with 5'PTO bonds would protect them from potential degradation. Curiously, 5'PTO barcodes resulted in significantly reduced PCR efficiency in comparison to the ones with regular phosphodiester bonds (noPTO) (**Figure 22B**). Potential explanations might be higher affinity of Lambda towards 5'PTO ends, or getting trapped by the PTO bonds, in turn depriving the PCR reaction of free primers. I did not run further experiments to test these hypotheses; yet, practically decided not to use the barcodes with 5'PTO bonds.

I next tested if adding extra nucleotides to the 5' of barcodes, namely "protection groups", can preserve the intactness of the actual barcode sequences. When I compared different protection groups attached to the same barcode with qPCR (part of which is shown in **Figure 23A**) there was not pronounced differences among them. In order to analyze the protection groups more systematically, we designed an NGS experiment where we assembled primers with the barcodes flanked by additional 5' trinucleotides in all possible combinations (NNN) to amplify a constant amount of gDNA template, to identify the best sequences that could "protect" the barcodes from trimming. CCA- had the highest relative frequency among all the 64 combinations tested (**Figure 23C**). We therefore inferred that this group should be the most resilient sequence against exonuclease trimming and decided to include it at the 5' of barcodes as a protection group in the subsequent experiments.

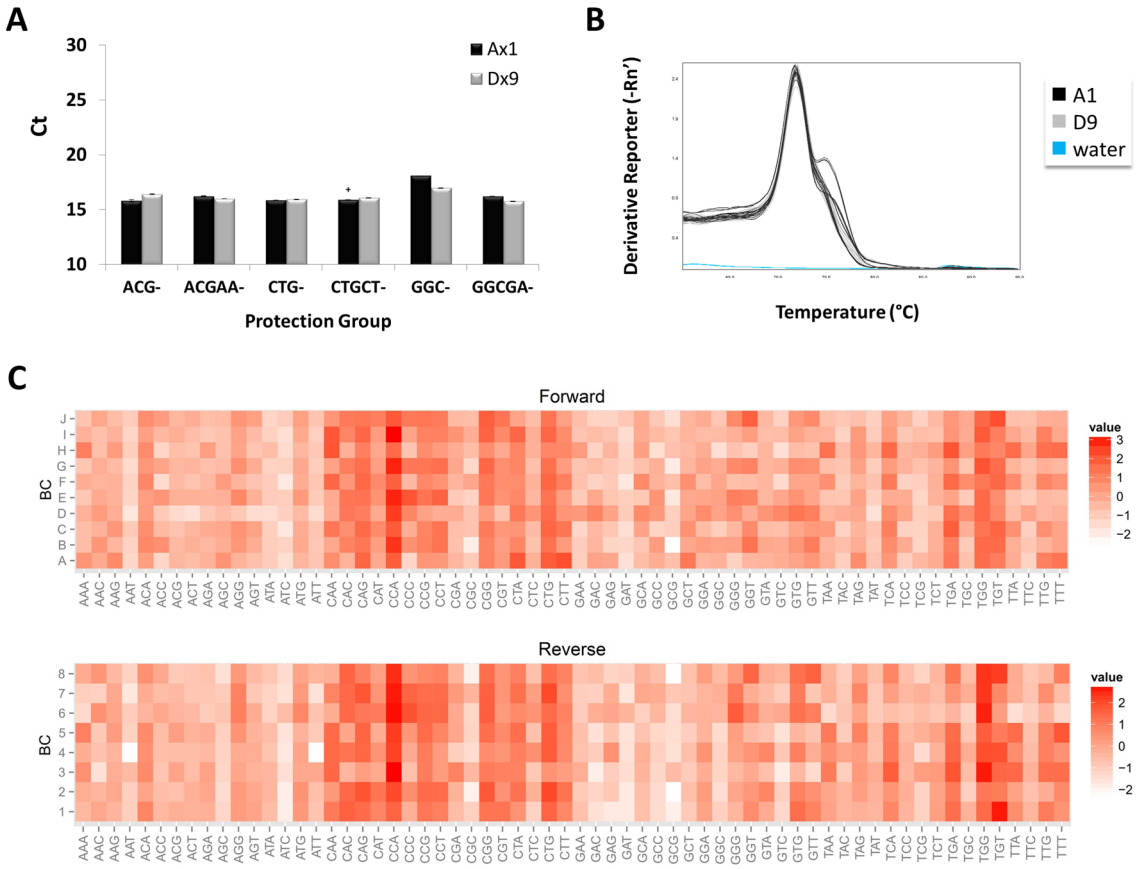



**Figure 23: Selecting a protection group for barcodes. (A-B)** Protection groups with different sequences and lengths resulted in very similar efficiencies when evaluated with qPCR (target: Amp9).

**(C)** A matrix of ten forward (A-J) and eight reverse (1-8) barcodes flanked with 5'NNN trinucleotides were assembled with the *BRCA* genotyping primer set to amplify constant amount of gDNA templates (20 pg/µL) from MCF7 cells. Heatmaps illustrate the sorted amplicons following NGS to each forward and reverse barcode. 5'CCA had the highest relative frequency among all the 64 trinucleotides tested, implying the highest resistance to λ-exo hydrolysis. Intensity range light to dark corresponds to low to high read numbers, respectively

### 3.1.4 Reverse transcription

For analyzing RNA samples with BART-Seq, a reverse transcription (RT) step is required in the workflow (**Figure 15**). For this, the RT mix that contains spike-ins was aliquoted to the wells, into which single cells were sorted (often with FACS) or bulk RNA samples were pipetted, and RT reaction was run. Given that the efficiency of RT can influence the accuracy of quantification of bulk samples and the extent of dropout events in single cells, I tested whether the modifications I made to adapt the protocol to our workflow has any influence on the efficiency of this reaction.

#### 3.1.4.1 RNase H treatment following reverse transcription

The commercial reverse transcriptases are engineered to have diminished RNase activity; consequently, an RNase H treatment step often follows the RT protocols to remove the template RNA molecules, since the RNA:cDNA hybrids might display higher stability compared to their homodimer counterparts, depending on their base composition (Gyi et al., 1998) and negatively influence the subsequent PCR. With the purpose of reducing the sample-intervention steps within the BART-Seq workflow, in particular when analyzing thousands of single cells, I explored whether the RNase H treatment could be skipped or combined with the subsequent PCR. Among the targets tested, only *SOX2* was negatively influenced when RNase H treatment was skipped (**Figure 24**). On the other hand, mixing the RNase H with the PCR reagents, even in lower concentrations, and running an additional step just before PCR restored its Ct values.

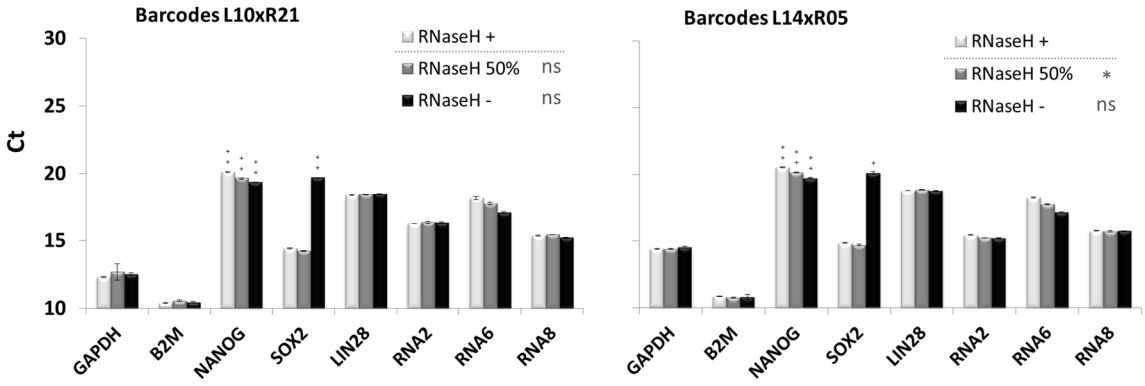
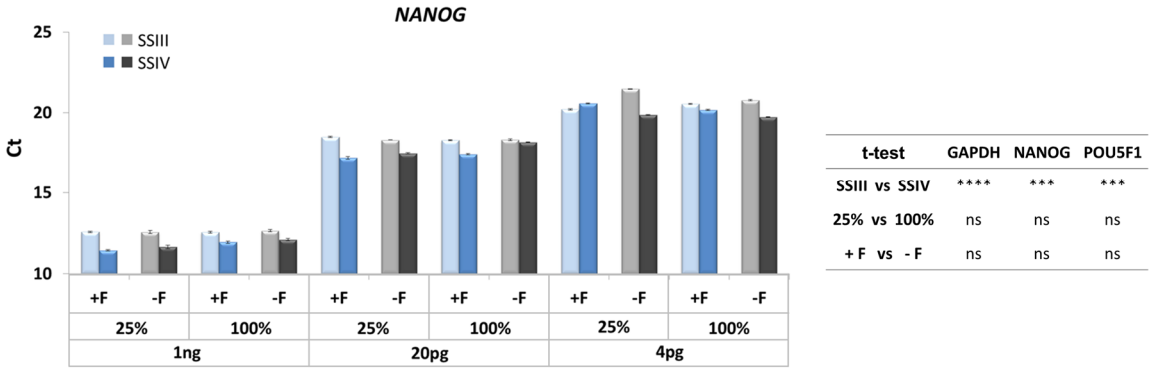



**Figure 24: RNase H treatment following reverse transcription.** Using 50% RNase H (RNaseH 50%) did not change the Ct values in comparison to the recommended concentration for the three different barcode combinations tested (two of them are shown). Skipping the RNase H addition (RNaseH -) did not influence the Ct values either, except for *SOX2*

### 3.1.4.2 Diluting and freeze-thawing reverse transcriptase

To make the mRNA of single cells available for reverse transcription following sorting, the workflow involves snap-freezing the cells together with the reverse transcription reagents and thawing for lysis, which could potentially damage the enzyme (Cao et al., 2003). I investigated whether this step reduces the efficiency of the reverse transcription. The efficiency of the snap-frozen reactions did not decrease as compared to the non-frozen ones (**Figure 25**). The same held true when I reduced the reverse transcriptase used per sample to 25% of the recommended concentration. The only significant reduction of Ct values was attained by using Superscript IV in comparison to Superscript III.




| t-test | GAPDH | NANOG | POU5F1 |
|---|---|---|---|
| SSIII vs SSIV | **** | *** | *** |
| 25% vs 100% | ns | ns | ns |
| + F vs - F | ns | ns | ns |

**Figure 25: Freeze-thawing or diluting the reverse transcriptase.** Snap freezing (+F) the reverse transcription mixture did not result in any significant differences in comparison to non-frozen (-F) samples for the template concentrations tested. Likewise, using 25% of the recommended enzyme concentration did not influence the efficiency. Superscript IV enzyme (SSIV) significantly improved the Ct values in comparison to SSIII. Very similar patterns were observed with *GAPDH* and *POU5F1*

## 3.1.5 Pre-amplification PCR

The PreAmp PCR is the step where the targets are co-amplified and samples are barcoded at the same time. I tested different parameters such as degree of multiplexing, primer concentrations, potential inhibitors, and so on, to increase the yield as much as possible to make it compatible for detecting minute amounts of template cDNAs from single cells, while reducing the consumption of multiplex PCR master mix, which constitutes one of the most expensive components of the workflow.

### 3.1.5.1 Multiplexing

The BART-Seq primers are relatively long oligonucleotides, ranging between ~38-50 nucleotides (barcode + linker + primer), which embrace a high potential of cross- or self-hybridization, especially when multiple primers are present in the same reaction. Therefore, I investigated whether multiplexing has any adverse effects on the uniformity and efficiency of barcode assembly or subsequent PCR. I compared multiplexing the primers ranging from 1 to 10 in parallel reactions and observed that increasing the number of multiplexed primers gradually did not influence the

efficiency of individual primers for the range tested (**Figure 26**).

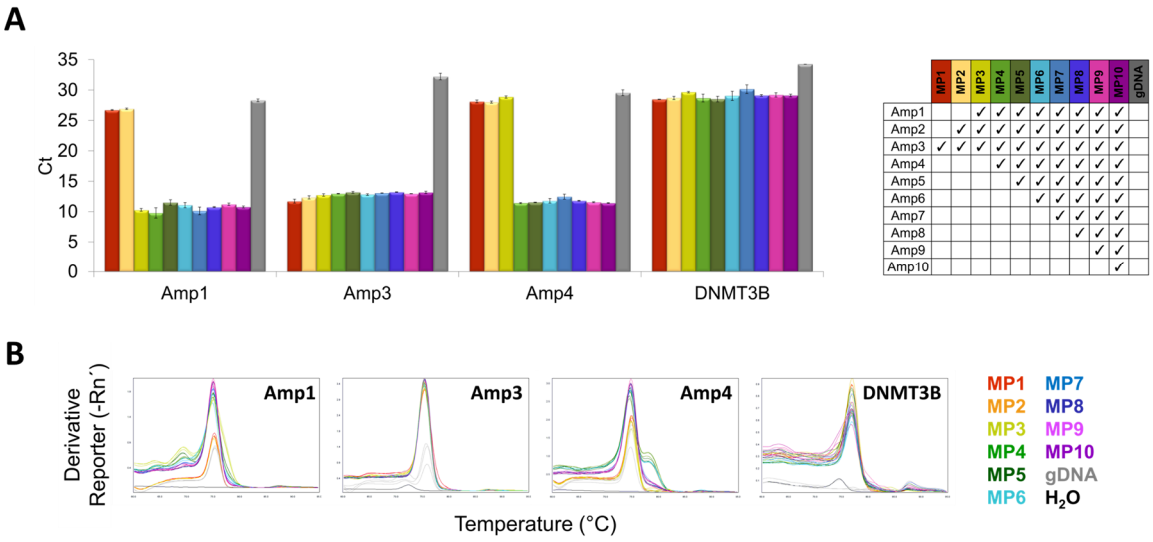



**Figure 26: Influence of multiplexing on the efficiency of barcode assembly and PCR. (A)** Multiplexing range from 1 to 10 was tested starting with Amp3 as singleplex, with the order shown in the right pane. The concentration of the individual primers was equal in all reactions, and the barcode concentration was matched to the total primer concentration. Total number of multiplexed primers did not influence the efficiency of individual primers. Non-pre-amplified gDNA and the non-targeted *DNMT3B* locus were used as negative controls. MCF7 gDNA was the template. **(B)** Melting curve signals of Amp3 was the same for the range of 1 to 10 primer pairs, while Amp1 and Amp4 exhibited weak signals in the reactions where they were not pre-amplified (MP1-2 and MP1-3, respectively)

### 3.1.5.2 Multiplex PCR master mix selection

Given that BART-Seq is based on multiplex pre-amplification, it requires a specialized master mix for the PCR. To find out which multiplex PCR master mix (MM) would perform better within the context of the workflow, I compared two kits, namely QiaGen and Platinum, and observed that the Platinum MM resulted in 25-fold higher amplification in average compared to QiaGen (**Figure 27**). Therefore, I integrated the Platinum MM into the workflow.

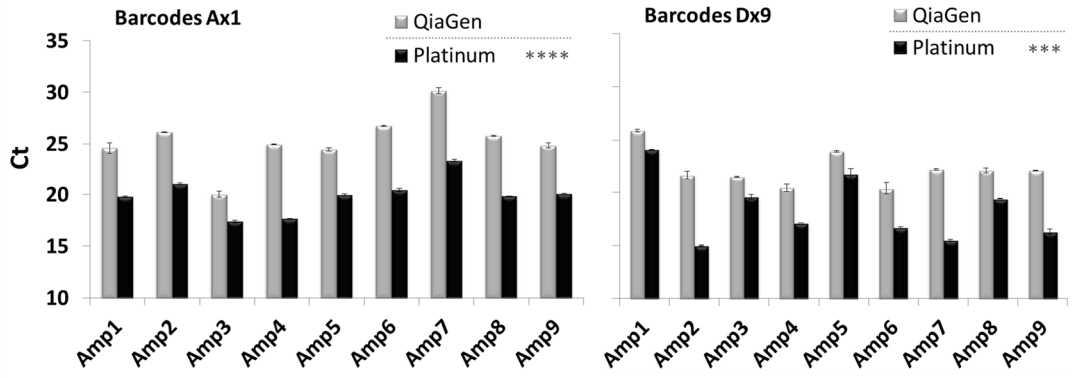



**Figure 27: Comparison of two multiplex PCR master mixes.** Platinum master mix yielded significantly higher efficiency compared to the QiaGen master mix for both barcode combinations tested. Melt curves did not indicate any non-specific amplification for both enzymes and barcode combinations (not shown)

### 3.1.5.3 PCR master mix dilution

The multiplex PCR MM is one of the most expensive reagents of the workflow; therefore, I tested whether we could attain similar efficiencies using reduced concentrations. Hypothetically, the enzyme concentration should not be a rate limiting factor for the PCR because single cells comprise tiny amounts of templates. Compared to the recommended dilution of Platinum master mix (100%), 67% and 50% dilutions resulted in slight reduction of PCR efficiencies (**Figure 28A**). It is known that $MgCl_2$ (up to a certain level) can increase the processivity of Taq DNA Polymerase (Henegariu et al., 1997). Accordingly, I supplemented the reactions with additional $MgCl_2$ to reach to the original concentrations (presumed to be 3 mM based on the equivalent MMs), and the Ct values became comparable to the recommended MM dilution. Further dilutions of the Platinum MM (50%, 33%, and 25%) resulted in similar efficiencies with additional $MgCl_2$ supplementation (to 3 mM, 4.5 mM, and 6 mM) (**Figure 28B**). Because some of the lower dilutions had slight melt curve irregularities, I decided to proceed with 50% of the recommended concentration with 6 mM final $MgCl_2$. Nevertheless, down to 25% dilution can be used for very large-scale experiments after testing the primers. Importantly, because the fidelity of Taq polymerase is inversely correlated with $Mg^{++}$ concentration (Eckert and Kunkel, 1990), the MM dilutions and $MgCl_2$ supplementation should be carefully adjusted for sensitive experiments, such as for mutation screening.

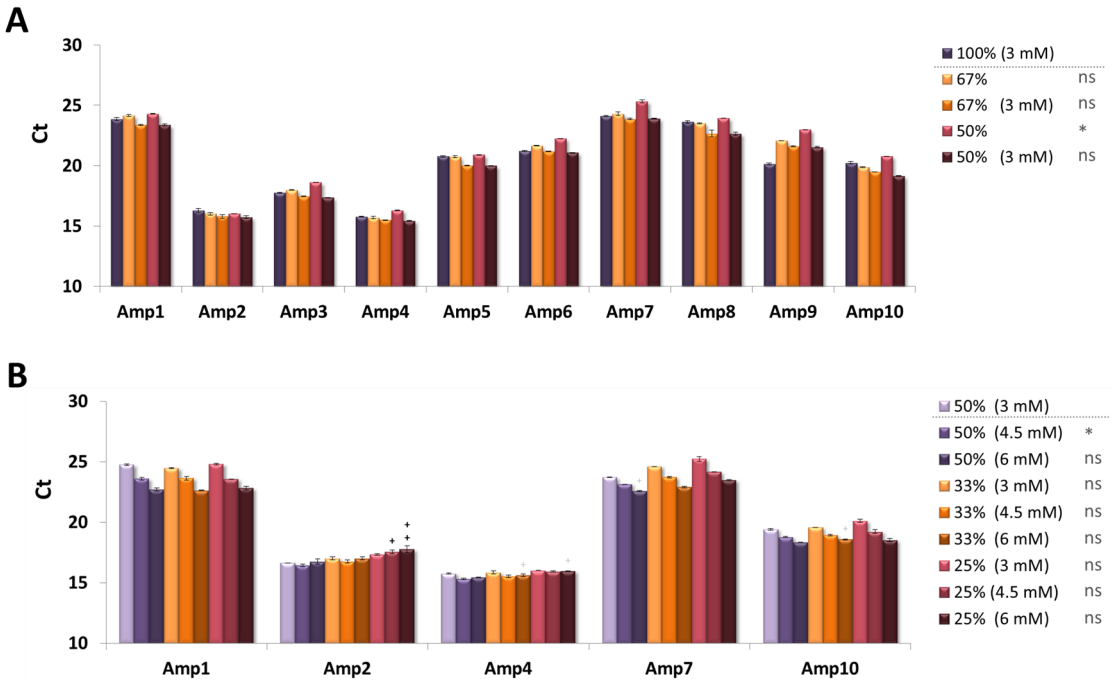



**Figure 28: Using reduced concentrations of the multiplex PCR master mix. (A)** Using 67% or 50% of the recommended concentrations of Platinum PCR MM resulted in slightly increased Ct values, which is restored by $MgCl_2$ supplementation (3 mM final). **(B)** Further dilutions of the MM down to 25% yielded comparable PCR efficiencies upon $MgCl_2$ supplementation (up to 6 mM final). The barcode combination D×9 resulted in very similar outcomes (not shown)

#### 3.1.5.4 Individual and total concentration of multiplexed primers

While the first set of primers we designed consisted of 10 targets, resulting in 0.025 µM of each primer and 0.5 µM total primers in the PCR, scaling up the primer sets brought about the question of whether the concentration of individual primers or the total concentration of multiplexed primers is more important for the efficiency of multiplex PCR. To test this, I ran experiments in a range of 0.025-0.1 µM individual primers, and 0.5-2 µM total primers, and observed that the concentration of individual primers has a higher influence on the PCR. Specifically, up to 0.03 µM of each primer resulted in good amplification, increasing it to 0.05 µM disrupted the melt curves of some targets, and to 0.1 µM totally abolished specific amplification of all the targets (results not shown). Therefore, I concluded that the total primer concentration does not adversely affect the specific amplification for the range tested (22 targets), as long as the optimum concentration of individual primers is ensured (in this case 0.03 µM maximum).

#### 3.1.5.5 Annealing temperature gradients

While the barcode+linker portion of the primers overhang in the initial cycles of PCR (**Figure 18**), they subsequently start to hybridize completely in the later cycles, resulting in two different melting temperatures (Tm). I tested whether increasing the annealing temperature (Ta) after the initial cycles of PCR would improve the PCR efficiency and specificity. For this, I compared various reactions with initial 3-5 cycles with 55-60 °C Ta and remaining cycles with 60-65 °C Ta. Because there was not any consistent improvement of the PCR efficiency with the gradients (data not shown), I decided to keep the Ta constant (58-60 °C) throughout the reaction for the subsequent experiments.

#### 3.1.5.6 RT/PCR ratio

Following cDNA synthesis from bulk RNA or single cells, the PCR components are added directly into the wells where the RT reaction took place. However, it is known that the presence of reverse transcriptase might have inhibitory effects on PCR efficiency especially with lower template concentrations, possibly because it can remain attached to the template (RNA or cDNA) despite heat inactivation, and mask the cDNAs from PCR (Chandler et al., 1998; Chumakov, 1994). Therefore, I compared the percentage of RT reaction volumes in the final PCR (RT/PCR) within a range of 10% to 30% (**Figure 29**). With QiaGen MM, PCR efficiency decreased gradually with increasing RT/PCR ratio. On the other hand, RT/PCR ratio up to 30% did not result any significant decrease with the Platinum MM; thus, I decided to use 20% RT/PCR in the subsequent experiments. Importantly, the slight decrease for the efficiency of the lowly expressed gene *LIN28A* above 25% supported the hypothesis of reduced effective template concentration mentioned above.

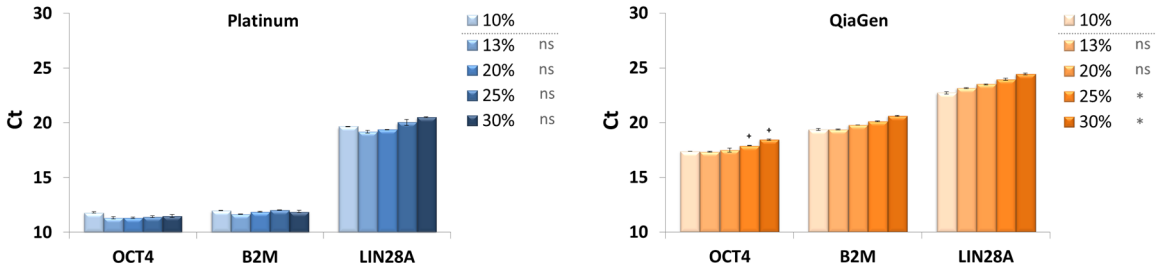



**Figure 29: Influence of the reverse transcription reaction volume ratio on the PCR.** The ratio of the RT reaction volume in the final PCR reaction (RT/PCR) was tested for the range of 10% to 30%. While the Platinum MM was not influenced for the genes tested, QiaGen master mix exhibited gradually decreasing efficiency with increasing RT/PCR volume ratios

### 3.1.6 Next-generation sequencing

The next step of the BART-Seq workflow following PreAmp PCR is pooling the amplicons from all the samples and preparing libraries for NGS. Because pooled samples from thousands of wells can reach up to tens of milliliters, Ethanol precipitation would be required so as to bring the volumes to the levels compatible with library preparation. Then, the libraries are prepared as described in **Materials and Methods**, and evaluated using Bioanalyzer before sequencing (**Figure 15**).

Following demultiplexing of the sequencing reads, we recurrently observed in the count matrices several reads assigned to negative control wells that did not contain any template RNA/DNA. Given the correlation between the magnitude of these background reads per target with the total number of reads that target received in the whole run, I hypothesized that this could be the result of index switching across the samples (Sinha et al., 2017). Due to the fact that BART-Seq libraries are inherently low diversity (consisting of PCR amplicons), I started to include the PhiX spike-in control[12] (5-15%) in the sequencing runs to mitigate this effect by increasing the diversity on the flow cell. Importantly, when I co-sequenced BART-Seq libraries (~35% of the run) together with non-BART-Seq samples, the reads assigned to control wells decreased significantly (e.g. 0 pg samples in **Figure 42C**). This showed that increasing the diversity of libraries can minimize read cross-assignment across the samples, thus enhance the accuracy of the results.

[12] https://www.illumina.com/products/by-type/sequencing-kits/cluster-gen-sequencing-reagents/phix-control-v3.html

### 3.1.7 Bioinformatics

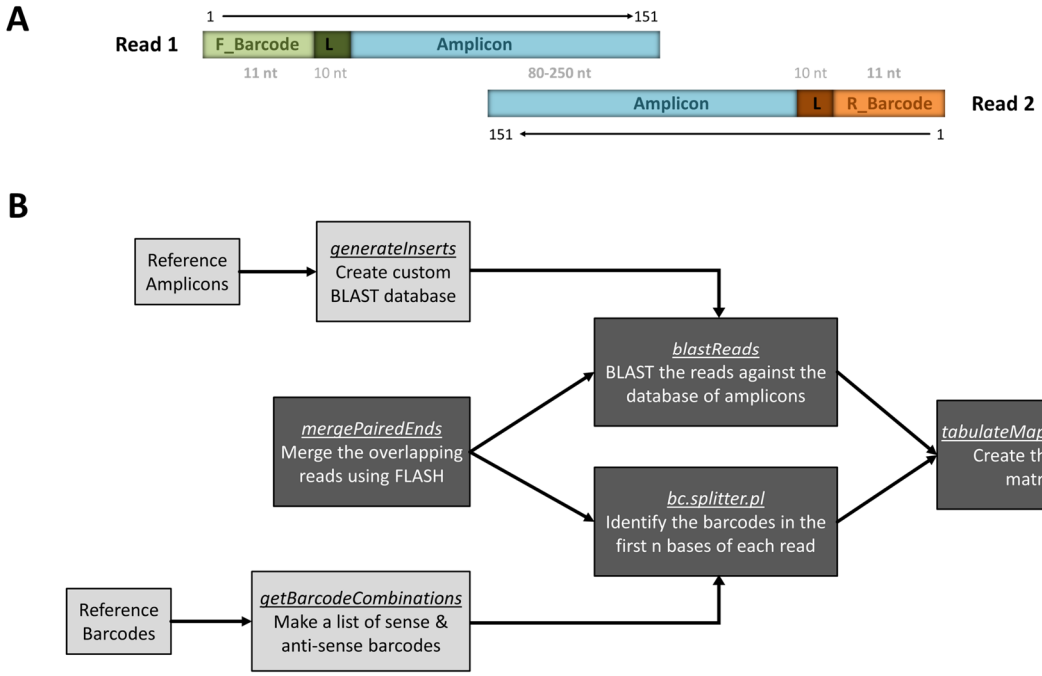
Bioinformatic analyses of the BART-Seq experiments consist of three main steps: **1)** demultiplexing the raw sequencing reads to generate count matrices, **2)** normalization and correction of the read counts, and **3)** analysis of the data to draw biological conclusions. This section provides a summary of the methods I implemented regarding these steps.

#### 3.1.7.1.1 Demultiplexing the RNA-Seq reads to count matrices

BART-Seq libraries consist of amplicons that contain linkers and barcodes on both ends, analyzed with paired-end sequencing (**Figure 30A**). As a result, read pairs in the FASTQ files must be processed by a tailor-made demultiplexing algorithm to identify the two barcodes and the amplicon contained between them. To this goal, we developed the methods and scripts for demultiplexing the NGS reads in collaboration with the Institute of Computational Biology of Helmholtz Center Munich.

#### 3.1.7.1.2 Merging the read pairs

The first demultiplexing algorithm developed for BART-Seq was based on merging the read pairs in the first step (*mergePairedEnds*) (**Figure 30B**). Then, the amplicons were BLASTed against the custom database consisting of expected amplicon sequences (*blastReads*). In parallel, first *n* nucleotides of each read (depending on the barcode length) were searched against the list of barcodes (*bc.splitter.pl*). Finally, using this information, the count matrices (as in **Figure 30C**) with the amplicon IDs and barcode combinations were generated. This algorithm worked very efficiently for the initial experiments (i.e. the MiSeq runs).




C

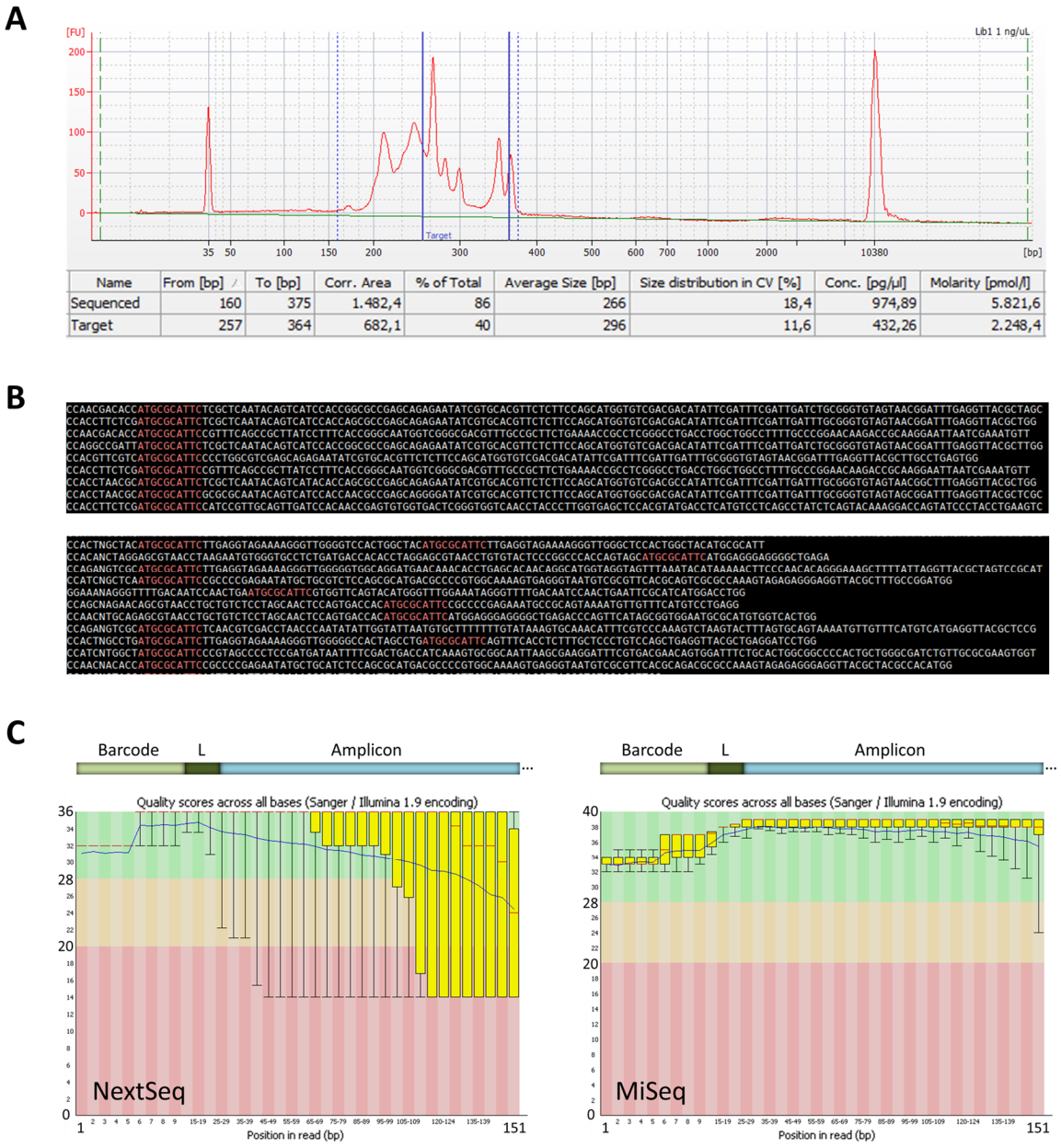
| bc_l | bc_r | B2M | CCND1 | CCNE1 | CER1 | DNMT3B | GAPDH | LIN28A | NANOG | POU5F1 | SOX2 | ZFP42 | RNA1 | RNA2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| L01 | R01 | 2443 | 349 | 209 | 1 | 1590 | 2481 | 12895 | 355 | 15830 | 1649 | 355 | 8055 | 644 |
| L01 | R03 | 2331 | 267 | 235 | 0 | 1497 | 2537 | 13326 | 385 | 15701 | 1793 | 362 | 8169 | 689 |
| L01 | R04 | 3108 | 412 | 239 | 0 | 2121 | 2952 | 6309 | 294 | 15333 | 2003 | 525 | 8224 | 1068 |
| L01 | R05 | 187 | 20 | 1 | 0 | 102 | 219 | 774 | 16 | 1242 | 120 | 25 | 7713 | 601 |
| L01 | R06 | 185 | 19 | 2 | 0 | 144 | 230 | 712 | 23 | 1257 | 138 | 28 | 9544 | 709 |
| L01 | R07 | 323 | 45 | 13 | 0 | 156 | 294 | 1660 | 40 | 1667 | 142 | 31 | 9656 | 932 |
| L01 | R08 | 88 | 4 | 5 | 0 | 76 | 148 | 418 | 8 | 329 | 33 | 3 | 6369 | 560 |
| L02 | R01 | 4337 | 664 | 295 | 4 | 2319 | 12207 | 21211 | 607 | 25527 | 2395 | 642 | 7220 | 1313 |
| L02 | R03 | 3502 | 537 | 241 | 2 | 2038 | 9936 | 19357 | 382 | 23859 | 2363 | 555 | 6120 | 865 |
| L02 | R04 | 4303 | 775 | 342 | 4 | 2459 | 11709 | 7805 | 528 | 21927 | 2486 | 636 | 6970 | 1334 |
| L02 | R05 | 230 | 40 | 6 | 0 | 168 | 959 | 1304 | 31 | 2208 | 151 | 62 | 6577 | 1267 |
| L02 | R06 | 367 | 51 | 7 | 0 | 218 | 1228 | 1197 | 29 | 2090 | 194 | 47 | 8238 | 1400 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

**Figure 30: The demultiplexing algorithm based on merging read pairs. (A)** Paired-end sequencing of the amplicon libraries results in partially overlapping reads each beginning with a barcode and a linker. **(B)** The earlier version of the demultiplexing pipeline merged the read pairs, and subsequently identified the barcodes and amplicons to create the count matrices. **(C)** Part of an example read count matrix

#### 3.1.7.1.3 De-multiplexing read pairs separately

Once I started to design larger scale experiments, i.e. analyzing thousands of samples, I decided to use the NextSeq instrument instead of MiSeq to ensure sufficient sequencing depth per sample. When we de-multiplexed these runs using the initial pipeline, the mapping percentage was unexpectedly low (down to 4% for some samples). Hence, I investigated the possible reasons of this problem from multiple aspects. First of all, re-inspecting the BioAnalyzer traces of the sequenced libraries so as to see whether the size range was accurate revealed a significant percentage of fragments that were out of the anticipated range (**Figure 31A**), implying that the expected amplicons comprised only approximately 40% of the

sequencing reads. This was because I had avoided performing a strict size selection as it would partially remove the target fragments as well, and there was the possibility that these smaller fragments could be shorter variants of some transcripts.




| Name | From [bp] | To [bp] | Corr. Area | % of Total | Average Size [bp] | Size distribution in CV [%] | Conc. [pg/µl] | Molarity [pmol/l] |
|---|---|---|---|---|---|---|---|---|
| Sequenced | 160 | 375 | 1.482,4 | 86 | 266 | 18,4 | 974,89 | 5.821,6 |
| Target | 257 | 364 | 682,1 | 40 | 296 | 11,6 | 432,26 | 2.248,4 |

**Figure 31: Investigation of a sequencing run with sub-optimal quality. (A)** Sample Bioanalyzer trace of a sequencing library. A significant percentage of fragments (dashed lines) were shorter than the anticipated range (257-364 nt) (solid lines). **(B)** Linker sequences were found in random positions in a low quality library (bottom panel) in contrast to the uniform localization in a high quality one (upper panel), when the raw reads in the FASTQ files were inspected. **(C)** Per base sequence quality of a library analyzed with a NextSeq instrument (left) showing decreased qualities towards read ends in contrast to the same library analyzed with a MiSeq instrument (right)

Next, we manually inspected the raw reads inside the FASTQ files to see whether the library contained correct amplicons. Many reads contained barcode and linker sequences at random positions (**Figure 31B**, lower panel), sometimes at multiple locations within one read. Moreover, in comparison to a high quality run (**Figure 31B**, upper panel), many reads were shorter than 151 nt (the maximum read length). Since these experiments consisted mainly of single cells, I hypothesized that those unusual reads might have originated from concatemerized excess primers, and because our size selection was not strict, they should be retained in the libraries and sequenced.

Nonetheless, we should still have recovered around 40% of the reads in theory, which was not the case. We then hypothesized that a failure in merging the read pairs due to base mismatches or quality trimming of the read ends might be the reason of data loss. To investigate this, we checked per base quality of the raw reads using FASTQC tool and observed a massive decrease after cycles 90-100 (**Figure 31C**, left). This could be a joint result of low diversity libraries and the two-channel chemistry of the NextSeq instrument, because sequencing the same library on MiSeq resulted in very high base qualities throughout the reads (**Figure 31C**, right). Having confirmed the hypothesis of failed merging, we decided to design a new algorithm that is robust to all these issues.

The new demultiplexing pipeline processes the read pairs separately (**Figure 32**), and combines the information at the last step, as follows: The reads are first trimmed based on the per base quality scores (*trim quality*). Then, the barcode is identified with reference to a table of barcode sequences and their potential single nucleotide variants, excluding potential overlaps between barcodes (*tag reads*). The next step aligns the reads to the amplicons, using a minimum pre-defined length of nucleotides to prevent potential false positives that may originate from primer-only reads or concatemers (*map reads*). Here, the anti-sense barcodes and linkers at the end of the amplicons shorter than 151 nt are excluded from alignment. Finally, a table containing the reads annotated with barcode and amplicon information is created (*R1*, *R2*), which is then summarized as a read count matrix (*spreadsheet*). When we re-processed the NextSeq run mentioned above using this pipeline, percentage of mapping increased from 4% to 42%, which agreed with the anticipated ratio based on the BioAnalyzer traces, suggesting that we could recover all the true amplicons from the experiment.

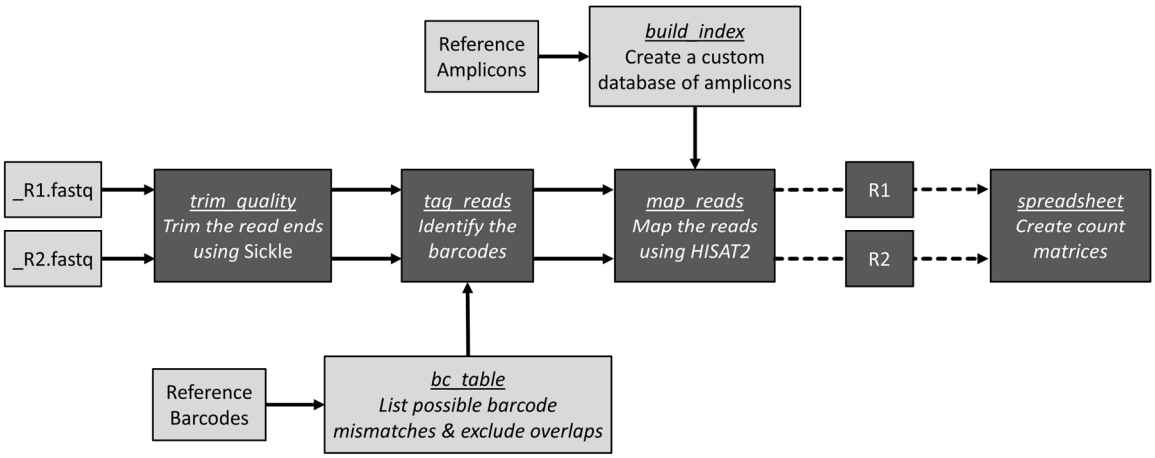



**Figure 32: The demultiplexing algorithm for processing the read pairs separately.** The new demultiplexing pipeline processes the read pairs separately and combines the information from the two reads in the last step to create a count matrix

### 3.1.7.2 Normalization of count matrices

For transcriptomics experiments, I routinely added fixed amounts of RNA spike-ins to each sample to calculate the technical variations that might arise within the workflow starting from the reverse transcription till the end of sequencing, and to normalize the data. Although the stochastic events at the molecular level cannot be estimated, the systematic effects within or across the samples such as RT efficiency, PCR efficiency, pipetting errors, degradation, evaporation and so on should theoretically influence all the amplicons within a sample the same way, and should be reflected by the spike-in reads. To confirm this, I analyzed an experiment where a mixture of bulk RNA and spike-ins was aliquoted into multiple wells, reverse transcribed, and tagged with 225 different barcode combinations. Sum of the gene reads and sum of the spike-in reads had a high correlation (R=0.8219) across the wells, indicating that the spike-ins can account for the majority of the technical variations (**Figure 33**).

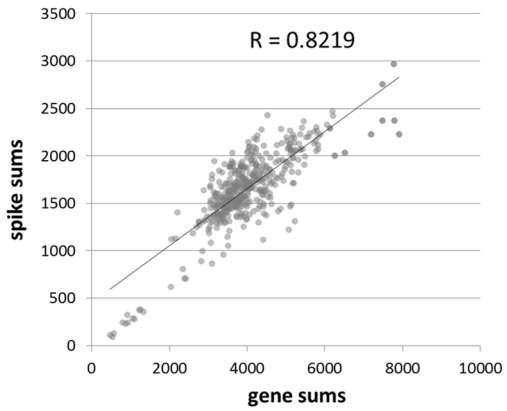



**Figure 33: Spike-in reads can estimate the technical variations.** Correlation of sum of the spike-in reads to sum of the gene reads (R=0.8219) per sample in an experiment where constant amount of bulk RNA & spike-in mixture was pre-amplified using 225 different barcode combinations (each represented with a dot)

#### 3.1.7.2.1 Barcode-primer combination effect

Besides the technical variations within the data, there were consistent differences in the efficiency of barcodes that influenced all the amplicons (including spike-ins) similarly, which could in theory be eliminated during normalization. However, an entirely unexpected phenomenon I discovered during the analyses was the variation of barcode efficiencies depending on the primer they are combined with, which I named "barcode-primer combination effect" (**Figure 34**). Because it perturbed the read counts of amplicons (including spike-ins) non-uniformly within each sample, it biased the scaling factors that were calculated using spike-ins (**Figure 34**, black dots), and in turn distorted the data during normalization. In order to solve this problem, I ran several analyses to empirically and computationally calculate and correct this effect in the existing experiments, and to find out the underlying mechanism in order to avoid it in the future experiments.

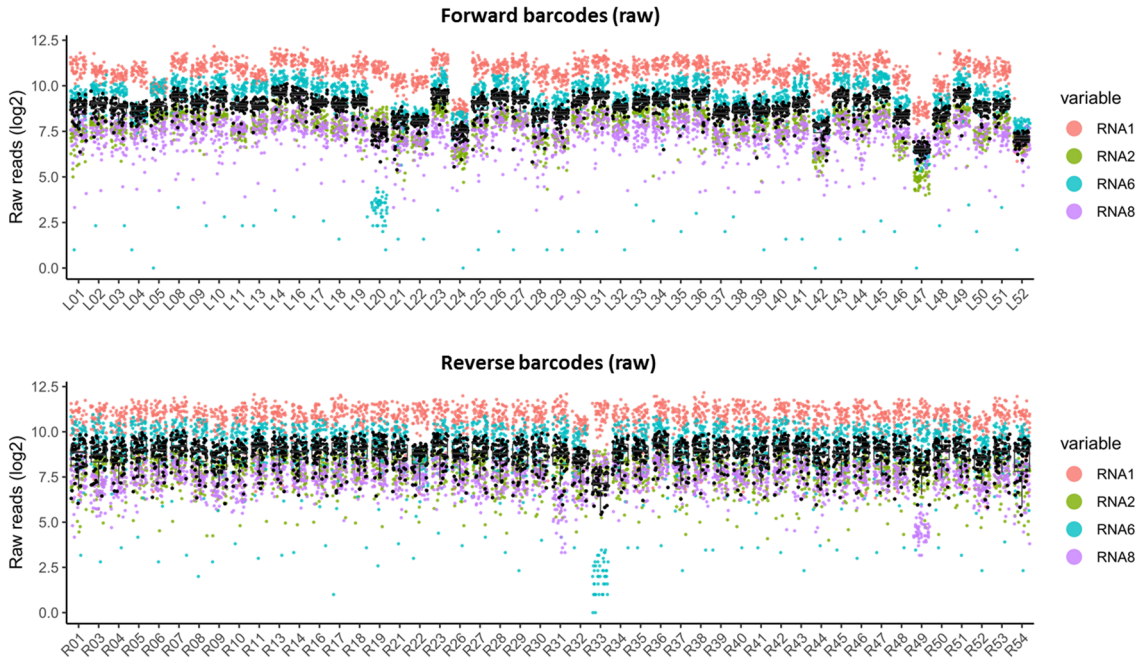



**Figure 34: Global and primer-specific variation of barcode efficiencies.** Two types of systematic variations existed in the read-count matrices. Global barcode efficiencies influenced all the targets similarly (more evident in the upper panel). Barcode-primer combination effects influenced the read counts of only some genes/spike-ins and biased the scaling factors (black dots) that were calculated using the spike-ins. e.g. R33-RNA6, R49-RNA8, L20-RNA6

#### 3.1.7.2.2 Correction and normalization

##### *Modeling and correcting the barcode-primer combination effects*

Given the consistent patterns the barcode-primer combination effects had across the count matrix, I hypothesized that it should in theory be possible to calculate and correct them. After trying numerous strategies, I finally decided to build negative binomial generalized linear models (nb-glm) from the spike-in reads and try to

estimate an efficiency coefficient for each combination. For this, I initially modeled the spike-ins using the full formula below:

$$full\_model: \quad read\ count \sim variable + cells + well.location + forward + reverse + forward{:}variable + reverse{:}variable \tag{1}$$

where I used the spike-in ID (*variable*), number of sorted cells to the well (*cells*), location of the well on the plate (side/corner/middle) (*location*), global efficiency of the barcodes (*forward*/*reverse*), and combination of barcodes and primers (*forward:variable, reverse:variable*) as the explanatory variables.

The distribution of the spike-in reads visually resembled a negative binomial distribution (Gierliński et al., 2015; Jiang et al., 2011). I verified the power of the fit by building a model from 95% of the data and using it to predict the test data (5%), which resulted in very high correlation coefficients (R=0.97 in average from 10 repetitions) (**Figure 35**). Next, I modelled the complete data (100%) (*full_model*), and tagged individual reads as outliers if they deviated from the predictions more than 2-fold, and filtered out the wells completely if they contained more than two outlier spike-ins. Subsequently, I replaced the remaining outliers with the predictions of the model that was built with the same formula using the filtered data.

The next step following the clean-up was correcting the barcode-primer combination effects. Importantly, I wanted to correct in this step "only" the barcode-primer combination effects (i.e. *forward:variable* and *reverse:variable)* while preserving the global ones, because they should still account for technical variations during normalization.

Because the model uses one of the elements of each explanatory variable as intercept, it was not possible to extract the full set of coefficients directly from the model. To overcome this, I developed a strategy using two basic models, one taking the combination effects into account (*model#1*) and one ignoring them (*model#0*):

$$\text{model\#1:} \quad read\ count \sim variable + forward + reverse + forward{:}variable + reverse{:}variable \tag{2}$$

$$\text{model\#0:} \quad read\ count \sim variable + forward + reverse \tag{3}$$

I calculated the *correction factors* by dividing the predictions of the *model#1* to the predictions of the *model#0*, which should theoretically reflect the variations that result only from combinations. Then, I corrected the raw reads by dividing them with these *correction factors*. I hypothesized that if the barcode-primer combination effects are corrected, only the variations that influence all the reads similarly should remain, thus the co-variation of the spike-ins should improve. To test this, I compared the pairwise correlations of spike-ins before and after the correction, and observed remarkable increase in the latter, indicating the success of the correction strategy (**Figure 36**). A shortened version of the R script is provided in **APPENDIX E.**

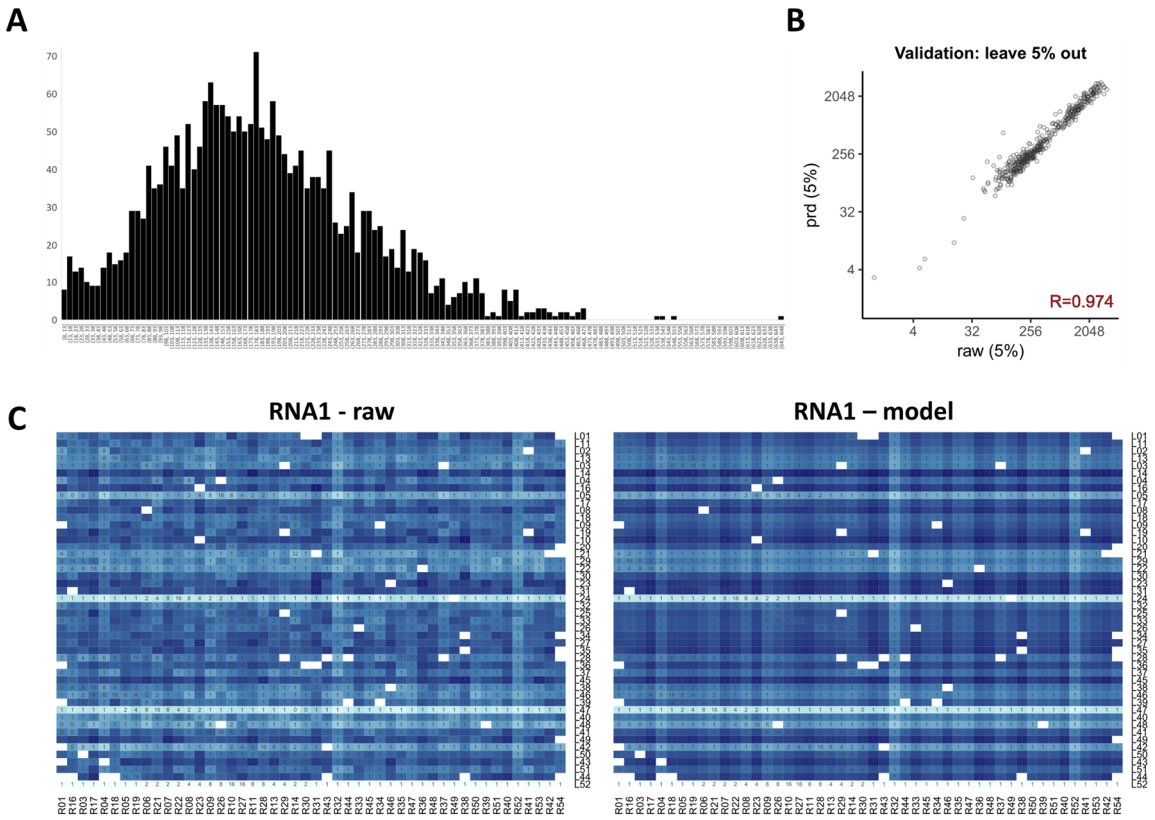



**Figure 35: Fitting a negative binomial generalized linear model to spike-in reads. (A)** Histogram of the raw reads counts assigned to one of the spike-ins (bin size: 5) **(B)** Predictions of the 5% of the data using the model built from 95% of the data correlates very highly with the raw reads (R=0.97, average of 10 repetitions). **(C)** Heatmaps visually comparing the raw reads of the spike-in RNA1 (left) with the predictions of the model (right)

### *Normalization*

The next step was the normalization of the count matrices using the corrected spike-in reads. Before the normalization, I removed all the barcodes exhibiting extremely low efficiencies in combination with any of the primers. Then, I normalized the reads by calculating scaling factors ($RNA_x$) using either corrected spike-ins (4) only or spike-ins and genes together (5) as follows:

$$\mathrm{RNA_x} = (2^{\left(\frac{1}{N}\Sigma_1^N \log_2(\mathrm{spike_n}+1)\right)} - 1)/median \quad (4)$$

or

$$\mathrm{RNA_x} = (2^{\left(\frac{1}{N}\Sigma_1^N \log_2(\mathrm{gene_n}+1)\right)} - 1)/median \quad (5)$$

I removed the samples if its scaling factor differed more than 10-fold from median of all scaling factors in order to prevent over-correction. Finally, I divided the raw read counts of the transcripts with the scaling factors. Variation of each spike-in across the data shrank to a 2-fold range after normalizing, demonstrating the efficiency of the correction and normalization strategy.

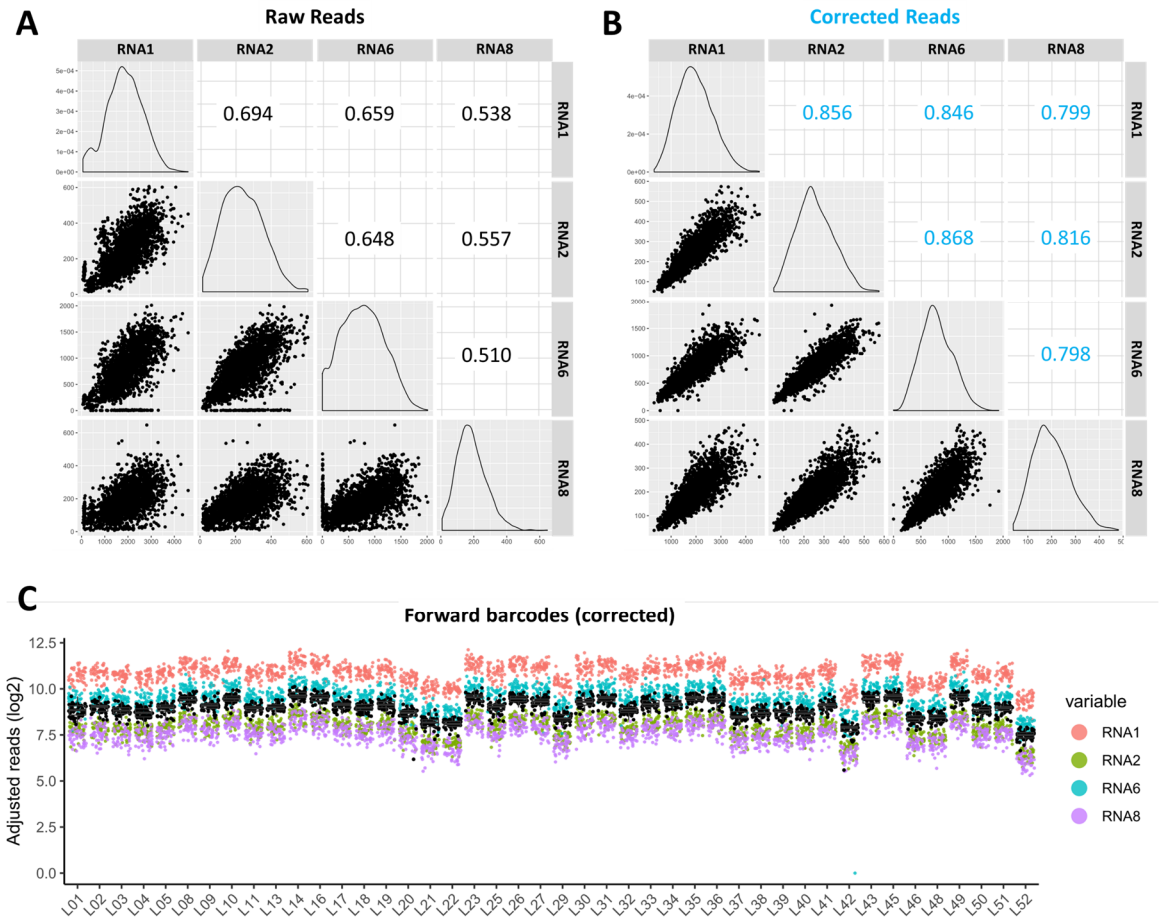



**Figure 36: Correction of the spike-in reads using model fits.** Pairwise correlations of spike-ins increased significantly (*P-value* < 0.0001) **(B)** after correction of the barcode-primer combination effects compared to **(A)** raw reads. **(C)** Improved co-variation of spike-in reads following correction was evident visually (compare to **Figure 34**, upper panel)

#### 3.1.7.2.3 Empirical analysis of barcode efficiencies

Although I could model and correct the sub-optimal barcode-primer combinations for the spike-ins that had constant input concentrations, it was not feasible to do the same for the genes that are heterogeneously expressed. Ideally, all the possible combinations of barcodes with all the gene and spike-in primers using constant template should allow to empirically estimate all such effects and to exclude them from the subsequent experiments. Yet, this would require a massive and expensive sequencing run. Because the effects tend to be consistent across the samples, I hypothesized that combining a few forward/reverse barcodes with all the barcodes of the opposite type should simulate the complete matrix. Therefore, I combined three forward barcodes with all the reverse barcodes and vice versa (**Figure 37A**), and assembled them with three different primer sets (**APPENDIX J, APPENDIX K**) to amplify constant amount of templates.

As I inspected the heatmaps or scatter plots of the raw data, the sub-optimal barcodes were readily noticeable even by eye (**Figure 37B**, top). In order to numerically determine global and primer-specific barcode inefficiencies, I built a

custom R script, which runs as follows: First, I adjusted any differences between the three common barcodes, and median centered each gene by subtraction (using log2-transformed reads) (**Figure 37B**, middle). Using this matrix, I calculated the global efficiency of each barcode as the average distance to the median. Then, I subtracted the global efficiencies from the barcodes, which should leave the barcode-primer combination effects (**Figure 37B**, bottom). Then, I set empirical thresholds to make a list of globally inefficient barcodes and barcode-primer combinations, which would serve a reference for the subsequent experiments, and for the computational analysis of barcode efficiencies.

Although I attempted to employ a more sophisticated method, for example modeling the data with nb-glm, whose predictions correlated with the raw data perfectly (i.e. R = 0.99); the usage of the first component of each variable as the intercept (e.g. all the Reverse barcodes were calculated relative to R01) rendered it useless to calculate the efficiencies accurately. As a result, I implemented the method above.

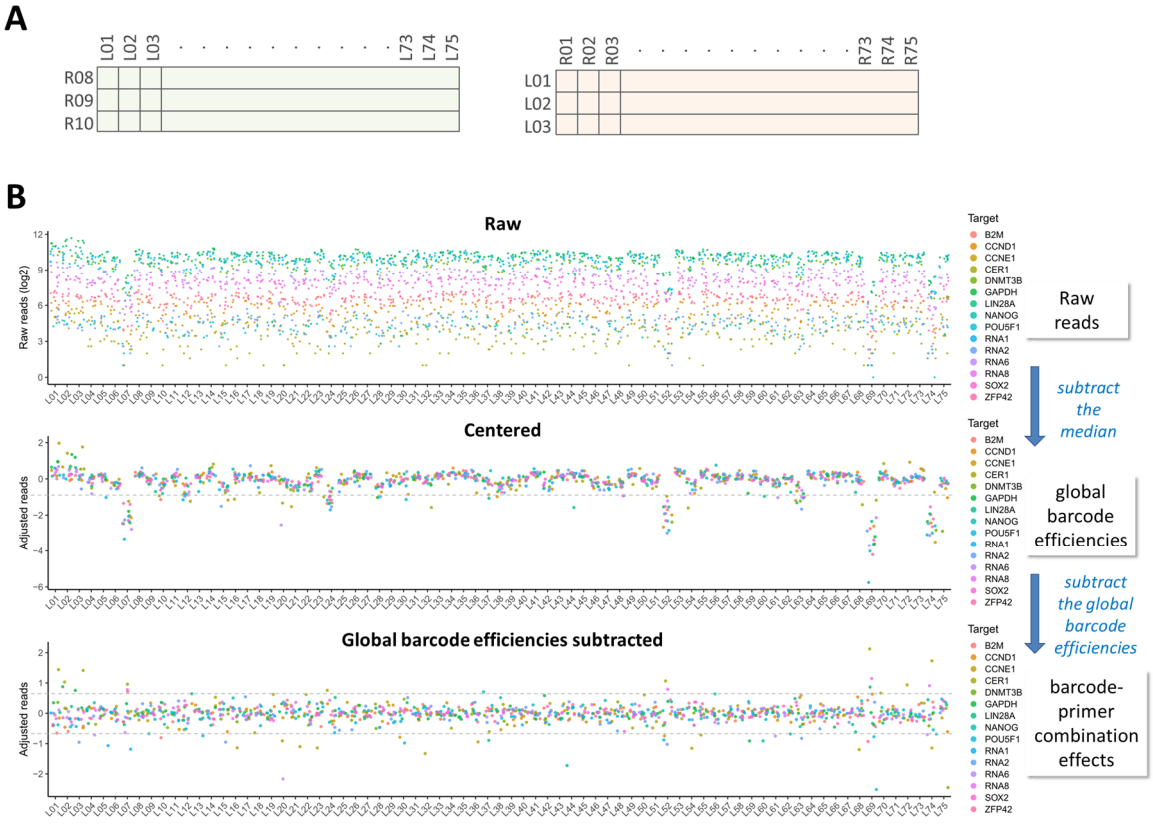



**Figure 37: Determining inefficient barcodes and barcode-primer combinations empirically.** **(A)** As a proxy for a matrix of all possible combinations of 75 Forward × 75 Reverse barcodes, three Reverse barcodes were combined with 75 Forward barcodes, and vice versa, to amplify a constant template. **(B) Upper:** Raw reads (log2) generated from bulk RNA templates using the barcode combinations in **A** in combination with pluripotency primers. **Middle**: Median centering each gene reveals global efficiency of the barcodes. **Bottom**: Subtracting the global barcode efficiencies from the median-centered reads reveals barcode-primer combination effects. Thresholds were decided empirically based on the distribution of the reads

#### 3.1.7.2.4 Computational analysis of barcode efficiencies

Because I postulated that there should be an underlying rule that can explain the reduced barcode-primer efficiencies, I tried to correlate the *in silico* calculated thermodynamic properties of barcodes and primers with the empirical findings from the previous section. For this, I initially predicted[13] secondary (2°) structure of the amplicon pairs that differed only in the 8 nt barcode on one end. Investigating amplicon pairs, one with an inefficient barcode-primer combination, and the other with an average combination pointed to the first 30-40 nucleotides, which correspond exactly to the barcode-linker-primer stretch. While the inefficient combinations formed a stable hairpin structure in this region, the average combinations had weaker and smaller hairpins (**Figure 38A**).

Following this observation, I shifted my focus to the barcode-linker-primer sequences. Folding the barcode-primer combinations that were known to be inefficient resulted in stronger 2° structures in comparison to the same primer in an average barcode combination (**Figure 38B**). Next, I simulated and calculated the minimum free energies (MFE) of all the possible primer-barcode combinations using the RNAfold software from the Vienna Package (R version 3.6.0). Computationally calculated MFEs of the experimentally identified inefficient barcode-primer pairs often fell below -10, whereas the average combinations of the same primers remained above this level (**Figure 38C**). An extreme confirmation from the opposite end was L19-EOMES that had much higher read counts than the average and a very high MFE. Overall, these results indicated that MFE can be a robust predictor of the barcoded primer efficiencies, which should be taken into account when designing the BART-Seq primers.

Finally, I investigated the barcodes that were globally inefficient regardless of the primer they are combined with (e.g. L07, L52, L69, L74), which consistently appeared in all the experiments. Comparing their 2° structures (barcode+linker sequences), MFEs, or similarity to human genome & transcriptome (BLAST) with the average barcodes did not reveal any differences. Surprisingly, when I analyzed the dimer formations, the only four barcodes forming self-dimers with 12 consecutive bonds were the ones that were found to be the most inefficient. This was because all four of them ended with "GA", increasing the complementarity of the linker, which turns out to be already self-complementary for 8 nt out of 10 (**Figure 39**). Self-dimerization possibly reduced the efficiency of the Klenow reaction, thus decreased the concentration of assembled primers with these barcodes, an important consideration when designing new barcode panels in the future.

[13] https://rna.urmc.rochester.edu/RNAstructureWeb/Servers/Predict1/Predict1.html

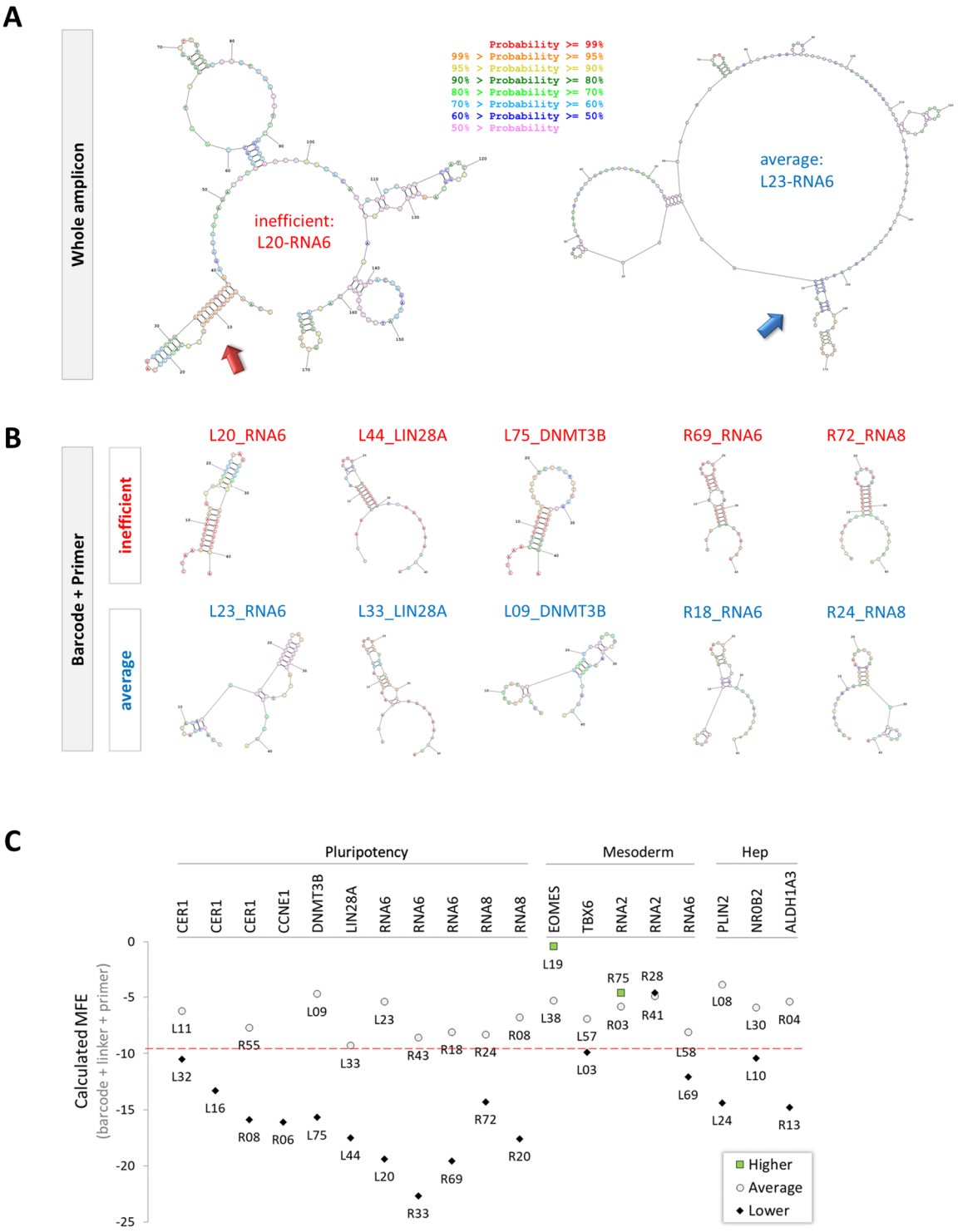



**Figure 38: Barcode-primer combination effect explained by minimum free energies. (A)** Folding analysis[13] of the whole amplicons that differ only in the 8 nt barcode (inefficient vs average barcode combination) point to the differences in the barcode-linker-primer stretch. **(B)** Folding analysis of only barcode-linker-primer sequences shows stronger 2° structures of the inefficient barcode combinations compared to average barcode combinations of the same primers. **(C)** Calculated minimum free energies (MFE) of the empirically determined (**Figure 37**) high-, average-, and low-efficiency barcode-linker-primer sequences. Low MFEs (< -10) corresponded to inefficient combinations, while over-efficient combinations had above-the-average MFEs (> -5) (e.g. L19-EOMES)

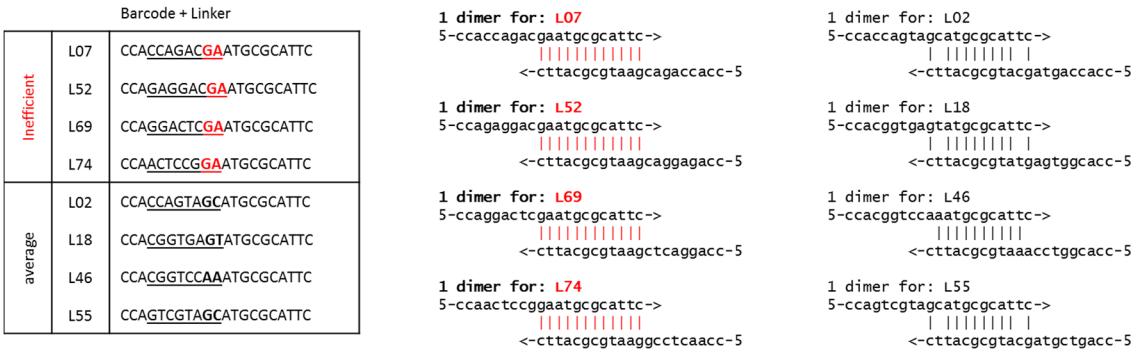



**Figure 39: Global barcode inefficiencies explained by stable dimerization.** The most inefficient four barcodes (**Figure 37B**) ends with GA dinucleotide, which causes them to form very strong dimers when combined with the linker sequence (12 consecutive bonds), possibly decreasing the effective amount of primers synthesized during the Klenow step

## 3.2 Applications of BART-Seq

The ultimate goal of this thesis was to establish a complete workflow that can be used to address specific biological questions. This section summarizes the applications of the BART-Seq method; first, to validate its power, and then to answer biological questions by analyzing single cells and bulk samples.

### 3.2.1 Validation of the barcode assembly

#### 3.2.1.1 Co-amplification of genomic targets

In order confirm that the assembly method can generate excess of barcoded primers to specifically enrich the targeted loci, I initially used qPCR to compare the enriched and non-enriched samples. For this, I used multiplex primers to co-amplify 10 genomic loci within the human *BRCA1* and *BRCA2* genes from human bulk gDNA. qPCR analysis indicated specific enrichment of all 10 loci (Amp1-10) with very similar efficiencies using four alternative barcode combinations, in contrast to non-pre-amplified gDNA, non-enriched loci (*MSX1*, *ZIC1*), or non-assembled reverse complementary primers (**Figure 40A**).

Next, we wanted to use NGS to confirm the target enrichment, which is indeed the ultimate objective of the workflow. To this goal, we scaled up the size of the barcode matrix to amplify the 10 *BRCA* loci from bulk gDNA samples of 96 patients, and sequenced the amplicons in a paired-end run (**Figure 40B**). Demultiplexing the reads mapped the amplicons exclusively to the barcode combinations that were used in the experiment, in contrast to the 18 additional "dummy" barcodes that were not part of the experiment, which received negligible number of reads, probably due to index switching. This proved the robustness of our barcode design, and specific enrichment of the target loci using the assembled multiplex primers.

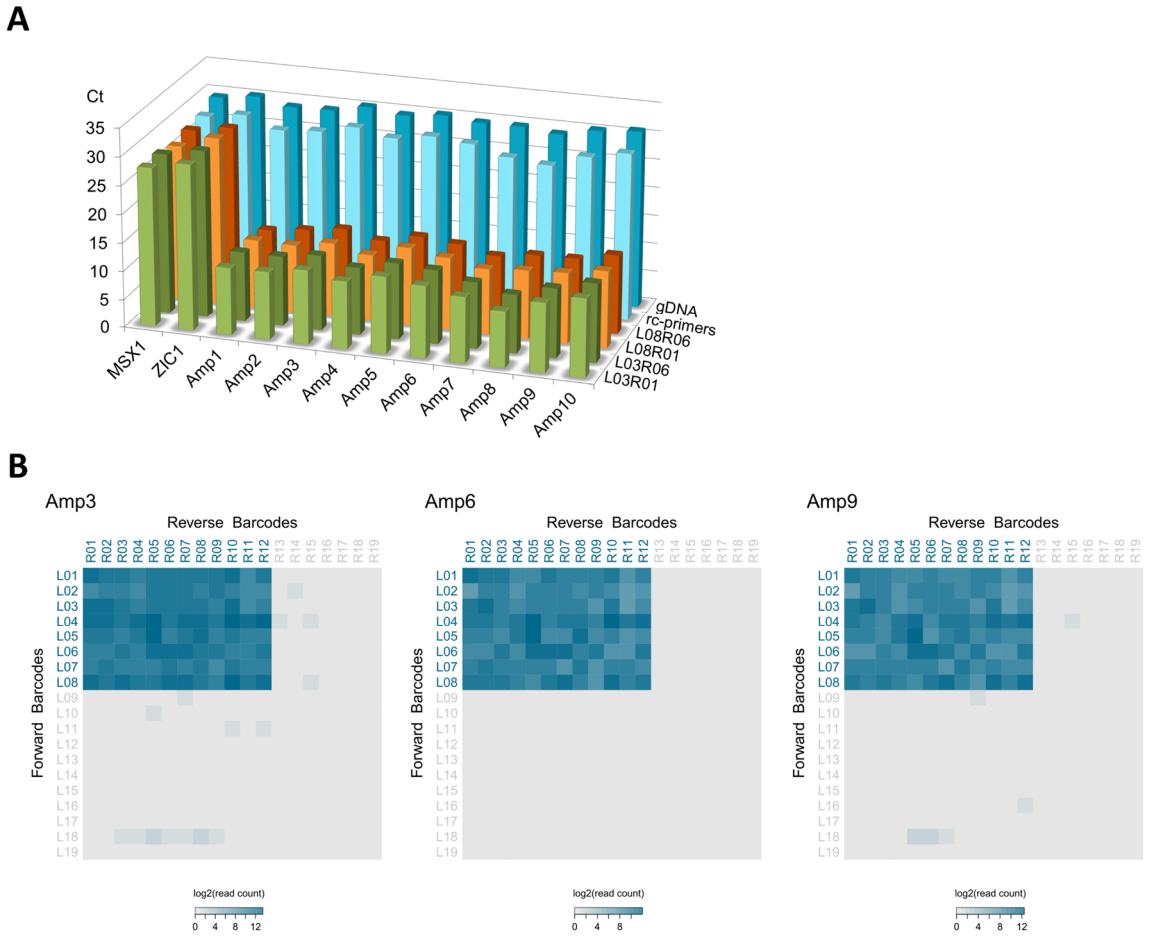



**Figure 40: Enrichment of genomic targets, assessed by qPCR and NGS. (A)** 10 loci in *BRCA1* and *BRCA2* genes were co-amplified from bulk gDNA templates using genotyping primers and four different barcode combinations (L03/L08 × R01/R06), and the amplicons were assessed by qPCR using nested primers. Different primer pairs and barcode combinations exhibited homogenous amplification. Non-pre-amplified gDNA, non-barcoded rc-primers, and non-targeted loci (*MSX1* and *ZIC1*) were negative controls. Melt curves showed unique peaks for all the targets (not shown). **(B)** 10 targets were co-amplified from 96 bulk gDNA samples, using 96 different barcode combinations (L01-L08 × R01-R12), and sequenced. Heatmaps show the number of amplicons assigned to three out of ten loci (Amp3, Amp6, and Amp9). L09-L19 and R13-R19 were control barcodes for demultiplexing, which were not used in the experiment. The sample-to-sample heterogeneity, which has a similar pattern in each amplicon, is mainly caused by different template concentrations.

### 3.2.2 RNA quantification

Because one of the main objectives of this project was to implement the BART-Seq for targeted transcriptomics, I next carried out experiments starting with RNA samples and tested whether we can quantify the abundance of RNA targets accurately.

#### 3.2.2.1 Pluripotency primer set

I initially confirmed that the assembled primers can enrich the RNA targets just like the genomic targets. For this, I designed a set of 15 primer pairs to amplify selected loci (5 of which exon spanning) from pluripotency and control gene

transcripts, as well as 4 exogenous RNA spike-in molecules (**APPENDIX J**). Nested qPCR validation of the amplicons derived from hESC cDNA indicated comparable enrichment, which were accompanied with unique melt curve peaks indicating specific amplification (**Figure 41**).

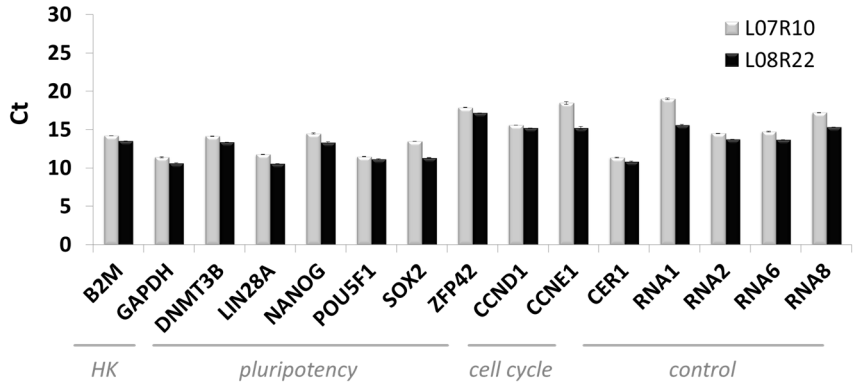



**Figure 41: qPCR evaluation of the pluripotency primer set.** Primers targeting 11 transcripts (pluripotency, housekeeping (HK), and cell cycle) and 4 RNA spike-ins were assembled with two different barcode combinations (L07×R10 and L08×R22, inefficient and efficient respectively) and used for pre-amplification of hESC bulk cDNA templates (100 pg/µl). Shown are the Ct values of pre-amplified samples assessed by nested primers, and error bars represent standard deviation of duplicates. Melting curves had unique peaks for all the targets (not shown)

#### 3.2.2.2 Quantifying transcripts from bulk RNA

Next, I asked how well BART-Seq performs in relative quantification of RNA templates. To test this, I prepared four-fold dilution series of bulk RNA isolated from H9 hESCs, combined with constant concentration of spike-ins, and aliquoted into 9 replicate wells for reverse transcription. Negative control wells contained only water and spike-ins. Each replicate well received primers assembled with different barcode combinations (**Figure 42A**). Following sequencing, coefficients of variation among the equimolar replicates were very low (std/avg <25%) even though they were at picogram levels (**Figure 42B**), indicating that the technical variations were minimal in the workflow. Only exceptions were a few genes in the lower end of the dilution series (i.e. 4 pg) due to very low averages, such as *CER1*, which is marginally expressed in undifferentiated cells.

Second, fitted regression models to the read counts and template concentrations had remarkably high correlations, with coefficients of determination ($R^2$) 0.96 on average (**Figure 42C**). A replicate experiment starting with bulk cDNA instead of RNA yielded very similar results (not shown). As an orthogonal validation, I analyzed the same cDNA samples with qPCR, and compared the results directly with the raw sequencing reads, which showed linear trends between the two methods (average $R^2$ of 0.99) (**Figure 42D**). Overall, these results demonstrated that the synthesized primers can co-amplify multiple RNA targets linearly for sequencing-based quantification with high precision.

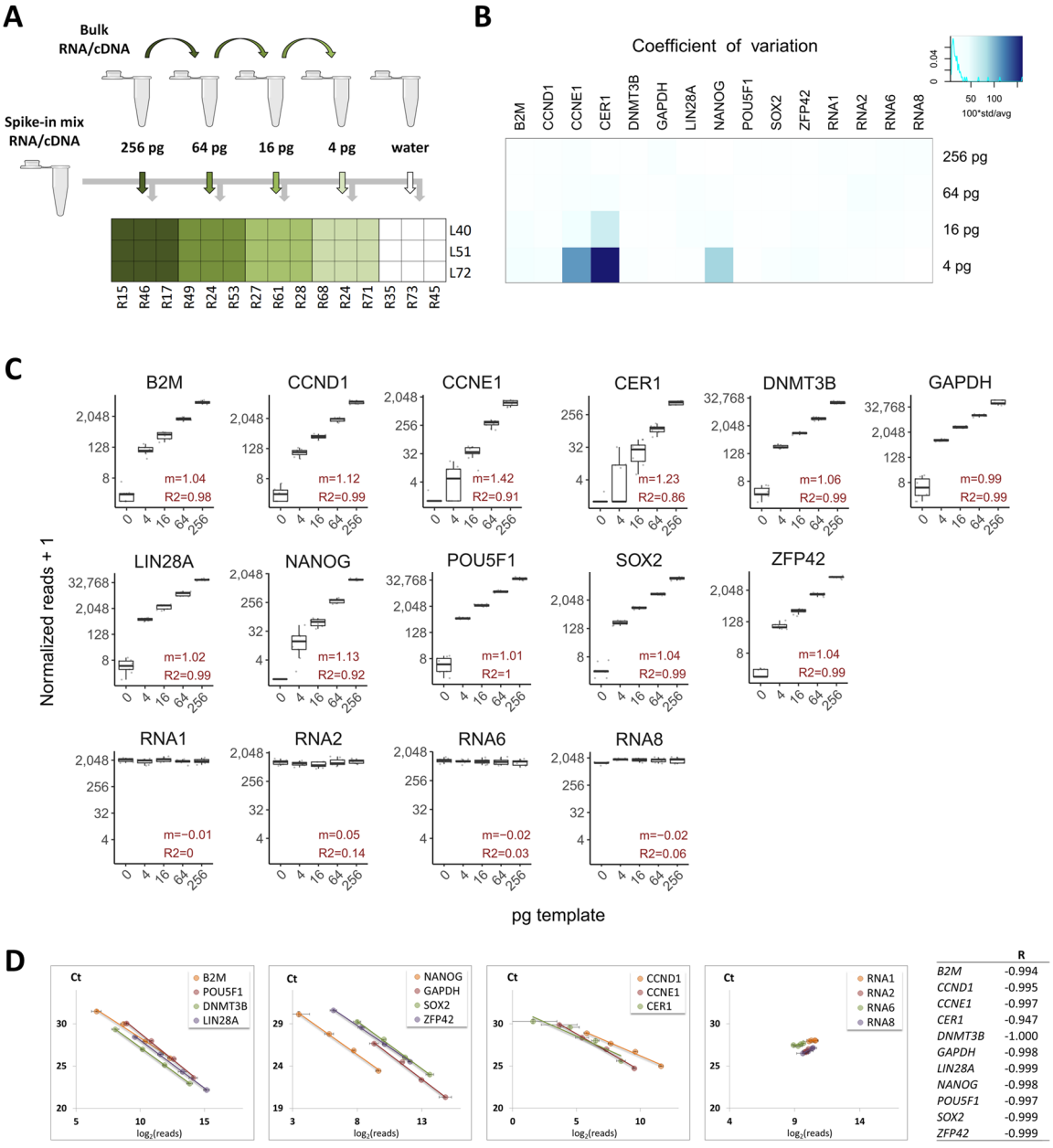



**Figure 42: Quantification of transcripts in isolated bulk RNA samples. (A)** Four-fold serial dilutions of bulk RNA isolated from hPSCs (Kunze et al., 2018) were combined with constant amount of spike-in RNA mixture, aliquoted into 9 replicate wells (4-256 pg/well), and reverse transcribed, each of which is indexed with a different barcode combination. Water mixed with spike-ins was a negative control. **(B)** The coefficients of variation (standard deviation divided by the average) of the normalized reads obtained from the RNA dilutions in **A,** calculated for the groups of nine samples with identical template concentration, had an average of less than 25%. The experiment was repeated by reverse transcribing the bulk RNA and spike-in mixture separately and combining respective bulk cDNA dilutions with spike-in mix cDNA, with very similar results (not shown). **(C)** Boxplots showing normalized read counts per target, plotted against template concentration. Coefficients of determination ($R^2$) were higher than 0.96 on average for the linear regression models fitted to the 4-256 pg sample groups. **(D)** cDNA dilution series was analyzed in parallel by qPCR. The plots show the correlation of BART-Seq results (average log2 read counts) with the qPCR (average Ct values). Vertical error bars represent three qPCR replicates, and horizontal error bars represent nine sequencing replicates. R values on the right are the corresponding coefficients of correlation per gene

Finally, I compared the quantitative power of BART-Seq with the contemporary global single-cell sequencing techniques (scRNA-Seq) reviewed by Ziegenhain et al. (2017). They used the known concentrations of 92 spike-ins and read counts thereof to calculate the linear model fits and the $R^2$ values. Similarly, I used average read counts of 11 genes across the count matrix instead of the known concentrations to calculate linear regressions per sample, as if each gene was a spike-in with a fixed concentration. BART-Seq results exhibited an outstanding accuracy (median $R^2$ of 0.98), placing it above all the single-cell transcriptomics methods reviewed (**Figure 43**).

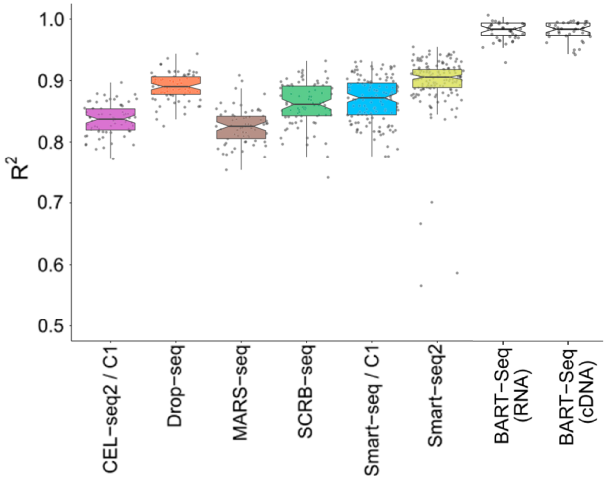
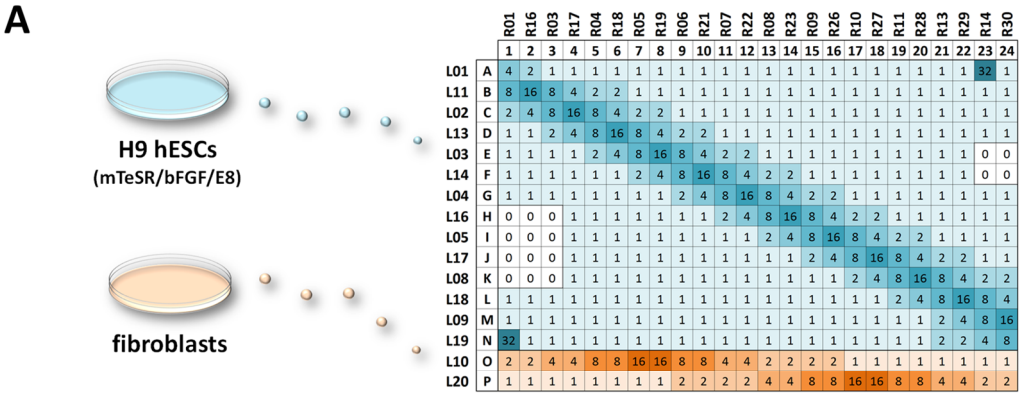
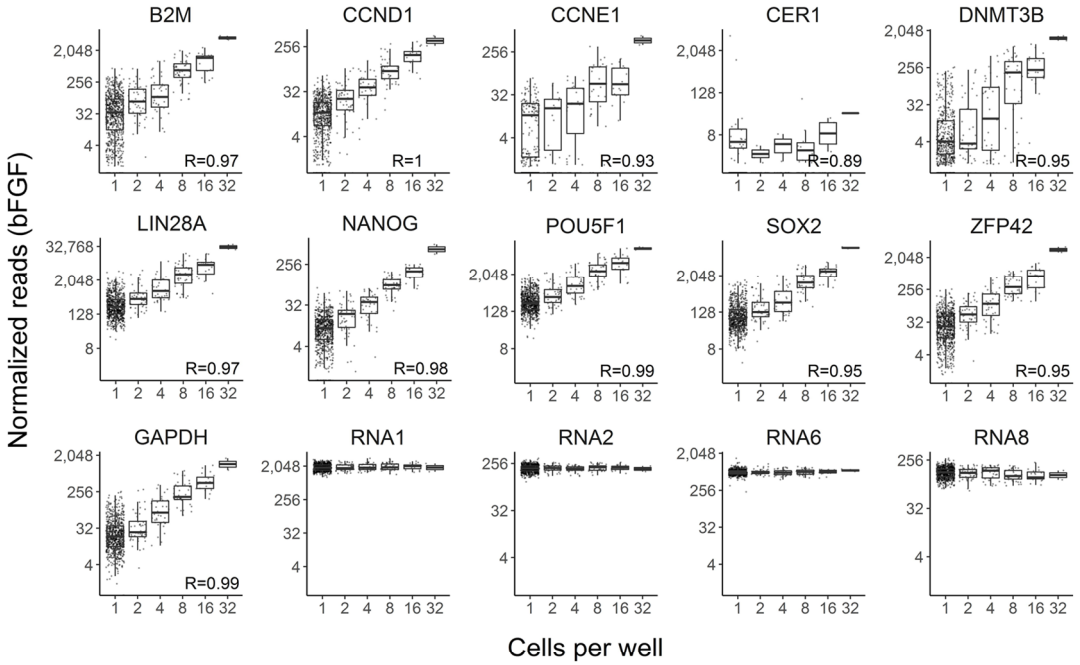



**Figure 43: Accuracy of BART-Seq as compared to other scRNA-Seq methods.** A plot adapted from Ziegenhain et al. (2017), displaying the adjusted $R^2$ values of linear regression models calculated using 96 ERCC spike-in expression values obtained using different global transcriptomics methods. The regression models calculated for BART-Seq samples using the average read counts of 11 genes across the bulk RNA dilution experiment (**Figure 42**) had a median $R^2$ value of 0.98

### 3.2.2.3 Quantifying transcripts from cells

Having confirmed its accuracy to quantify isolated bulk RNA samples, I next assessed BART-Seq in measuring the transcripts directly from cells. I analyzed the hESCs sorted into 384-well plates with two-fold increments (1-32 cells per well), and correlated the transcript counts with the number of cells. The plates were pre-filled with the reverse transcription (RT) reaction mix that contained four RNA spike-ins (**Figure 44A**), and BJ fibroblasts were included as control. The number of sorted cells per well and the corresponding read counts showed very high correlations, while the spike-in values remained constant (**Figure 44B**). Single cells, as expected, had the highest transcriptional heterogeneity. Analyzing the same set of genes in sorted fibroblasts showed a clear distinction of their expression profiles (**Figure 44C**), despite receiving some reads of the pluripotency genes, which should be predominantly caused by index switching (Sinha et al., 2017). Considering that only *DNMT3B* primers was exon spanning among the pluripotency targets, the others may be residually amplified from the gDNA of fibroblasts as well. Taken together, these analyses show that BART-Seq can be used for directly analyzing gene expression in numerous single cells and produce quantitative results within a broad dynamic range.


A
H9 hESCs
(mTeSR/bFGF/E8)
fibroblasts
R01 R16 R03 R17 R04 R18 R05 R19 R06 R21 R07 R22 R08 R23 R09 R26 R10 R27 R11 R28 R13 R29 R14 R30
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
L01 A 4 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 32 1
L11 B 8 16 8 4 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
L02 C 2 4 8 16 8 4 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
L13 D 1 1 2 4 8 16 8 4 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1
L03 E 1 1 1 1 2 4 8 16 8 4 2 2 1 1 1 1 1 1 1 1 1 1 0 0
L14 F 1 1 1 1 1 1 2 4 8 16 8 4 2 2 1 1 1 1 1 1 1 1 0 0
L04 G 1 1 1 1 1 1 1 1 2 4 8 16 8 4 2 2 1 1 1 1 1 1 1 1
L16 H 0 0 0 1 1 1 1 1 1 1 2 4 8 16 8 4 2 2 1 1 1 1 1 1
L05 I 0 0 0 1 1 1 1 1 1 1 1 2 4 8 16 8 4 2 2 1 1 1 1 1
L17 J 0 0 0 1 1 1 1 1 1 1 1 1 1 1 2 4 8 16 8 4 2 2 1 1
L08 K 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 2 4 8 16 8 4 2 2
L18 L 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 4 8 16 8 4
L09 M 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 4 8 16
L19 N 32 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 4 8
L10 O 2 2 4 4 8 8 16 16 8 8 4 4 2 2 2 2 1 1 1 1 1 1 1 1
L20 P 1 1 1 1 1 1 1 1 2 2 2 2 4 4 8 8 16 16 8 8 4 4 2 2


B
B2M
CCND1
CCNE1
CER1
DNMT3B
LIN28A
NANOG
POU5F1
SOX2
ZFP42
GAPDH
RNA1
RNA2
RNA6
RNA8
R=0.97
R=1
R=0.93
R=0.89
R=0.95
R=0.97
R=0.98
R=0.99
R=0.95
R=0.95
R=0.99
Normalized reads (bFGF)
Cells per well

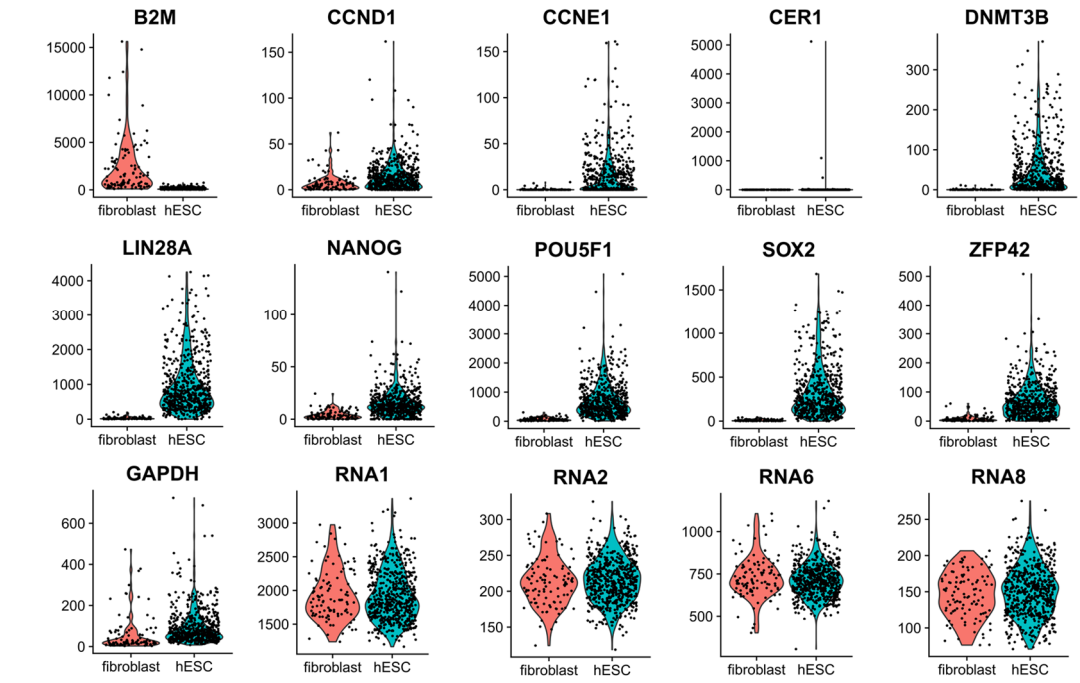

C
B2M
CCND1
CCNE1
CER1
DNMT3B
LIN28A
NANOG
POU5F1
SOX2
ZFP42
GAPDH
RNA1
RNA2
RNA6
RNA8
fibroblast
hESC

**Figure 44: Quantification of transcripts directly from cells. (A)** Part of the barcode matrix used for analyzing single (1) and multiple (2-32) hESCs maintained on different media (mTeSR™1, KSR-bFGF, and E8), and BJ fibroblasts. Negative control wells (0) did not receive any sorted cells. Prior to sorting, all wells (including negative controls) were pre-filled with 2 µl of RT mixture containing fixed concentrations of four RNA spike-ins. Over 4500 wells representing two biological replicates were analyzed as two libraries, and sequenced using Illumina NextSeq for a total of 23.5 million processed paired reads. **(B)** Normalized read counts plotted against the number of cells sorted per well (n=858 samples from KSR-bFGF medium are shown). Correlation coefficients (R) between the cell counts and the median of corresponding reads are shown. **(C)** Violin plots illustrating gene expression differences between hESCs and fibroblasts. Samples include single cells and calculated 1-cell values of multi-cell wells. Fibroblasts had higher *B2M* expression (Drukker et al., 2002), whereas pluripotency and cell cycle genes had notably higher expression in the hESCs

### 3.2.3 Single-cell analyses

After confirming that the BART-Seq is suitable for targeted quantitative single-cell transcriptomics, I used it to investigate specific processes related to self-renewal or lineage commitment of human pluripotent stem cells (hPSCs), which are summarized in this section.

#### 3.2.3.1 Influence of maintenance media on the pluripotency state of hESCs

There are many commercial and homemade formulations to maintain the hPSCs in a self-renewing state, as described in the **Introduction**. I hypothesized that they might endow the hPSCs with different flavors of pluripotency due to diverse sets of constituents (**Table 2**), which may stimulate the signaling pathways differently. To test this, I measured the expression levels of the core pluripotency network of transcription factors using BART-Seq in single H9 hESCs, which were cultured on mTeSR™1, KSR-bFGF, or E8 at least five passages, using the pluripotency primer set (**APPENDIX J**).

Non-linear dimensionality reduction (UMAP) revealed two major subpopulations of cells exhibiting naïve-like -*NANOG*$^{HIGH}$ *ZFP42* (*REX1*)$^{HIGH}$- and primed-like -*LIN28A*$^{HIGH}$ *DNMT3B*$^{HIGH}$- profiles (**Figure 45**) (Pastor et al., 2016; Theunissen et al., 2014; Warrier et al., 2017; Zhang et al., 2016). The cell cycle-related gene *CCND1* strongly correlated with *NANOG* (and *ZFP42*), although it is not a known marker for the ground-state pluripotency. Remarkably, mTeSR™1-treated cells were found primarily in the primed-like cluster, whereas majority of the E8-treated cells were localized to the ground state-like (naïve) cluster; suggesting that these growth conditions shifted the hESCs along the pluripotency axis.

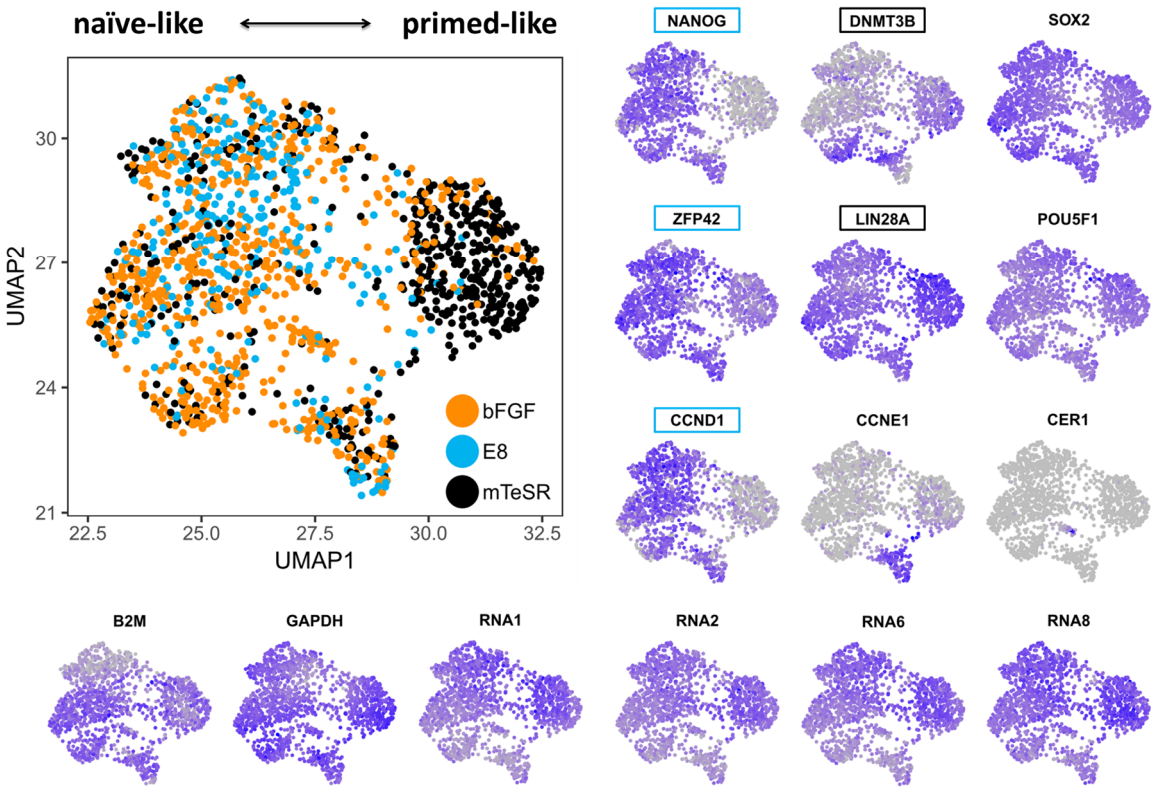
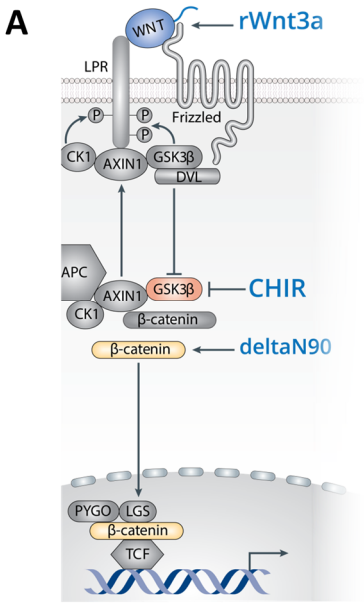



**Figure 45: Transcriptional profiles of single hESCs cultured on different media.** UMAP visualization of single hESCs (n=1550) treated with three media (● mTeSR™1, ● bFGF, ● E8). Expression of the genes underlying the distribution (11 genes) and spike-ins are shown. E8- and mTeSR™1-cultured cells formed distinguishable clusters resembling naïve (*NANOG*, *ZFP42*) and primed (*DNMT3B*, *LIN28A*) pluripotent states, respectively. The results are based on two biological repetitions

### 3.2.3.2 Stimulation of the Wnt pathway in hESCs

Wnt signaling has key functions in gastrulation and early lineage segregation events in the developing embryo, as described in the **Introduction**. Although numerous *in vitro* differentiation protocols involve activation of the pathway with different stimulants (Lindsley et al., 2006; Loh et al., 2016), it is not clear whether they truly mimic the canonical pathway. Therefore, I wanted to test the hypothesis that chemical inhibitors of GSK3 or ectopic expression of ß-catenin can substitute the ligands of the Wnt pathway (Barth et al., 1997; Blauwkamp et al., 2012). To this goal, I analyzed hESCs differentiated using recombinant Wnt3a protein, treatment with a chemical inhibitor of GSK3 (CHIR99021), or via Dox-induced ectopic expression of the constitutively active ß-catenin (ΔN90-ß-catenin), for 22 targets including early gastrulation and housekeeping genes, and 3 RNA spike-ins, at 0, 24 and 72 hours of treatment (**Figure 46**, **APPENDIX K**). The validation of the primers with nested qPCR and BART-Seq can be found in **APPENDIX B**.




B

| Target genes |
|---|
| **lateral mesoderm:** *EOMES, HAND1, MESP1* |
| **paraxial/(pre-)somitic mesoderm:** *CDX2, HOXA1, MSGN1, PAX3, TBX6* |
| **neural crest:** *ZIC1, MSX1* |
| **pan-primitive streak:** *EVX1, GSC, MIXL1, T* |
| **anterior primitive streak:** *NODAL* |
| **endoderm:** *FOXA2, SOX17* |
| **control:** *GAPDH, NANOG*, 3 RNA spike-ins |

**Figure 46: Stimulation of the Wnt/β-catenin pathway at different stages of the cascade. (A)** hESCs were treated by recombinant Wnt3a or CHIR99021 (CHIR), or a transgenic hESC line was induced with Doxycycline to induce the expression of ß-cateninΔN90 for 72 hours. Single cells were sampled at 0, 24, and 72 hours. A total of 4324 cells from three biological replicates were analyzed in a single Illumina NextSeq mid-output run. **(B)** The primer set used for the analysis (**APPENDIX K**)

Initially, I inspected the expression of the same gene set in the global sequencing results of bulk RNA following 72 hour of stimulation, and observed a striking similarity between ΔN90-ß-catenin and CHIR99021, which differed from the rWnt3a treatment. Single-cell data from BART-Seq analysis corresponding to the same time point showed remarkable resemblance to the global RNA-Seq results notwithstanding a significant degree of cellular heterogeneity (**Figure 47**).

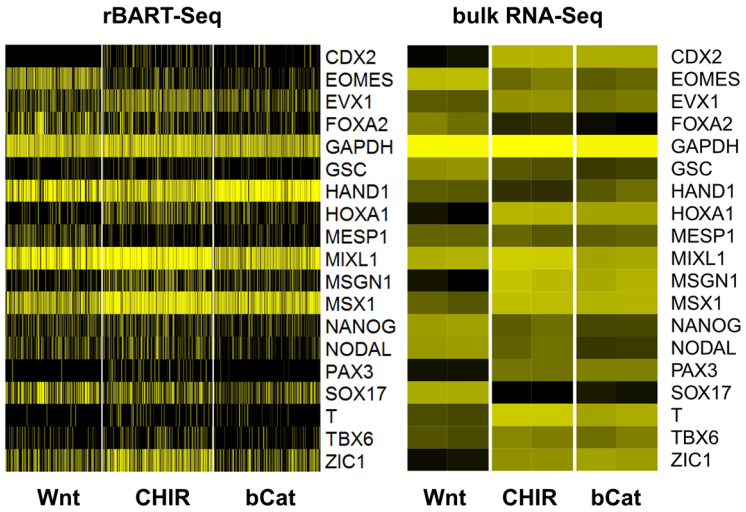



**Figure 47: Comparison of the BART-Seq results with the bulk RNA-Seq results.** Heatmaps of the expression of 19 genes after 72 hours of treatment with the three inducers of the Wnt pathway. **Left**: single cells analyzed with BART-Seq. **Right**: TPM values (Transcripts per Million) obtained from bulk RNA-Seq analysis of a replicate experiment (two independent replicates per condition). Very similar patterns were observed between the BART-Seq results and bulk RNA-Seq

Next, I analyzed the pairwise gene correlations in single cells that were stimulated for 24 hours, and observed that they were grouped in two major clusters exhibiting *NANOG*, *NODAL*, *EOMES*, *FOXA2* and *MESP1*, *MSX1*, *SOX17*, *ZIC1*, *TBX6*, *HOXA1*, *HAND1*, *MSGN1* gene signatures, respectively (**Figure 48**, left). They reflected the emergence of two cell subpopulations at 24 h in the dimensionality reduction (tSNE) analysis (**Figure 48**, right; **APPENDIX C**), which likely correspond to the proximal and the distal region of the embryo, respectively, as indicated by the topography of the expression of orthologous genes in the mouse embryo[14]. Pan primitive streak markers *GSC*, *EVX1*, *MIXL1* correlated with both groups, while *MIXL1* was expressed at a higher level in the distal-like group. With respect to the influence of different stimulations of the Wnt/ß-catenin pathway, distinct clusters were observed after 72 hours, and rWnt3a treatment exhibited definitive endoderm-like and lateral plate mesoderm-like cells, with *FOXA2*$^{HIGH}$ *SOX17*$^{HIGH}$ and *HAND1*$^{HIGH}$ *MESP1*$^{HIGH}$ *EOMES*$^{HIGH}$ profiles, respectively. The latter population dominated the rWnt3a progeny in a replicate experiment (**APPENDIX D**). Taken together, CHIR99021 seems to limit the diversity of primitive streak-like progeny that differentiates from hESCs compared to the ligand of the pathway Wnt3a, an effect that was also validated using constitutively active ß-catenin.

### 3.2.4 Bulk analyses

Besides single-cell transcriptomics, BART-Seq is applicable also to the analysis of bulk RNA or gDNA samples. This section summarizes the two proof of concept projects I contributed that involved analysis of bulk samples using BART-Seq.

#### 3.2.4.1 Genotyping the patients for *BRCA* mutations

Because BART-Seq workflow can enrich genomic targets, we hypothesized that we should be able to detect mutations if they are located within the loci targeted by BART-Seq primers (**Figure 49**). *BRCA1* and *BRCA2* are breast and ovarian cancer susceptibility genes with a strong hereditary component. The Jewish Ashkenazi population in Israel is a carrier of 10 founder mutations in *BRCA1* and *2*, which reside within the loci targeted by the primer sets we had designed for developing and optimizing the barcode assembly method (Kaufman et al., 2006; Laitman et al., 2012; Lerer et al., 1998) (**APPENDIX L**). We received genomic DNA samples obtained from 96 breast cancer patients of Jewish Ashkenazi descent that have been previously tested for a panel of 10 hereditary mutations by Sanger sequencing and other conventional assays. To test our hypothesis, we co-amplified 10 *BRCA1* and *BRCA2* loci from these samples using a matrix of 12 forward and 8 reverse barcodes, and analyzed the pooled amplicons with NGS.

[14] http://www.picb.ac.cn/hanlab/itranscriptome/Home/

When we compared the genotyping results of all 1920 multiplexed alleles (spanning 10 amplicons from 96 patient samples with 2 alleles each), we saw that 95 out of 96 patients (~99%) were classified in agreement with the results of the clinical lab (**Figure 49B**). In the misclassified sample, the expected mutation (Mut2) was detected together with an unexpected mutation (Mut7). When I analyzed this sample with Sanger sequencing for these two mutations, only Mut2 was present, in agreement with the known genotype (**Figure 49C**), indicating a false positive. Because this experiment was demultiplexed with the earlier pipeline, which used BLAST for identifying amplicons, the allowed mismatches might partially have hindered true identification of the mutations. Hence, using string matching for the exact mutation region (e.g. 5-10 nucleotides before and after the mutation) can enhance the genotyping rates for future analyses. Collectively, these results showed that BART-Seq performs well also in targeted genomics applications.

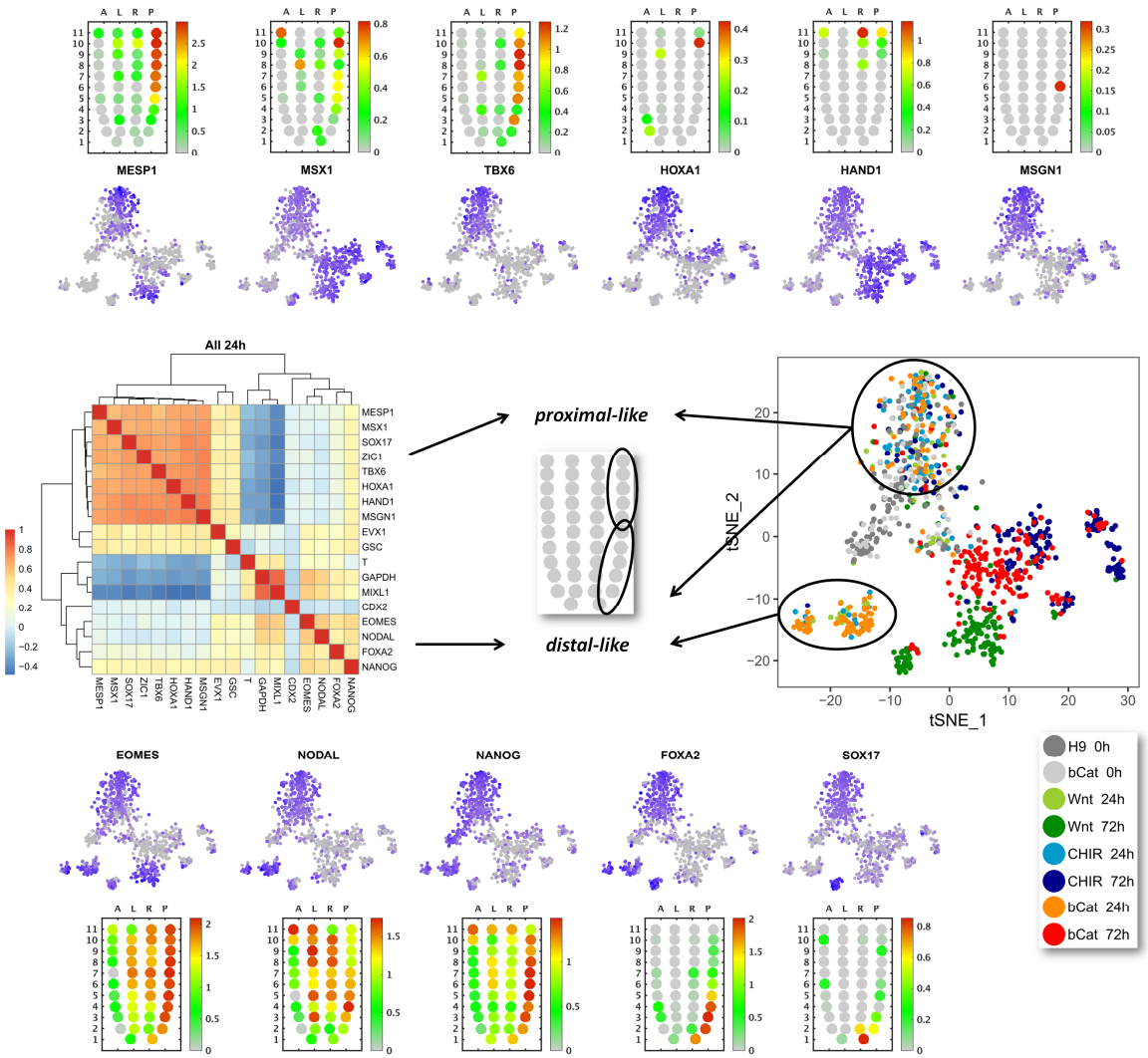



**Figure 48: Cell populations that emerge upon stimulation of the Wnt/β-catenin pathway.** A heatmap of the pairwise gene correlations in single cells after 24 h from the three treatments (left) and two-dimensional representation (tSNE) of all the single cells sampled at 0, 24, and 72 h, based on the expression of 19 genes (right). Expression of selected genes underlying the tSNE plot is shown in the upper and lower panels (additional data in **APPENDIX C**). The corn plots were derived from the

iTranscriptome database[15] that represent the localization of the same transcripts in epiblast stage mouse embryos (E6.5-E7.5)

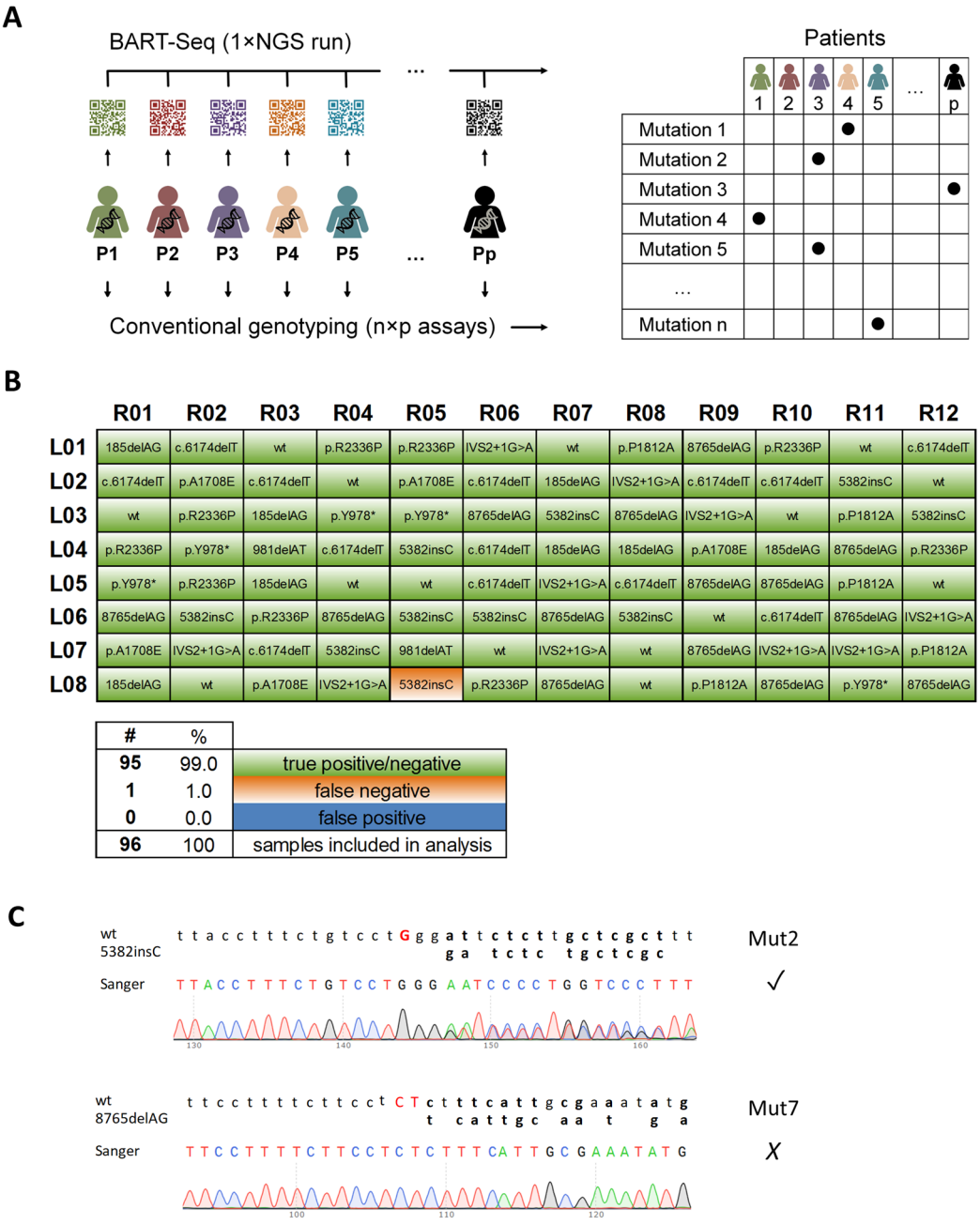



**Figure 49: Genotyping cancer patients using BART-Seq. (A)** Schematic representation of the application of BART-Seq for genotyping to replace mutation-specific assays. **(B)** Genotypes of 96 breast cancer patients corresponding to 10 *BRCA1* and *BRCA2* mutations. Correspondence of BART-Seq results to the known genotypes is marked by green sectors (true positives), and the statistics is provided below. **(C)** Although BART-Seq detected two mutations in one of the samples, Sanger sequencing identified only the mutation 5382insC (Mut2) but not the 8765delAG (Mut7)

[15] http://www.picb.ac.cn/hanlab/itranscriptome/Home/

### 3.2.4.2 Compound screening on hepatocytes

Because BART-Seq is capable of analyzing tens of transcriptional targets in hundreds to thousands of samples, we hypothesized that it should be suitable also for pharmaceutical applications, for example for screening the transcriptional response of cells to a compound library. To this goal, we collaborated with the pharmaceutical company AstraZeneca to screen the dose responses of human iPS-derived hepatocytes to a library of agonists using BART-Seq. We screened 17 selected genes in the bulk cDNA samples synthesized from lysates of cells treated with 3-fold dilutions of compounds. We analyzed the same cDNA samples with qPCR in parallel. Even though this preliminary experiment partially failed due to incomplete lysis of the cells (because the spike-ins had rather consistent results), we could obtain dose-response curves that closely paralleled the qPCR results for some of the genes and compounds, as shown in **Figure 50**. This pilot project showed that BART-Seq can be adapted to a variety of research applications.

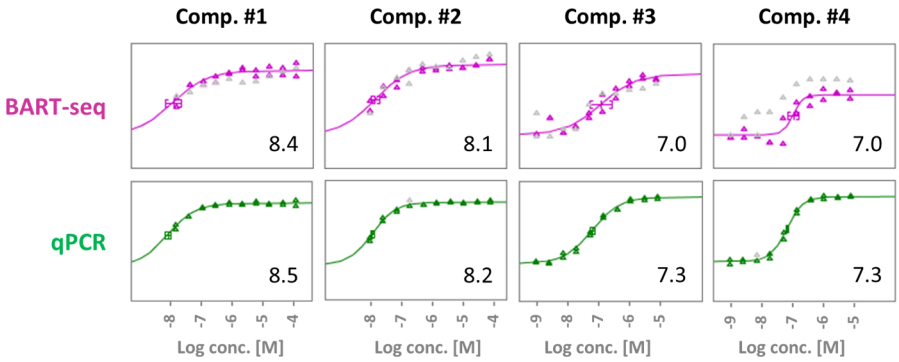



**Figure 50: Compound screening on hepatocytes using BART-Seq.** We analyzed the cDNAs derived from hiPSC-derived hepatocytes treated with three-fold dilution series of a compound library, using BART-Seq and qPCR in parallel. Sample results show the response of *ABCG1* gene to four of the compounds. BART-Seq plots are based on log-normalized counts and qPCR results are based on ΔΔCt (% of control) values

# 4 DISCUSSION

It was not long after the introduction of the NGS technology that the first single-cell RNA-Seq (scRNA-Seq) analysis was published (Tang et al., 2009). Over the past ten years the number of publications that accommodate single-cell experiments has grown exponentially, as well as the number of cells that can be analyzed in parallel (Svensson et al., 2017) (**Figure 5**). Although a variety of scRNA-Seq techniques exist today, essentially all have certain limitations, which converge mainly upon two major issues: shallow coverage (depth) and high costs. One of the reasons of shallow coverage is the attempt of global (unbiased) techniques to sequence the whole transcriptome in a very large number of cells. As a result, moderately or lowly expressed genes receive either zero or very small number of reads given a fixed sequencing capacity (L. Lun et al., 2016), while many reads are consumed by housekeeping and uninteresting genes. This can be partly circumvented by sequencing samples with much greater depth, albeit with considerably higher costs.

The central premise of my thesis is that the coverage and quantitative accuracy per gene should increase if the same sequencing capacity could be focused on a small number of genes. To achieve this in a way that is compatible with single cell applications and for a reasonable cost, I established a workflow (named BART-Seq) that combines enrichment of pre-selected transcripts or genomic loci using barcoded primers for multiplex PCR, and NGS. In parallel, I developed customized computational methods to analyze the sequencing data generated using the technique.

Because BART-Seq is a novel targeted sequencing technique, it is important to discuss here questions about its comparability to other approaches, to analyze its pros and cons, possible improvements, and the biological lessons learnt. Specifically, I address a series of questions outlined in the sections below starting with specific insights about the path I took during the development phase.

## 4.1 Development and Optimization of the BART-Seq Workflow

During the development of BART-Seq, I placed special emphasis and diligence in optimizing the individual reaction steps, which resulted in the end an easy to follow and robust workflow that can readily accommodate single-cell analysis.

### 4.1.1 Barcode assembly

With the purpose of ensuring full conversion of oligonucleotide building blocks to single-stranded barcoded primers, the initial parameter I tested was the duration of the assembly steps. By gradually increasing the duration of Klenow reaction I found that the efficiency reaches to a maximum after 60 minutes (**Figure 20**), and the

Lambda treatment had the maximum at 30 minutes (**Figure 22A**). I refrained from longer incubations to avoid possible degradation of the newly synthesized primer ends by Klenow by its 3'→5' exonuclease activity[10], and trimming of the barcodes by Lambda exonuclease with its residual activity towards 5'OH ends (Little, 1967). A notable observation was the essence of removing the anti-sense primers, as their presence reduced the efficiency of PreAmp PCR up to tens of folds and resulted in non-specific amplification (**Figure 21D**), possibly by reducing the availability of sense primers in the reaction (Kapley et al., 2000). Using an exonuclease for anti-sense primer removal required different oligonucleotide modifications, i.e. T7 was used with barcodes with 6 consecutive phosphorothioate bonds to prevent degradation and Lambda was used with anti-sense primers with 5'P ends to promote their hydrolysis (**Figure 21A**). The cost for these additional modifications is comparable and the two enzymes exhibited similar efficiencies; however, the distortion of melt curves with T7 and the inability to heat-inactivate it[16] prompted me to proceed with the Lambda exonuclease.

I also took measures to maintain the intactness of barcode sequences through the assembly reactions, which would eventually cause read cross-contamination due to the trimmed sequences. Screening all possible trinucleotides (NNN) flanking the 5' of barcodes revealed that CCA- had the highest resistance among all the 64 sequences (**Figure 23C**). A previous study also reported lower degradation rates of GC-rich DNA by Lambda exonuclease compared to AT-rich DNA (Foulk et al., 2015). Interestingly, our data contained many GC-rich trinucleotides with low frequency (**Figure 11**), whereas a consistent pattern was the high frequency C nucleotide as the first base, which might be the key position for resistance against degradation. Based on these results, I flanked all the BART-Seq barcodes by the 5'CCA protection group. A validation of the efficacy of this modification was the increase in genotyping rates from 96% to 99% when the same samples were analyzed using the barcodes without and with the CCA- group (**Figure 49**).

### 4.1.2 Reverse transcription

The major cause of the dropout events in single-cell experiments is the failure to capture and reverse transcribe a fraction of mRNA molecules, which may cause incorrect designation of a gene's expression status (Stegle et al., 2015). Whereas the target mRNA molecules are significantly enriched with the BART-Seq protocol, their inadequate capturing might weaken the linearity of quantification. Hence, when adapting the RT protocol for cost-efficiency and practicality, I ensured that the efficiency of reverse transcription remains the same. For instance, the workflow involves snap-freezing the cells together with the reverse transcription reagents to lyse the sorted cells, which could damage the enzyme (Cao et al., 2003). I verified that this step did not reduce the efficiency of RT reaction in my protocol (**Figure 25**), possibly because either freezing does not damage this enzyme, or there is still

[16] https://international.neb.com/products/m0263-t7-exonuclease#Product%20Information

plenty of the intact enzyme left in the reaction to process small template amounts (down to single cell levels). Using 25% of the recommended concentration of reverse transcriptase did not affect the efficiency at all (**Figure 25**). This adjustment reduced the overall cost of large-scale experiments substantially. A significant improvement of RT efficiency was evident when I used the newer version of the enzyme (Superscript IV), which should be preferred for the future experiments.

To reduce the number of steps in the protocol, I omitted the RNase H treatment following RT. Even though the presence of RNA:DNA heterodimers could potentially reduce the PCR efficiency, I did not observe any difference with the treated samples, except for one gene (**Figure 24**). On the other hand, adding the RNase H to the PCR reagents (even in lower concentrations) and running an additional cycle before PCR could restore the efficiency for the affected gene. Hence, this optional step can be considered for the future single-cell experiments that involve lowly expressed genes.

It is essentially possible to use the barcoded-primers also for priming the reverse transcription. A preliminary experiment I conducted suggested that RT and PCR could be run successively within the same mixture (RT+PCR) including components of both reactions (**APPENDIX A**). Although I did not run further experiments in this regard within this project, this one-step reaction could potentially be optimized further to achieve similar efficiencies with separate reactions.

### 4.1.3 Pre-amplification PCR

Within the frame of BART-Seq, the number of multiplexed primers (tested up to 10) did not influence the efficiency of PreAmp PCR (**Figure 26**) as long as the concentration of individual primers are kept in the optimal range, which I found to be 0.025-0.030 μM final (which translates to 0.25-0.30 μM in the Klenow reaction) (**Figure 19**). Nevertheless, for larger primer sets reducing individual primer concentrations can be considered to prevent overcrowding, which might lead to non-specific amplification products or concatemers. Even though I explored alternating the annealing temperature (Ta) in the early and late phases of the PCR to enhance specific amplification (to account for the hybridization status of barcode-linker sequences) it did not seem to have an influence on the efficiency (**Figure 18A**).

As the BART-Seq workflow entails addition of PCR reagents directly on top of the RT reaction, another concern to be addressed with respect to efficiency was the amount of reverse transcriptase carried over into the PCR, as it can inhibit the amplification by remaining attached to the newly synthesized cDNA (Chandler et al., 1998; Chumakov, 1994). While gradually increasing the ratio of RT reaction volume within the PCR (RT/PCR) did not cause any reduction in the efficiency of Platinum MM; QiaGen MM exhibited decreasing trends (**Figure 29**). Despite both master mixes contain Taq polymerase, modifications of the enzymes or different ingredients of each master mix could be the reason of this difference. Besides, the Platinum MM had several folds higher efficiency (**Figure 27**, **Figure 29**) and ~20%

lower cost compared to the QiaGen MM, hence, I decided to integrate it into the workflow.

Because the multiplex PCR MM is one of the most expensive reagents of the workflow, I reduced the concentration of MM down to 25% of the manufacturer guidelines, considering that the enzyme should not be the rate limiting component for the single-cell experiments. Notably, this did not decrease the quality or robustness of amplification when I supplemented the reactions with $MgCl_2$ up to 6 mM final (**Figure 28**). Although $MgCl_2$ increases the processivity of the *Taq* polymerase to a certain extent if well-balanced with dNTP concentrations (Henegariu et al., 1997), special attention should be paid for the sequence-sensitive experiments, such as mutation screening, as the fidelity of the *Taq* polymerase is inversely correlated with the $Mg^{++}$ concentrations (Eckert and Kunkel, 1990).

Despite mixing the primers equimolarly in the experiments presented here, they had varied efficiencies even in the presence of same concentration of targets, as in the case of genomic loci (**Figure 20)**. An arduous measure would be adjusting the concentration of individual primers to compensate for this variation, as discussed by Henegariu et al. (1997). This can be considered for the cases where a few primers have much higher efficiency, or a few genes have much higher expression compared to the others and consume majority of the reads. Nevertheless, such an adjustment should be done cautiously in order to avoid reaching the plateau phase of PCR due to limiting primer concentrations, especially for the highly expressed genes (Kainz, 2000), which may hinder the quantitativeness of the method.

## 4.2 Bioinformatics

Each stage of the BART-Seq workflow goes hand in hand with computational components, such as designing the primers and barcodes, processing the raw reads to count matrices, and analyzing the data. While some of those were previously established, such as the primer designing tool, my project involved implementing majority of the analysis methods from scratch and assisting development of updated versions of the existing tools.

### 4.2.1 Primer design

PrimerSelect[17] is the custom-made open-access web tool developed in collaboration with the Institute of Computational Biology of Helmholtz Center Munich for designing multiplex primers for the BART-Seq method. It requires as the input the sequence of the regions that will be targeted by the primers and a configuration file (**APPENDIX H**) to specify parameters for the prospective primer set, such as melting temperatures, size range, GC content. Besides ensuring the compatibility of

[17] http://icb-bar.helmholtz-muenchen.de/primerselect

multiplexed primers, the tool forces the primers to end with an Adenine (A) base, because the Klenow polymerase template-independently adds an A to the newly synthesized strand (Clark, 1988). Accordingly, we order the reverse complementary primers excluding the 5'T, so that they are still complementary to their targets following the Klenow fill-in reaction (**Figure 17**, **Table 9, Appendices I-L**).

Designing primers for the genomic loci is easier because there is a single target sequence. On the other hand, selecting the target transcriptomic loci was a challenge, in particular for the genes that express several mRNA isoforms. Because my focus in this project was not to distinguish transcript isoforms, I aimed to select the regions that would capture majority of the expression from each gene. To achieve this, I implemented a method using bulk RNA-Seq data obtained from the cell types and differentiation stages that were relevant to my experiment. After aligning them to the human genome, I generated coverage maps that indicate the distribution of the reads to the genomic sequence regardless of the transcript variants, and selected the loci that receive the highest number of reads per gene for designing primers (**Materials and Methods**, section **2.4.3**).

### 4.2.2 Demultiplexing the sequencing reads

Due to the unique structure of BART-Seq libraries, in which dual barcodes are contained by the paired reads, we developed a tailor-made algorithm for translating the raw sequencing reads to count matrices, in collaboration with the Institute of Computational Biology of Helmholtz Center Munich. The initial demultiplexing pipeline (**Figure 30**) worked smoothly for the earlier experiments that were conducted on bulk templates, which generated fairly pure amplicons and good read qualities as they were analyzed on a MiSeq instrument.

After starting to analyze thousands of single cells, I preferred a NextSeq instrument in order to have higher coverage, as it can generate several folds more reads than the MiSeq. Yet, we experienced troubles merging the read pairs unlike the previous experiments. This turned out to be due to a significant reduction of base calling qualities towards the read ends (**Figure 31C**). Consequently, we decided to develop a more robust algorithm that would bypass this problem by processing the read pairs separately (**Figure 32**).

In the new version of the demultiplexing algorithm, we took into account several considerations that we either already knew in the earlier version or learned through trial and error during the analyses.

- Barcodes may contain single nucleotide errors when they are purchased, or may harbor mutations during the workflow or sequencing. Therefore, we allow mismatches in barcode identification to minimize false negatives. At the same time, we minimize barcode misidentification by excluding mismatches that could potentially arise from more than a single barcode to prevent read cross-contamination.

- To avoid missing the barcodes that might be trimmed by a few bases (i.e. at the CCA- protection group), we allow some positional flexibility instead of searching the barcode sequence at exactly the nucleotide positions 4-11. Using 8 nt barcode sequences instead (excluding the CCA-) as a reference did not make a significant difference in the percentage of mapping.
- Reads from short amplicons (< 172 bp) contain part of or complete second linker and barcode at the other end, which may preclude their proper alignment with the reference amplicons. Accordingly, we identify and trim those sequences before the alignment.
- We apply quality trimming to remove the bases with decreased qualities at the read ends (e.g. the NextSeq experiments) (**Figure 31C**).
- Some of the reads might contain junk sequences that consist of primer concatemers or non-specific PCR products (**Figure 31B**). Thus, we enforce a minimum alignment length that is longer than the primer sequence to assign it as a true amplicon.

Once we addressed these considerations, the mapping ratio of one of the NextSeq runs increased from 4% to 42%, demonstrating the power of the new algorithm. We provide its source code in GitHub[18] (see **2.5 Availability of Data and Materials**).

We traditionally designed the BART-Seq primers in a way that the resulting barcoded amplicons comply with 2×150 bp sequencing run with partially overlapping read pairs (**Figure 30A**), so that they could subsequently be merged. The primers targeted 80-250 bp loci, which, following addition of the linker and the barcode (10+11 bp) to both ends, would result in an amplicon range within 122-292 bp. Nonetheless, it is potentially possible to sequence an amplicon library in the range of 100-N bp with single end N nt sequencing. With the new demultiplexing pipeline, the size restriction becomes even less strict. For example, longer amplicons than the 122-292 bp range should not be a problem if identification of the exact amplicon sequence between the primers is not required, because merging is not necessary. Likewise, it should be possible to analyze the libraries with shorter reads, e.g. 2×75 bp paired-end, because they will still contain ~30 bp after the primers, that should be sufficient for assigning them to a target. Nevertheless, avoiding too long amplicons would be safer as they might reduce the total number of reads to be sequenced by taking up more space on the flow cell at optimal density, and result in lower base qualities as recently reviewed by Tan et al. (2019).

### 4.2.3 Using exogenous spike-ins for normalization and filtering

Various methods were introduced for normalizing the single-cell data using quantiles, total read counts, spike-ins, non-differentially expressed genes, pooled read counts, and so on. This is because, compared to bulk RNA-Seq, expression data from single cells is subject to higher variability due to biological (e.g. transcriptional

[18] https://github.com/theislab/bartseq-pipeline

bursting, cell size, cell type) or technical (e.g. sequencing depth, dropouts) factors, which holds true for the housekeeping genes as well (Luecken and Theis, 2019; Vallejos et al., 2017). The assumptions made for the global single-cell RNA-Seq data, such as constant amount of total mRNA per cell or majority of the genes not being differentially expressed (Lun et al., 2017) does not apply to BART-Seq due to the small number of genes analyzed and significant difference in their magnitudes, rendering the existing normalization methods impractical.

RNA spike-ins are commonly used in particular with the plate-based methods, to estimate the technical variations, even though they cannot normalize the differences in total mRNA content of cells. I, too, included RNA spike-ins in my experiments (**Figure 33**) by adding them at the reverse transcription phase, and including the primers targeting them in the multiplex primer sets. Because I could not include the whole set (e.g. 96) but only a few of the ERCC spike-ins, special attention was required in order not to skew the data while trying to normalize it.

A highly critical consideration was precise adjustment of the ratio of spike-ins to the cellular mRNAs. They may consume majority of the reads and preclude biological conclusions with very high ratios, whereas they may not suffice for normalization with lower ratios by intermingling with the noise (Lun et al., 2017; Vallejos et al., 2017). Hence, I carefully calculated per experiment the expected spike-in percentages based on the previous sequencing runs and adjusted the concentration of spike-ins accordingly.

Another challenge was the calculation of scaling factors for normalization. Arithmetic or geometric mean of non-transformed spike-in reads was not robust enough to outliers. Using median spike-in as the scaling factor was not an option either, because there were a maximum of four of them in my experiments, magnitudes of which could differ more than ten-fold despite fine-tuning the input amounts. Adjusting their overall magnitudes in the count data would bias the scaling factor towards the smaller spike-ins, whose variations became magnified upon re-sizing. Eventually, I came up with a formula to calculate the arithmetic mean of log2-transformed spike-ins, and back-transform this value to calculate size factors (**Formula (4)** or **(5)**, **Materials and Methods**, section **2.4.12.2**). When I applied it to normalize the count matrices, the variation of the spike-ins shrunk to a two-fold range, demonstrating the power of the approach. Even though I used three to four spike-ins in this project, including more would benefit the future experiments for the robustness of scaling factors. For bulk RNA analyses, more housekeeping genes can be included instead, as an alternative normalization approach.

Spike-ins were useful also to filter the wells containing insufficient number of reads due to failed sorting or lysis of the cells. I used the ratio of spike-ins to the total reads in the known empty wells (0 cell) as a filtering reference for failed samples. Conversely, some wells would receive cell doublets, which I filtered by estimating the two-cell expression levels by referring to the median read count of each gene across the matrix (**Figure 14**).

### 4.2.4 Barcode-primer combination effect

I discovered during data analyses that the amplification efficiency per gene had specific patterns depending on the barcode identities (barcode-primer combination effect), which were evident in both qPCR and sequencing data (**Figure 34**). Including spike-ins with constant amounts in the sequencing experiments facilitated this finding, because they made the patterns more obvious in comparison to the variably expressed genes. I initially sought to correct them, as I had already conducted a couple of large-scale experiments with substantial costs. Later, I also identified the underlying causes, which will considerably enhance the future projects by enabling *in silico* pre-selection of the optimal barcode and primer sets to minimize these effects.

Consistency of the patterns across the count matrices facilitated modeling and correcting them. Negative binomial generalized linear models could robustly estimate the read counts (**Figure 35**); yet, the use of the first component of each explanatory variable as intercept precluded calculation of their own coefficients. I circumvented this by building two models, one taking into account the interactions and one ignoring them, and dividing the predictions of the first model to that of the second to estimate the combination effects (**Formulas (2)** and **(3)**), and correct the data, as described in detail in **Materials and Methods**, section **2.4.12.1**.

Subsequently, I ran an experiment dedicated to calculating the combination effects for all the barcodes and primer sets we already had (**Figure 37**). The results helped me to empirically identify the globally inefficient barcodes and barcode-primer combinations. One use of this information was to flag the barcodes that are inefficient in a combination, for future reference. Another usage was to use them for further *in silico* analyses. Predicting the 2° structures of the whole amplicons containing average or inefficient combinations pointed to the barcode-linker-primer regions as the primary source of variations. Analysis of only the barcode-linker-primer sequences indicated that the stability of the 2° structures highly correlated with the inefficiency of the combination (**Figure 38**). I subsequently calculated the minimum free energies (MFE) of the known good/bad combinations, which revealed a clear pattern, where MFEs lower than -10 were all inefficient barcode-primer pairs. Parallel to that, some of the over-efficient combinations had higher-than-average MFEs.

These findings indicate that the barcode-primer combination effect is likely to be the result of stable hairpin formation of the primers, which reduces their availability in the reaction, thus the efficiency of PCR. PrimerSelect tool calculates the compatibility of multiplexed (nested) primers with each other, but not the combined sequences with the barcodes. To address this, there could be two potential improvements of the tool. The software can prompt the user to provide the list of barcodes and linkers that will be used in the experiment, so that the primers can be selected accordingly. Yet, this can restrict the primer selection considerably, which is already limited by the amplicon size range and the requirement to end with an Adenine base. A better option would be to design the primers as usual, simulate

their combinations with the whole barcode set, rank the MFEs of barcode-primer combinations, and eliminate barcodes that could potentially generate an inefficient combination. This can be either integrated into the PrimerSelect tool or provided as an accompanying script.

On the other hand, I discovered the roots of global barcode inefficiencies in an earlier step, the Klenow reaction. Presence of the GA dinucleotide as the last two bases of barcodes seems to increase the complementarity of barcode-linker dimers, eventually reducing the quantity of synthesized primers (**Figure 39**). This can simply be surpassed by excluding those barcodes from the panel and taking this knowledge into account when designing new barcode and linker sets.

Even though the barcode-primer combination effects also influenced the gene reads, I did not attempt to correct them in this study due to the complexity caused by their variable expression patterns. Yet, I excluded from the analysis the whole columns or rows with known inefficient combinations. Nevertheless, the MFE method should be helpful when designing new experiments to filter out the barcodes that might potentially have a bad combination with one of the primers, as stated above.

## 4.3 Applications of BART-Seq

### 4.3.1 RNA quantification

BART-Seq is a powerful method for quantifying template RNAs. To verify that, I initially measured a dilution series of isolated bulk RNA from hESCs, which showed remarkably high correlations between read counts and template concentrations (average $R^2$ of 0.96) (**Figure 42**). Importantly, the linear range spanned single cell levels (i.e. 4-16 pg). Direct analysis of cells (1-32 cells per well), too, revealed a notable correlation with the number of sorted cells and read counts ($R > 0.95$ for most genes) (**Figure 44**). When compared to the contemporary global scRNA-Seq techniques, the sensitivity of BART-Seq (median $R^2$ of 0.98) exceeded all the methods reviewed by Ziegenhain et al. (2017) (**Figure 43**).

BART-Seq has advantages over other targeted techniques as well. While the hybridization-based targeted methods require high amount of starting material that is not compatible with gene expression analysis in single cells, the amplification-based ones are not suited for quantitative precision (Hodges et al., 2007; Mercer et al., 2014; Ozsolak and Milos, 2011), both of which could be addressed with BART-Seq.

### 4.3.2 Influence of maintenance media on the pluripotency state of hESCs

With the purpose of finding out whether maintenance media might influence the expression of the core pluripotency network, I analyzed single hESCs cultured on

mTeSR™1, KSR-bFGF, or E8 using BART-Seq. Two main clusters emerged in the UMAP distribution, exhibiting naïve-like (*NANOG*$^{HIGH}$ *ZFP42*/*REX1*$^{HIGH}$) and primed-like (*LIN28A*$^{HIGH}$ *DNMT3B*$^{HIGH}$) pluripotency profiles (**Figure 45**) (Pastor et al., 2016; Theunissen et al., 2014; Warrier et al., 2017). These clusters were predominated by E8- and mTeSR™1-cultured cells, respectively, implying that they were transcriptionally conditioned by these media.

mTeSR™1 medium contains several components including cholesterol and lipids, that are not present in the E8 (**Table 2**). A previous study reported a switch in the lipid metabolism between the naïve and primed states of hPSCs (Sperber et al., 2015). A later study shows that, in contrast to the primed cells, E8 medium induces a formative pluripotent state that is in between the naïve and primed states (Cornacchia et al., 2019). These cells exhibit decreased *DNMT3B* and increased *NANOG* expression, which parallel my findings, and downregulate TGF-ß-related genes. It is suggested that lipid availability has a role in naïve-to-primed transition, and lipid deprivation might induce naïve-like features via suppression of endogenous ERK. The same study reported increased propensity of E8-cultured hPSCs for neuroectodermal differentiation over mesoderm or endoderm. This supports the reported cases in my lab having difficulties in differentiating the E8-cultured hPSCs towards mesoderm (e.g. cardiac) in comparison to mTeSR™1 (unpublished), and highlights the importance of matching the growth media to different differentiation protocols (Lee et al., 2017).

Another interesting observation was the strong correlation of *CCND1* expression with *NANOG* (and *ZFP42*), although it is not a known marker for the ground-state pluripotency. Human naïve hPSCs have a shorter G1 phase in comparison to the primed cells (Coronado et al., 2013), and Cyclin D1 is known to function in G1/S transition. Therefore, it is likely that increased *CCND1* expression in the naïve-like subpopulation relates to the higher proliferation rate of these cells.

### 4.3.3 Stimulation of the Wnt pathway with different inducers

In order to test whether different inducers that are used to activate the Wnt pathway yield the same transcriptional outcomes (Loh et al., 2016; Mendjan et al., 2014), I analyzed cell populations that emerge upon treating the hESCs with recombinant Wnt3a (rWnt3a) protein, a small molecule inhibitor of GSK3 (CHIR99021), or ectopically expressing ß-catenin for 0, 24 and 72 hours. Pairwise gene correlations at 24 hours resulted in two main groups, regardless of the inducer. The distribution of these transcripts in the developing mouse embryo (E6.5-E7.5) resembled the distal and proximal regions (**Figure 48**, **APPENDIX C**), while the pan-primitive streak markers *GSC*, *EVX1*, and *MIXL1* correlated with both groups, and *MIXL1* was higher in the distal-like group. One of the two cell clusters at 24 h in the tSNE plot expressed the genes related to both distal and proximal regions and was enriched for CHIR-induced cells, and the second expressed distal-like genes only and was enriched for ß-catenin-induced cells. It can be speculated that this is because ß-catenin is downstream to the GSK3 in the cascade, therefore may not be

able to activate all the Wnt-responsive genes. The cells isolated at 72 hours, especially rWnt3a-induced ones, formed distinguishable clusters, too. A cell subpopulation co-expressing endoderm markers *SOX17* and *FOXA2* emerged only with rWnt3a-induction, possibly because Wnt3a, located upstream of the other two inducers, could give rise to a more diverse set of progeny. This subpopulation was absent in a replicate experiment though (**APPENDIX D**). One interesting observation after 72 hours was that *FOXA2* and *MSX1* were mutually exclusive, which localize to the floor plate (Lek et al., 2010) and the roof plate (Eggenschwiler and Anderson, 2000) of the developing neural tube respectively, hence they are possibly inversely regulated.

As an orthogonal validation, I checked the same set of genes at 72 hours in a bulk RNA-Seq analysis of a replicate experiment (**Figure 47**). Notwithstanding a significant degree of cellular heterogeneity in the single-cell data, majority of the genes had very similar patterns with the bulk data. Importantly, ß-catenin and CHIR-inductions resembled each other but diverged from the rWnt3a, as mentioned above.

In conclusion, activating the Wnt pathway using different inducers does not seem to yield the same cell subpopulations. Supposedly, it should not be possible to adjust the intracellular effective concentration of each inducer to the same levels, which might partially account for these differences. Besides, as they are located at different levels of the pathway, they should be involved in diverse feedback or feedforward loops, which cannot generate the exact same outcomes when ectopically activated. These inferences are based on only this experiment, therefore further research would be required to reach more decisive conclusions. Nonetheless, they confirm that BART-Seq is useful for gaining novel insights into gene expression patterns or verifying the existing knowledge.

### 4.3.4 Bulk analyses

Besides single-cell transcriptomics, BART-Seq can be used for analyzing transcriptomic or genomic targets in bulk samples as well. Here, I presented two proof of concept experiments. One of the applications was screening the gDNA samples from 96 breast cancer patients for the 10 founder mutations on the *BRCA1* and *BRCA2* genes, which resulted in 99% genotyping accuracy based on mutations known from Sanger sequencing (**Figure 49**). Another application was screening the transcriptional response of hepatocytes to a compound library as part of a collaboration project with the pharmaceutical company AstraZeneca. We analyzed 17 target genes, some of which correlated well with the qPCR analysis of the same samples (**Figure 50**). Nonetheless, it was not possible to derive nice dosage response curves for most of the genes. We hypothesized that this was either a result of incomplete lysis of the sampled cells, or inhibition of the PCR by the components in the lysis buffer. Efficient amplification of spike-ins unlike the gene targets supports the prior hypothesis. Due to the time restriction we could not pursue this project further, yet it exemplified a potential application. Taken together, these

experiments demonstrated that BART-Seq can substitute the conventional methods such as Sanger sequencing or qPCR for high-throughput genomic or transcriptomic screening, by reducing the costs and labor.

## 4.4 Advantages of BART-Seq

### 4.4.1 A targeted approach for quantitative -omics

The unbiased (global) scRNA-Seq techniques distribute the sequencing reads to thousands of genes in thousands of cells, often leading to zero-inflated count matrices, where the distinction between the unexpression and dropouts becomes profoundly blurred. In that case, mathematical approaches are often required to "estimate" the expression of the genes using the pooled read counts (Luecken and Theis, 2019). While this level of information can accurately classify cell types, which is adequate for many research questions, it may not suffice to explore the changes in the expression levels of genes, when the gene of interest cannot be even detected at times. Therefore, focusing the sequencing capacity to a small number of targets can provide deeper and quantitative information. Here, I show that BART-Seq is highly accurate in linear quantification of the template concentrations, down to single cell levels (**Figure 42**), which outperforms the eminent technologies (CEL-Seq2/C1, Drop-seq, MARS-seq, SCRB-seq, Smart-seq/C1, Smart-seq2) recently reviewed by Ziegenhain et al. (2017) (**Figure 43**). There are also notable imaging based-techniques such as MERFISH (Chen et al., 2015) or seqFISH (Shah et al., 2017), which provides higher resolution information including localization of the transcripts; yet, they are often laborious and expensive. Besides, they infer the transcripts from hybridization of probes, like the qPCR-based Fluidigm Biomark, whereas BART-Seq provides direct sequence information.

Methods exist for target enrichment by hybridization such as NimbleGen arrays (Hodges et al., 2007) or RNA CaptureSeq (Mercer et al., 2014), which often require high amounts of starting material. There are also methods based on multiplexed PCR amplification (Blomquist et al., 2013; Herbold et al., 2015; Ståhlberg et al., 2016; Ion AmpliSeq from Life Technologies), MIP-based methods using circularizable probes (Tao et al., 2018), or parallel amplification of multiple targets like the RainStorm platform (Tewhey et al., 2009). Nonetheless, these targeted approaches are neither suited for amplifying small amounts of templates, such as from single cells, nor for high degree of sample parallelization due to high costs or technical limitations. One of the few methods for targeted scRNA-Seq is CytoSeq, which combines oligo(dT) capturing with gene-specific primers to analyze 100+ genes (Fan et al., 2015). RAGE-Seq combines targeted nanopore sequencing of full-length transcripts with the short-read sequencing (Singh et al., 2019), yet, accessibility of third-generation sequencers might preclude its widespread usage for the time being. Also, the quantitative range of these methods remains to be determined. In conclusion, BART-Seq can offer targeted validation for specific

research questions that require higher resolution information with quantitative sensitivity.

### 4.4.2 Sequence coverage

Majority of the scRNA-Seq methods are based on tag counting (i.e. only 3' or 5' end of transcripts) because they add the barcodes at the reverse transcription stage, often to the poly(A) tails, which are used also to capture the transcripts (**Table 1**). Yet, this does not allow analyzing the regions that are far from the transcript ends, for example detecting different isoforms. There are only a few methods, such as Smart-seq or RamDA-seq that can sequence whole transcripts (Hayashi et al., 2018; Picelli et al., 2013). Even though BART-Seq does not offer full-length coverage, it does not depend on the poly(A) tail because the barcodes are attached not to RT but to PCR primers, which enables amplifying any part of the transcripts once they are reverse transcribed simply using oligo(dT) primers, random hexamers, or even barcoded primers directly. Thereby, it allows analyzing a wider range of transcripts unlike the majority of the existing technologies, including lncRNAs that play important roles in development and disease (Esteller, 2011).

### 4.4.3 An economical method

The existing scRNA-Seq technologies often require expensive instrumentation and consumables (Macosko et al., 2015; Picelli et al., 2013; Zheng et al., 2017). Whereas the droplet-based, microfluidic, or nanowell systems offer massive reduction of reagents consumed per cell (Ziegenhain et al., 2017), the overall expenses turn out to be still high due to non-customizable size and number of samples that can be analyzed in parallel, each of which would require the use of different preparation kits for every iteration of time point, treatment, and biological replicate (Macosko et al., 2015; Zheng et al., 2017). In comparison, BART-Seq libraries are highly versatile; they can consist of tens of samples or thousands of samples, only a few or hundreds of different conditions can be assayed in parallel, which can be easily adjusted by just up/downscaling the size of the barcode matrix based on the experimental needs. This allows adjusting the sample size according to the available budget. The full cost of analysis per sample (i.e. a single-cell or bulk gDNA/cDNA in one well of a 384-well plate) is approximately 1 EUR including sequencing, which is more affordable than many of the plate- or microfluidics-based approaches (Ziegenhain et al., 2017).

### 4.4.4 Versatility and accessibility

BART-Seq has a simple workflow that can be performed with common laboratory equipment and basic reagents, making it available to any research group with access to a next-generation sequencing instrument. The barcode assembly and target enrichment are straightforward, without any intermittent purification steps, which is a plus for the robustness of quantification. The whole workflow takes

approximately one week to complete, starting from barcode assembly till the generation of count matrices.

As a plate-based technique, BART-Seq is not restricted with the cell sizes in a certain range, unlike the droplet- or microfluidic-based techniques (Loh et al., 2016; Macosko et al., 2015; Sanchez-Freire et al., 2012). It is possible to analyze single cells isolated using different methods (e.g. LCM, FACS, single-cell printer), and bulk samples or sensitive cells can simply be pipetted into the wells. Sample properties can be indexed (e.g. fluorescent intensity, phenotype, patient ID) and linked to the expression profiles or genotypes, an information that is often lost with the techniques that utilize pools of cells and barcoded droplets. Another advantage of being a plate-based method would be the presence of excess reagents per sample, which can be speculated to be useful for capturing as many molecules as possible.

## 4.5 Limitations of BART-Seq

Just like any other technique, BART-Seq comes with certain limitations as well. As I showed here in the form of barcode-primer combination effects, the efficiency of primers is likely to have some degree of heterogeneity, even though they can be kept within a negligible range by *in silico* pre-selection. Besides, the method is not fully immune to the complications related to multiplex PCR, such as uneven efficiency of amplification, non-specific products, or cross-hybridization of the pooled primers, though the PrimerSelect tool minimizes them. The amplification-based nature of BART-Seq may bring about questions regarding PCR artifacts, which is addressed by other technologies using UMIs to tag and count the original mRNA molecules (**Table 1**). Unless a two-step PCR is not implemented into BART-Seq, it is not possible to include UMIs in the workflow, because the same primers that attach the UMIs to the amplicons cannot be used for further amplification; as the UMI part would not match in the next rounds of PCR. Considering the simplicity as the basic merit of BART-Seq, attempting to synthesize such complex primers would abolish this advantage. Given the close-to-perfect linear RNA quantification of the targets down to picogram levels (**Figure 42**), absence of UMIs may be neglected.

Be it due to the thermodynamic restrictions in pooling numerous primers, or the availability of PCR resources, there is a hypothetical limit to the number of primers that can be multiplexed. More importantly, the more targets analyzed in parallel are, the lower the average number of reads they would receive; besides the possibility of a few over-efficient primers or highly abundant genes consuming the majority of the reads, just like the unbiased methods. Even though it is hard to specify an upper limit to multiplexing (e.g. 50-100), the expected yield of the sequencing run and the desired average sequencing depth per target can be useful references for making a decision.

Unlike the global techniques that can lead to novel discoveries on gene regulatory networks, BART-Seq is a hypothesis-driven approach, in that, the list of targets to

be analyzed must be known in advance. In essence, it can be considered a sequencing-based version of a highly parallelized qPCR. This is not a limitation though for the experiments that aim to assay a known set of loci, such as screening for certain mutations, or CRISPR targets. Finally, BART-Seq is not yet an automated method; hence requires a lot of hands on time.

Lastly, a minor concern regarding scRNA-Seq experiments would be the possibility of amplification of the targets also from genomic DNA. I observed this with *NANOG* primers, which generated amplicons in negligible amounts from a pseudogene (*NANOGP1*) as well, i.e. 1-2% of the real amplicons. The pluripotency transcripts assigned to fibroblasts might be partially caused by the same reason (**Figure 44C**). In cases where accuracy is crucial, designing primers that span exons would be a solution.

## 4.6 Further Applications

High-throughput sequencing of defined sets of transcripts could be very useful for numerous studies that involve parallel analysis of massive arrays of samples. The application areas include probing of mechanisms; single-cell analysis; validating and complementing results obtained by genome-wide approaches, such as the Human Cell Atlas Project (Regev et al., 2017); and screening in genome engineering, drug development, and toxicology assays. I have already presented here a range of possible BART-Seq applications. For example, it suits high-throughput targeted genomics, as demonstrated by screening breast cancer patients. It can be useful for the pharmaceutical industry for screening transcriptional response of certain genes to a compound library cost-effectively, as we explored with the AstraZeneca project. Also, it is convenient to assay in parallel tens or hundreds of different time points or conditions, as exemplified on a smaller scale by the Wnt pathway stimulation experiment.

Furthermore, BART-Seq can aid precision medicine by screening a large population of individuals for the mutations that are known to influence the action of a certain drug and personalize the treatments accordingly. We have not tested the method for bacteria or viruses yet, but it might be useful for testing patient samples (e.g. blood, mucous, urine), or other liquids for the presence of certain pathogens. Moreover, BART-Seq can be combined with the CRISPR/Cas9 technology for probing the introduced modifications in the targeted genomic loci and correlating them with the transcriptional response of selected sets of genes, either at the clonal or single-cell level. Besides, multiple primer pairs can be included in the same primer pool for different transcript isoforms to study their regulation in changing physiological conditions, for example during stem cell differentiation. Primer sets can include non-poly(A) targets as well, such as lncRNAs. I can foresee that the possibilities to use BART-Seq for highly parallelized targeted applications is much more than the examples listed here.

## 4.7 Conclusions

I focused this project into the development of the novel method BART-Seq, and its applications. I optimized individual steps of the workflow to increase the efficiency and reduce the costs, and implemented it for targeted transcriptomics, including that of single cells. In parallel, I established, and also assisted the development of bioinformatics tools for designing the experiments and analyzing the results.

BART-Seq overcomes many limitations of other targeted or single-cell sequencing approaches, such as shallow coverage, lack of quantitativeness, or high costs. It is a highly precise, inexpensive, simple, and scalable method that can be readily used by any research group with basic laboratory equipment and reagents, and access to a next-generation sequencing device. The accompanying bioinformatics tools are available with open access to the scientific community.

To my knowledge, BART-Seq is the first targeted sequencing technology that is applicable to both transcriptomics of single cells and genomics/transcriptomics of bulk samples. Thereby, it will serve a practical alternative or a companion to the existing technologies in a wide spectrum of research fields.

# REFERENCES

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651–1656.

Aho, A.V., and Corasick, M.J. (1975). Efficient string matching: an aid to bibliographic search. Commun. ACM *18*, 333–340.

Alwine, J.C., Kemp, D.J., and Stark, G.R. (1977). Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. U. S. A. *74*, 5350–5354.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature *226*, 1209–1211.

Barth, A.I., Pollack, A.L., Altschuler, Y., Mostov, K.E., and Nelson, W.J. (1997). NH2-terminal deletion of beta-catenin results in stable colocalization of mutant beta-catenin with adenomatous polyposis coli protein and altered MDCK cell adhesion. J. Cell Biol. *136*, 693–706.

Ben-Haim, N., Lu, C., Guzman-Ayala, M., Pescatore, L., Mesnard, D., Bischofberger, M., Naef, F., Robertson, E.J., and Constam, D.B. (2006). The Nodal Precursor Acting via Activin Receptors Induces Mesoderm by Maintaining a Source of Its Convertases and BMP4. Dev. Cell *11*, 313–323.

Blauwkamp, T.A., Nigam, S., Ardehali, R., Weissman, I.L., and Nusse, R. (2012). Endogenous Wnt signalling in human embryonic stem cells generates an equilibrium of distinct lineage-specified progenitors. Nat. Commun. *3*, 1070.

Blomquist, T.M., Crawford, E.L., Lovett, J.L., Yeo, J., Stanoszek, L.M., Levin, A., Li, J., Lu, M., Shi, L., Muldrew, K., et al. (2013). Targeted RNA-Sequencing with Competitive Multiplex-PCR Amplicon Libraries. PLOS ONE *8*, e79120.

Cao, E., Chen, Y., Cui, Z., and Foster, P.R. (2003). Effect of freezing and thawing rates on denaturation of proteins in aqueous solutions. Biotechnol. Bioeng. *82*, 684–690.

Cao, J., Packer, J.S., Ramani, V., Cusanovich, D.A., Huynh, C., Daza, R., Qiu, X., Lee, C., Furlan, S.N., Steemers, F.J., et al. (2017). Comprehensive single cell transcriptional profiling of a multicellular organism by combinatorial indexing. BioRxiv 104844.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature *566*, 496–502.

Chandler, D.P., Wagnon, C.A., and Bolton, H. (1998). Reverse Transcriptase (RT) Inhibition of PCR at Low Concentrations of Template and Its Implications for Quantitative RT-PCR. Appl. Environ. Microbiol. *64*, 669–677.

Chen, G., Gulbranson, D.R., Hou, Z., Bolin, J.M., Ruotti, V., Probasco, M.D., Smuga-Otto, K., Howden, S.E., Diol, N.R., Propson, N.E., et al. (2011). Chemically defined conditions for human iPS cell derivation and culture. Nat. Methods *8*, 424–429.

Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. Science *348*, aaa6090.

Chumakov, K.M. (1994). Reverse transcriptase can inhibit PCR and stimulate primer-dimer formation. Genome Res. *4*, 62–64.

Clark, J.M. (1988). Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. Nucleic Acids Res. *16*, 9677–9686.

Cornacchia, D., Zhang, C., Zimmer, B., Chung, S.Y., Fan, Y., Soliman, M.A., Tchieu, J., Chambers, S.M., Shah, H., Paull, D., et al. (2019). Lipid Deprivation Induces a Stable, Naive-to-Primed Intermediate State of Pluripotency in Human PSCs. Cell Stem Cell *25*, 120-136.e10.

Coronado, D., Godet, M., Bourillot, P.-Y., Tapponnier, Y., Bernat, A., Petit, M., Afanassieff, M., Markossian, S., Malashicheva, A., Iacone, R., et al. (2013). A short G1 phase is an intrinsic determinant of naïve embryonic stem cell pluripotency. Stem Cell Res. *10*, 118–131.

Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., Stephan, D.A., et al. (2008). Identification of genetic variants using bar-coded multiplexed sequencing. Nat. Methods *5*, 887–893.

Dakhore, S., Nayer, B., and Hasegawa, K. (2018). Human Pluripotent Stem Cell Culture: Current Status, Challenges, and Advancement. Stem Cells Int. *2018*.

Deamer, D., Akeson, M., and Branton, D. (2016). Three decades of nanopore sequencing. Nat. Biotechnol. *34*, 518–524.

Doble, B.W., and Woodgett, J.R. (2003). GSK-3: tricks of the trade for a multi-tasking kinase. J. Cell Sci. *116*, 1175–1186.

Drukker, M., Katz, G., Urbach, A., Schuldiner, M., Markel, G., Itskovitz-Eldor, J., Reubinoff, B., Mandelboim, O., and Benvenisty, N. (2002). Characterization of the expression of MHC proteins in human embryonic stem cells. Proc. Natl. Acad. Sci. U. S. A. *99*, 9864–9869.

Eberwine, J., Sul, J.-Y., Bartfai, T., and Kim, J. (2014). The promise of single-cell sequencing. Nat. Methods *11*, 25–27.

Eckert, K.A., and Kunkel, T.A. (1990). High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. Nucleic Acids Res. *18*, 3739–3744.

Eggenschwiler, J.T., and Anderson, K.V. (2000). Dorsal and Lateral Fates in the Mouse Neural Tube Require the Cell-Autonomous Activity of the open brain Gene. Dev. Biol. *227*, 648–660.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., et al. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. Science *323*, 133–138.

Esteller, M. (2011). Non-coding RNAs in human disease. Nat. Rev. Genet. *12*, 861–874.

Faial, T., Bernardo, A.S., Mendjan, S., Diamanti, E., Ortmann, D., Gentsch, G.E., Mascetti, V.L., Trotter, M.W.B., Smith, J.C., and Pedersen, R.A. (2015). Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. Development *142*, 2121–2135.

Fan, H.C., Fu, G.K., and Fodor, S.P.A. (2015). Combinatorial labeling of single cells for gene expression cytometry. Science *347*, 1258367.

Foulk, M.S., Urban, J.M., Casella, C., and Gerbi, S.A. (2015). Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome Res. *25*, 725–735.

Garalde, D.R., Snell, E.A., Jachimowicz, D., Heron, A.J., Bruce, M., Lloyd, J., Warland, A., Pantic, N., Admassu, T., Ciccone, J., et al. (2016). Highly parallel direct RNA sequencing on an array of nanopores. BioRxiv 068809.

Gierliński, M., Cole, C., Schofield, P., Schurch, N.J., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G., Owen-Hughes, T., et al. (2015). Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics *31*, 3625–3630.

Gyi, J.I., Lane, A.N., Conn, G.L., and Brown, T. (1998). The orientation and dynamics of the C2'-OH and hydration of RNA and DNA.RNA hybrids. Nucleic Acids Res. *26*, 3104–3110.

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

Hangauer, M.J., Vaughn, I.W., and McManus, M.T. (2013). Pervasive Transcription of the Human Genome Produces Thousands of Previously Unidentified Long Intergenic Noncoding RNAs. PLOS Genet. *9*, e1003569.

Hayashi, T., Ozaki, H., Sasagawa, Y., Umeda, M., Danno, H., and Nikaido, I. (2018). Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat. Commun. *9*, 619.

Henegariu, O., Heerema, N. a., Dlouhy, S. r., Vance, G. h., and Vogt, P. h. (1997). Multiplex PCR: Critical Parameters and Step-by-Step Protocol. BioTechniques *23*, 504–511.

Herbold, C.W., Pelikan, C., Kuzyk, O., Hausmann, B., Angel, R., Berry, D., and Loy, A. (2015). A flexible and economical barcoding approach for highly multiplexed amplicon sequencing of diverse target genes. Front. Microbiol. *6*.

Higuchi, R., Fockler, C., Dollinger, G., and Watson, R. (1993). Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. Bio/Technology *11*, 1026.

Hodges, E., Xuan, Z., Balija, V., Kramer, M., Molla, M.N., Smith, S.W., Middle, C.M., Rodesch, M.J., Albert, T.J., Hannon, G.J., et al. (2007). Genome-wide *in situ* exon capture for selective resequencing. Nat. Genet. *39*, 1522–1527.

Hwang, B., Lee, J.H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp. Mol. Med. *50*, 96.

Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., and Oliver, B. (2011). Synthetic spike-in standards for RNA-seq experiments. Genome Res. *21*, 1543–1551.

Joshi, N.A., and Fass, J.N. (2011). Sickle: a sliding-window, adaptive, quality-based trimming tool for FastQ files. (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle.

Kainz, P. (2000). The PCR plateau phase – towards an understanding of its limitations. Biochim. Biophys. Acta BBA - Gene Struct. Expr. *1494*, 23–27.

Kalisky, T., Oriel, S., Bar-Lev, T.H., Ben-Haim, N., Trink, A., Wineberg, Y., Kanter, I., Gilad, S., and Pyne, S. (2018). A brief review of single-cell transcriptomic technologies. Brief. Funct. Genomics *17*, 64–76.

Kapley, A., Lampel, K., and Purohit, H.J. (2000). Thermocycling steps and optimization of multiplex PCR. Biotechnol. Lett. *22*, 1913–1918.

Karczewski, K.J., and Snyder, M.P. (2018). Integrative omics for health and disease. Nat. Rev. Genet. *19*, 299–310.

Kaufman, B., Laitman, Y., Carvalho, M.A., Edelman, L., Menachem, T.D., Zidan, J., Monteiro, A.N., and Friedman, E. (2006). The P1812A and P25T BRCA1 and the 5164del4 BRCA2 Mutations: Occurrence in High-Risk Non-Ashkenazi Jews. Genet. Test. *10*, 200–207.

Kent, W.J. (2002). BLAT—The BLAST-Like Alignment Tool. Genome Res. *12*, 656–664.

Kerr, C., and Sadowski, P.D. (1972). Gene 6 Exonuclease of Bacteriophage T7 I. PURIFICATION AND PROPERTIES OF THE ENZYME. J. Biol. Chem. *247*, 305–310.

Kim, D., Langmead, B., and Salzberg, S.L. (2015). HISAT: a fast spliced aligner with low memory requirements. Nat. Methods *12*, 357–360.

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., and Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. Nat. Methods *9*, 72–74.

Klenow, H., and Henningsen, I. (1970). Selective Elimination of the Exonuclease Activity of the Deoxyribonucleic Acid Polymerase from Escherichia coli B by Limited Proteolysis*. Proc. Natl. Acad. Sci. U. S. A. *65*, 168–175.

Köster, J., and Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. Bioinformatics *28*, 2520–2522.

Krendl, C., Shaposhnikov, D., Rishko, V., Ori, C., Ziegenhain, C., Sass, S., Simon, L., Müller, N.S., Straub, T., Brooks, K.E., et al. (2017). GATA2/3-TFAP2A/C transcription factor network couples human pluripotent stem cell differentiation to trophectoderm with repression of pluripotency. Proc. Natl. Acad. Sci. *114*, E9579–E9588.

Kulski, J.K. (2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. Gener. Seq. - Adv. Appl. Chall.

Kunze, C., Börner, K., Kienle, E., Orschmann, T., Rusha, E., Schneider, M., Radivojkov - Blagojevic, M., Drukker, M., Desbordes, S., Grimm, D., et al. (2018). Synthetic AAV/CRISPR vectors for blocking HIV-1 expression in persistently infected astrocytes. Glia *66*, 413–427.

L. Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. *17*, 75.

Laitman, Y., Simeonov, M., Herskovitz, L., Kushnir, A., Shimon-Paluch, S., Kaufman, B., Zidan, J., and Friedman, E. (2012). Recurrent germline mutations in BRCA1 and BRCA2 genes in high risk families in Israel. Breast Cancer Res. Treat. *133*, 1153–1157.

Lee, T.I., and Young, R.A. (2013). Transcriptional Regulation and its Misregulation in Disease. Cell *152*, 1237–1251.

Lee, J.H., Daugharthy, E.R., Scheiman, J., Kalhor, R., Ferrante, T.C., Terry, R., Turczyk, B.M., Yang, J.L., Lee, H.S., Aach, J., et al. (2015). Fluorescent *in situ* sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues. Nat. Protoc. *10*, 442–458.

Lee, J.-H., Laronde, S., Collins, T.J., Shapovalova, Z., Tanasijevic, B., McNicol, J.D., Fiebig-Comyn, A., Benoit, Y.D., Lee, J.B., Mitchell, R.R., et al. (2017). Lineage-Specific Differentiation Is Influenced by State of Human Pluripotency. Cell Rep. *19*, 20–35.

Lek, M., Dias, J.M., Marklund, U., Uhde, C.W., Kurdija, S., Lei, Q., Sussel, L., Rubenstein, J.L., Matise, M.P., Arnold, H.-H., et al. (2010). A homeodomain feedback circuit underlies step-function interpretation of a Shh morphogen gradient during ventral neural patterning. Development *137*, 4051–4060.

Lerer, I., Wang, T., Peretz, T., Sagi, M., Kaduri, L., Orr-Urtreger, A., Stadler, J., Gutman, H., and Abeliovich, D. (1998). The 8765delAG mutation in BRCA2 is common among Jews of Yemenite extraction. Am. J. Hum. Genet. *63*, 272–274.

Li, V.S.W., Ng, S.S., Boersema, P.J., Low, T.Y., Karthaus, W.R., Gerlach, J.P., Mohammed, S., Heck, A.J.R., Maurice, M.M., Mahmoudi, T., et al. (2012). Wnt Signaling through Inhibition of ß-Catenin Degradation in an Intact Axin1 Complex. Cell *149*, 1245–1256.

Lindsley, R.C., Gill, J.G., Kyba, M., Murphy, T.L., and Murphy, K.M. (2006). Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. Development *133*, 3787–3796.

Little, J.W. (1967). An exonuclease induced by bacteriophage lambda. II. Nature of the enzymatic reaction. J. Biol. Chem. *242*, 679–686.

Livesey, F.J. (2003). Strategies for microarray analysis of limiting amounts of RNA. Brief. Funct. Genomic. Proteomic. *2*, 31–36.

Loh, K.M., Chen, A., Koh, P.W., Deng, T.Z., Sinha, R., Tsai, J.M., Barkal, A.A., Shen, K.Y., Jain, R., Morganti, R.M., et al. (2016). Mapping the Pairwise Choices Leading from Pluripotency to Human Bone, Heart, and Other Mesoderm Cell Types. Cell *166*, 451–467.

Lowe, R., Shirley, N., Bleackley, M., Dolan, S., and Shafee, T. (2017). Transcriptomics technologies. PLoS Comput. Biol. *13*.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Mol. Syst. Biol. *15*, e8746.

Lun, A.T.L., Calero-Nieto, F.J., Haim-Vilmovsky, L., Göttgens, B., and Marioni, J.C. (2017). Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. Genome Res. *27*, 1795–1806.

Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. J. Mach. Learn. Res. *9*, 2579–2605.

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

Mamanova, L., Coffey, A.J., Scott, C.E., Kozarewa, I., Turner, E.H., Kumar, A., Howard, E., Shendure, J., and Turner, D.J. (2010a). Target-enrichment strategies for next-generation sequencing. Nat. Methods *7*, 111–118.

Mamanova, L., Andrews, R.M., James, K.D., Sheridan, E.M., Ellis, P.D., Langford, C.F., Ost, T.W.B., Collins, J.E., and Turner, D.J. (2010b). FRT-seq: amplification-free, strand-specific transcriptome sequencing. Nat. Methods *7*, 130–132.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.-J., Chen, Z., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature *437*, 376.

Marín de Evsikova, C., Raplee, I.D., Lockhart, J., Jaimes, G., and Evsikov, A.V. (2019). The Transcriptomic Toolbox: Resources for Interpreting Large Gene Expression Data within a Precision Medicine Context for Metabolic Disease Atherosclerosis. J. Pers. Med. *9*, 21.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. ArXiv180203426 Cs Stat.

Mendjan, S., Mascetti, V.L., Ortmann, D., Ortiz, M., Karjosukarso, D.W., Ng, Y., Moreau, T., and Pedersen, R.A. (2014). NANOG and CDX2 Pattern Distinct Subtypes of Human Mesoderm during Exit from Pluripotency. Cell Stem Cell *15*, 310–325.

Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat. Protoc. *9*, 989–1009.

Metzker, M.L. (2010). Sequencing technologies — the next generation. Nat. Rev. Genet. *11*, 31–46.

Nichols, J., and Smith, A. (2009). Naive and Primed Pluripotent States. Cell Stem Cell *4*, 487–492.

Nikiforov, T.T., Rendle, R.B., Kotewicz, M.L., and Rogers, Y.H. (1994). The use of phosphorothioate primers and exonuclease hydrolysis for the preparation of single-stranded PCR products and their detection by solid-phase hybridization. Genome Res. *3*, 285–291.

Ozsolak, F., and Milos, P.M. (2011). RNA sequencing: advances, challenges and opportunities. Nat. Rev. Genet. *12*, 87–98.

Pan, X., Durrett, R.E., Zhu, H., Tanaka, Y., Li, Y., Zi, X., Marjani, S.L., Euskirchen, G., Ma, C., LaMotte, R.H., et al. (2013). Two methods for full-length RNA sequencing for low quantities of cells and single cells. Proc. Natl. Acad. Sci. *110*, 594–599.

Pastor, W.A., Chen, D., Liu, W., Kim, R., Sahakyan, A., Lukianchikov, A., Plath, K., Jacobsen, S.E., and Clark, A.T. (2016). Naïve human pluripotent cells feature a methylation landscape devoid of blastocyst or germline memory. Cell Stem Cell *18*, 323–329.

Perea-Gomez, A., Vella, F.D.J., Shawlot, W., Oulad-Abdelghani, M., Chazaud, C., Meno, C., Pfister, V., Chen, L., Robertson, E., Hamada, H., et al. (2002). Nodal Antagonists in the Anterior Visceral Endoderm Prevent the Formation of Multiple Primitive Streaks. Dev. Cell *3*, 745–756.

Picelli, S. (2017). Single-cell RNA-sequencing: The future of genome biology is now. RNA Biol. *14*, 637–650.

Picelli, S., Björklund, Å.K., Faridani, O.R., Sagasser, S., Winberg, G., and Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat. Methods *10*, 1096–1098.

Piétu, G., Mariage-Samson, R., Fayein, N.-A., Matingou, C., Eveno, E., Houlgatte, R., Decraene, C., Vandenbrouck, Y., Tahi, F., Devignes, M.-D., et al. (1999). The Genexpress IMAGE Knowledge Base of the Human Brain Transcriptome: A Prototype Integrated Resource for Functional and Computational Genomics. Genome Res. *9*, 195–209.

Porreca, G.J., Zhang, K., Li, J.B., Xie, B., Austin, D., Vassallo, S.L., LeProust, E.M., Peck, B.J., Emig, C.J., Dahl, F., et al. (2007). Multiplex amplification of large sets of human exons. Nat. Methods *4*, 931–936.

Raj, A., Peskin, C.S., Tranchina, D., Vargas, D.Y., and Tyagi, S. (2006). Stochastic mRNA Synthesis in Mammalian Cells. PLOS Biol. *4*, e309.

Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al. (2017). The Human Cell Atlas. ELife *6*.

Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., et al. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science *360*, 176–182.

Rothberg, J.M., Hinz, W., Rearick, T.M., Schultz, J., Mileski, W., Davey, M., Leamon, J.H., Johnson, K., Milgrew, M.J., Edwards, M., et al. (2011). An integrated semiconductor device enabling non-optical genome sequencing. Nature *475*, 348–352.

Saliba, A.-E., Westermann, A.J., Gorski, S.A., and Vogel, J. (2014). Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. *42*, 8845–8860.

Salomonis, N., Schlieve, C.R., Pereira, L., Wahlquist, C., Colas, A., Zambon, A.C., Vranizan, K., Spindler, M.J., Pico, A.R., Cline, M.S., et al. (2010). Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. Proc. Natl. Acad. Sci. *107*, 10514–10519.

Sanchez-Freire, V., Ebert, A.D., Kalisky, T., Quake, S.R., and Wu, J.C. (2012). Microfluidic single-cell real-time PCR for comparative analysis of gene expression patterns. Nat. Protoc. *7*, 829–838.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J. Mol. Biol. *94*, 441–448.

Sasagawa, Y., Nikaido, I., Hayashi, T., Danno, H., Uno, K.D., Imai, T., and Ueda, H.R. (2013). Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. *14*, R31.

Satija, R., Farrell, J.A., Gennert, D., Schier, A.F., and Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nat. Biotechnol. *33*, 495–502.

Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. Science *270*, 467–470.

Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2017). seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus. Neuron *94*, 752-758.e1.

Shapiro, E., Biezuner, T., and Linnarsson, S. (2013). Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat. Rev. Genet. *14*, 618–630.

Sheng, K., Cao, W., Niu, Y., Deng, Q., and Zong, C. (2017). Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat. Methods *14*, 267–270.

Silva, J., and Smith, A. (2008). Capturing Pluripotency. Cell *132*, 532–536.

Singh, M., Al-Eryani, G., Carswell, S., Ferguson, J.M., Blackburn, J., Barton, K., Roden, D., Luciani, F., Phan, T.G., Junankar, S., et al. (2019). High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat. Commun. *10*, 3120.

Sinha, R., Stanley, G., Gulati, G.S., Ezran, C., Travaglini, K.J., Wei, E., Chan, C.K.F., Nabhan, A.N., Su, T., Morganti, R.M., et al. (2017). Index switching causes "spreading-of-signal" among multiplexed samples in Illumina HiSeq 4000 DNA sequencing. BioRxiv 125724.

Smith, Z.D., Sindhu, C., and Meissner, A. (2016). Molecular features of cellular reprogramming and development. Nat. Rev. Mol. Cell Biol. *17*, 139–154.

Sperber, H., Mathieu, J., Wang, Y., Ferreccio, A., Hesson, J., Xu, Z., Fischer, K.A., Devi, A., Detraux, D., Gu, H., et al. (2015). The metabolome regulates the epigenetic landscape during naive-to-primed human embryonic stem cell transition. Nat. Cell Biol. *17*, 1523–1535.

Ståhlberg, A., Krzyzanowski, P.M., Jackson, J.B., Egyud, M., Stein, L., and Godfrey, T.E. (2016). Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing. Nucleic Acids Res. *44*, e105–e105.

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

Svensson, V., Vento-Tormo, R., and Teichmann, S.A. (2017). Exponential scaling of single-cell RNA-seq in the last decade. ArXiv170401379 Q-Bio.

Takahashi, K., and Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. Cell *126*, 663–676.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., and Yamanaka, S. (2007). Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors. Cell *131*, 861–872.

Takashima, Y., Guo, G., Loos, R., Nichols, J., Ficz, G., Krueger, F., Oxley, D., Santos, F., Clarke, J., Mansfield, W., et al. (2014). Resetting Transcription Factor Control Circuitry toward Ground-State Pluripotency in Human. Cell *158*, 1254–1269.

Tan, G., Opitz, L., Schlapbach, R., and Rehrauer, H. (2019). Long fragments achieve lower base quality in Illumina paired-end sequencing. Sci. Rep. *9*, 1–7.

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., et al. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods *6*, 377–382.

Tao, L., Raz, O., Marx, Z., Biezuner, T., Amir, S., Milo, L., Adar, R., Onn, A., Chapal-Ilani, N., Berman, V., et al. (2018). A biological-computational human cell lineage discovery platform based on duplex molecular inversion probes. BioRxiv 191296.

Temin, H.M., and Mizutani, S. (1970). Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. Nature *226*, 1211–1213.

Tesar, P.J., Chenoweth, J.G., Brook, F.A., Davies, T.J., Evans, E.P., Mack, D.L., Gardner, R.L., and McKay, R.D.G. (2007). New cell lines from mouse epiblast share defining features with human embryonic stem cells. Nature *448*, 196–199.

Tewhey, R., Warner, J.B., Nakano, M., Libby, B., Medkova, M., David, P.H., Kotsopoulos, S.K., Samuels, M.L., Hutchison, J.B., Larson, J.W., et al. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. Nat. Biotechnol. *27*, 1025–1031.

Theunissen, T.W., Powell, B.E., Wang, H., Mitalipova, M., Faddah, D.A., Reddy, J., Fan, Z.P., Maetzel, D., Ganz, K., Shi, L., et al. (2014). Systematic Identification of Culture Conditions for Induction and Maintenance of Naive Human Pluripotency. Cell Stem Cell *15*, 471–487.

Thomas, S., Underwood, J.G., Tseng, E., Holloway, A.K., and Subcommittee, on behalf of the B.T.B.C.I. (2014). Long-Read Sequencing of Chicken Transcripts and Identification of New Transcript Isoforms. PLOS ONE *9*, e94650.

Thomson, J.A., Itskovitz-Eldor, J., Shapiro, S.S., Waknitz, M.A., Swiergiel, J.J., Marshall, V.S., and Jones, J.M. (1998). Embryonic Stem Cell Lines Derived from Human Blastocysts. Science *282*, 1145–1147.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat. Biotechnol. *32*, 381–386.

Turner, E.H., Lee, C., Ng, S.B., Nickerson, D.A., and Shendure, J. (2009). Massively parallel exon capture and library-free resequencing across 16 genomes. Nat. Methods *6*, 315–316.

Vallejos, C.A., Risso, D., Scialdone, A., Dudoit, S., and Marioni, J.C. (2017). Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat. Methods *14*, 565–571.

Vallier, L., Touboul, T., Chng, Z., Brimpari, M., Hannan, N., Millan, E., Smithers, L.E., Trotter, M., Rugg-Gunn, P., Weber, A., et al. (2009a). Early Cell Fate Decisions of Human Embryonic Stem Cells and Mouse Epiblast Stem Cells Are Controlled by the Same Signalling Pathways. PLOS ONE *4*, e6082.

Vallier, L., Mendjan, S., Brown, S., Chng, Z., Teo, A., Smithers, L.E., Trotter, M.W.B., Cho, C.H.-H., Martinez, A., Rugg-Gunn, P., et al. (2009b). Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. Development *136*, 1339–1349.

Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W. (1995). Serial Analysis of Gene Expression. Science *270*, 484–487.

Waddington, C.H. (1957). The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser. Strategy Genes Discuss. Some Asp. Theor. Biol. Append. H Kacser.

Ware, C.B. (2017). Concise Review: Lessons from Naïve Human Pluripotent Cells. STEM CELLS *35*, 35–41.

Warrier, S., Jeught, M.V. der, Duggal, G., Tilleman, L., Sutherland, E., Taelman, J., Popovic, M., Lierman, S., Lopes, S.C.D.S., Soom, A.V., et al. (2017). Direct comparison of distinct naive pluripotent states in human embryonic stem cells. Nat. Commun. *8*, ncomms15055.

Weinberger, L., Ayyash, M., Novershtern, N., and Hanna, J.H. (2016). Dynamic stem cell states: naive to primed pluripotency in rodents and humans. Nat. Rev. Mol. Cell Biol. *17*, 155–169.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

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

Yiangou, L., Ross, A.D.B., Goh, K.J., and Vallier, L. (2018). Human Pluripotent Stem Cell-Derived Endoderm for Modeling Development and Clinical Applications. Cell Stem Cell *22*, 485–499.

Yilmaz, A., and Benvenisty, N. (2019). Defining Human Pluripotency. Cell Stem Cell *25*, 9–22.

Ying, Q.-L., Wray, J., Nichols, J., Batlle-Morera, L., Doble, B., Woodgett, J., Cohen, P., and Smith, A. (2008). The ground state of embryonic stem cell self-renewal. Nature *453*, 519–523.

Yu, J., Vodyanik, M.A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J.L., Tian, S., Nie, J., Jonsdottir, G.A., Ruotti, V., Stewart, R., et al. (2007). Induced Pluripotent Stem Cell Lines Derived from Human Somatic Cells. Science *318*, 1917–1920.

Zhang, J., Ratanasirintrawoot, S., Chandrasekaran, S., Wu, Z., Ficarro, S.B., Yu, C., Ross, C.A., Cacchiarelli, D., Xia, Q., Seligson, M., et al. (2016). LIN28 Regulates Stem Cell Metabolism and Conversion to Primed Pluripotency. Cell Stem Cell *19*, 66–80.

Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nat. Commun. *8*, ncomms14049.

Zhu, Y. y., Machleder, E. m., Chenchik, A., Li, R., and Siebert, P. d. (2001). Reverse Transcriptase Template Switching: A SMART™ Approach for Full-Length cDNA Library Construction. BioTechniques *30*, 892–897.

Ziegenhain, C., Vieth, B., Parekh, S., Reinius, B., Guillaumet-Adkins, A., Smets, M., Leonhardt, H., Heyn, H., Hellmann, I., and Enard, W. (2017). Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol. Cell *65*, 631-643.e4.

Ziegenhain, C., Vieth, B., Parekh, S., Hellmann, I., and Enard, W. (2018). Quantitative single-cell transcriptomics. Brief. Funct. Genomics *17*, 220–232.

α Uzbas F, Opperer F, Shaposhnikov D, Drukker M. BART-seq: cost-effective massively parallel targeted sequencing for genomics and transcriptomics. GSE107723. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107723 (2019).

α Shaposhnikov D. Total RNA sequencing of a time course treatment of human embryonic stem cells with CHIR99021, recombinant Wnt3a, and a time course activation of constitutively active beta-catenin expression. GSE130381. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE130381 (2019).

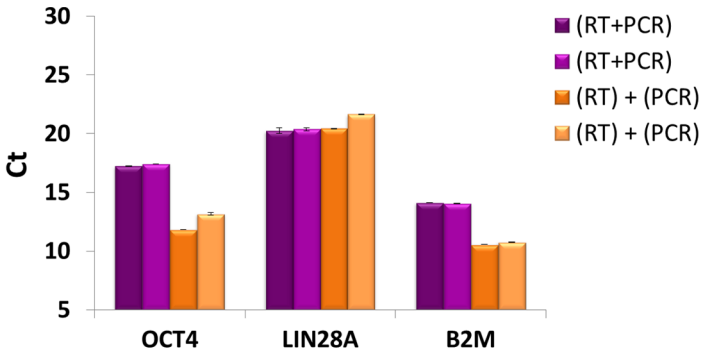
α Sass S, Angerer P, Uzbas F, Batra R, Müller N. Software required for Bart-Seq technology. Github. https://doi.org/10.5281/zenodo.3252205 (2019).

α Angerer P. Demultiplexing pipeline for BARTSeq. Github. https://doi.org/10.5281/zenodo.3251773 (2019).

α PrimerSelect - Bartender software suite, accessed 2 Nov 2019, <http://icb-bar.helmholtz-muenchen.de/primerselect>.

# APPENDIX A One-step RT+PCR

With an attempt to reduce the number of steps within the workflow, I ran a preliminary experiment to see whether the reverse transcription and PCR can be combined in a single reaction mixture. To this goal, I compared the reaction containing all the components of reverse transcription and PCR (42% v/v ratio of RT/PCR) to the regular two-step protocol (20% v/v ratio of RT/PCR). Among the tested genes, the efficiency of *B2M* and *OCT4* amplification was reduced when the reactions are combined (RT+PCR), while *LIN28A* remained the same. Given that the difference between the combined and decoupled reactions was only a few cycles in this preliminary experiment, and because I hypothesize that the reduced efficiency might be partially due to high RT/PCR v/v ratio, further enzyme concentrations and volume ratios could be tested to see whether the combined reaction can be efficient enough to replace the two-step protocol.

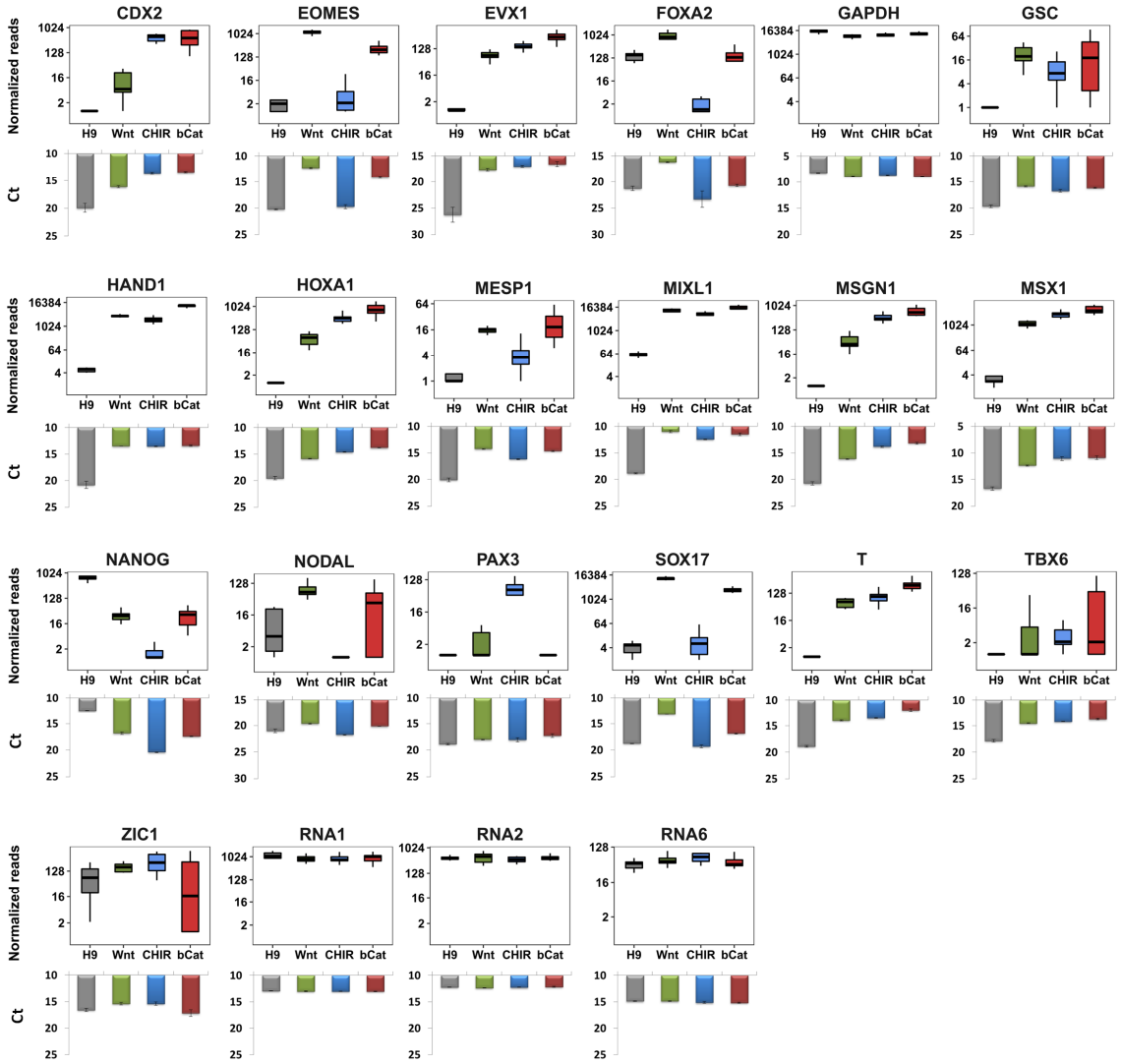



**Combining reverse transcription and PCR in a single reaction.** Comparison of the one-step -(RT+PCR)- and two-step -(RT)+(PCR)- reverse transcription and PCR. Concentration of the reverse transcriptase during reverse transcription and the template amounts in the PreAmp PCR were equal in both reactions. Shades of the same colors indicate the technical replicates and error bars indicate the qPCR replicates.

## APPENDIX B Validation of the mesoderm primer set with qPCR and BART-Seq using bulk RNA samples

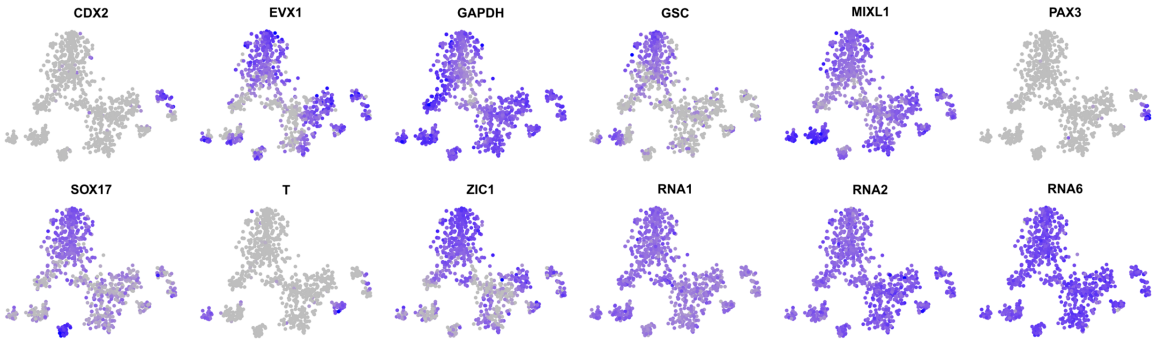
Mesoderm (Wnt stimulation) primers were used to pre-amplify bulk RNA samples isolated at 0 h and 72 h of the Wnt pathway stimulation (**Figure 46**), which were analyzed either with qPCR or BART-Seq.

Upper panels show the normalized reads from RNA samples that were aliquoted (50 pg per well), barcoded with 10 different combinations, and analyzed with the BART-Seq protocol together with single cells. Lower panels show the Ct values obtained with qPCR using nested primers to analyze the pre-amplified bulk samples from a biological replicate experiment. The two analyses displayed remarkably similar patterns, which also resembled the previous bulk RNA-Seq results (**Figure 47**).

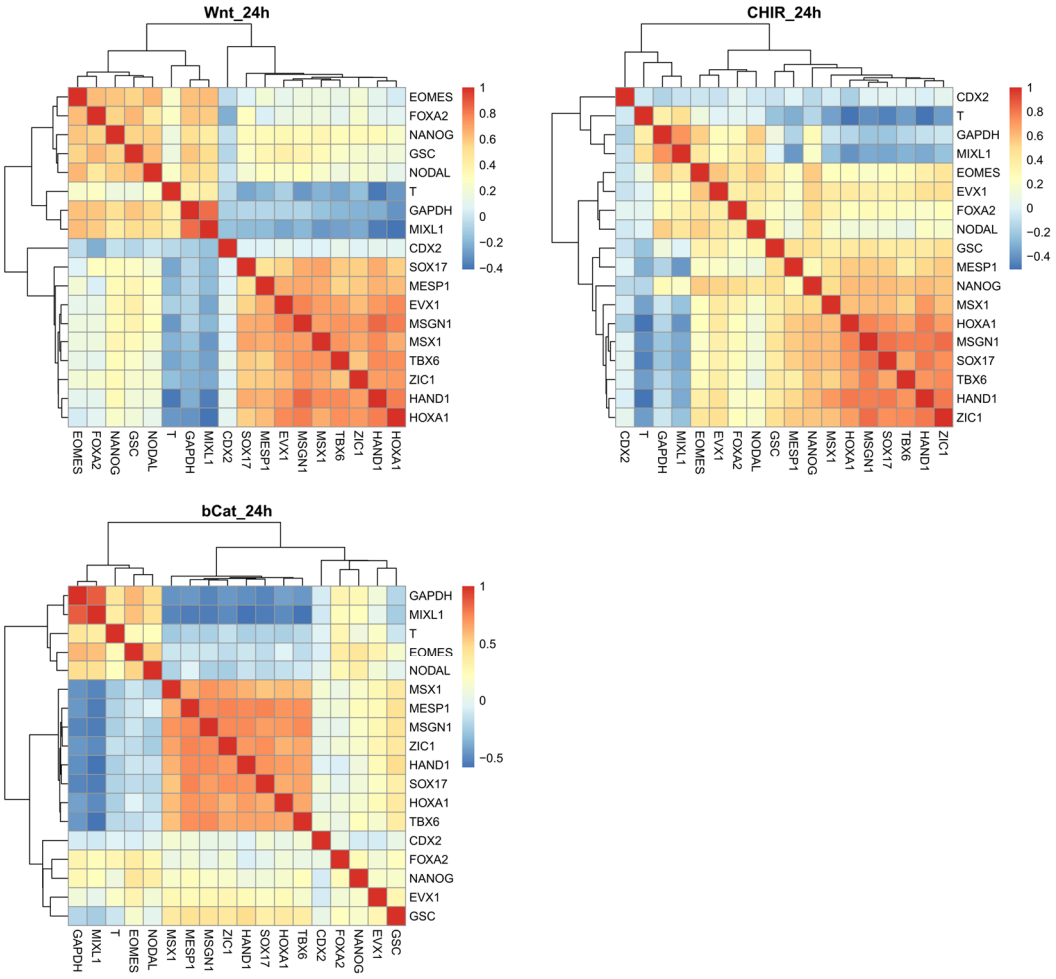


# APPENDIX C Additional data for the Wnt pathway stimulation experiment

tSNE distribution of the rest of the genes not shown in (**Figure 48**):

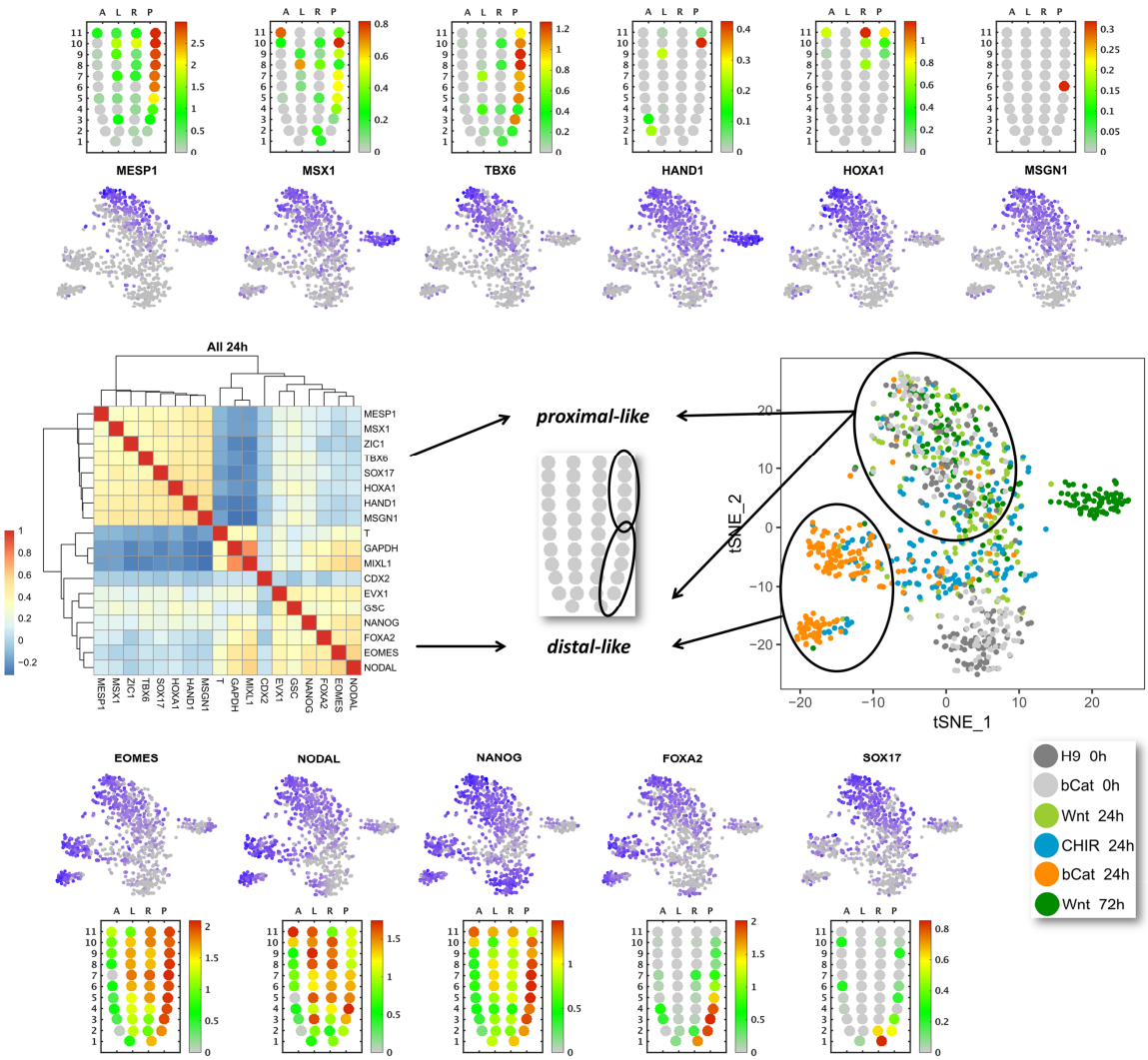



Heatmaps of the pairwise gene correlations at 24 h for each treatment separately (related to **Figure 48**):



## APPENDIX D Biological repetition of the Wnt stimulation experiment

Biological repetition of the Wnt stimulation experiment (Related to **Figure 48**). Heatmap of the pairwise gene correlations at 24 h calculated based on single cells from the three treatments (left) and two-dimensional representation (tSNE) of all the single cells sampled at 0 h, 24 h, and 72 h, based on the expression of 19 genes (right). The 72 h time point contains cells only from rWnt3a treatment due to the loss of samples from the other two conditions. Distribution of some selected genes underlying the tSNE plot is shown in the upper and lower panels.



# APPENDIX E Simplified R code for the correction & normalization of the data

```
library(readxl)
library(reshape2)
library(MASS)   # for glm
library(plyr)   # For count

# Read the count matrix
dat = read_excel("./Raw Data/2019-02-14 NGS15 Read Counts Formatted.xlsx", sheet = "Lib1",
col_names = TRUE, col_types = NULL, na = "", skip = 0)

# Retrieve (all) the gene names and spike-in names
genes = colnames(dat)[min(grep("^B2M|^RNA|^ZFP42",colnames(dat))) :
max(grep("^B2M|^RNA|^ZFP42",colnames(dat)))]
spikes = colnames(dat)[grep("^RNA",colnames(dat))]


#--------------      CLEAN UP THE DATA      ------------------------------------------------------

#--------------      Manual Clean-up
datclean = dat
columns = c("left","right","well.y","well.x","cell.count","cells","well.location")

# Subset the relevant columns and the spike-in part of the original data
spikes.all = data.frame(dat[c(columns,spikes)])

# Remove the very bad barcodes and very bad combination with genes: L28 (DNMT3B), L44 (LIN28A), R23
(POU5F1)
spikes.all = subset(spikes.all, !(left %in% c("L24","L47","L28","L44")))
spikes.all = subset(spikes.all, !(right %in% c("R23")))

# Melt the data frame
melted.spikes = melt(spikes.all,id.vars=columns, na.rm=TRUE)

#--------------      Determine the outliers that do not fit into the full model

# Full Model
mp2 = glm.nb(value ~ variable + left + right + cells + well.location + left:variable + right:variable, data =
melted.spikes)

melted.spikes$prd = predict.glm(mp2, melted.spikes, type="response")   # Predictions
melted.spikes$res = abs(residuals(mp2))       # Absolute value of residuals
melted.spikes$rem = "no"                      # Remove? First set everything to "no"
melted.spikes$rem[melted.spikes$res > 2] = "yes" # Remove if the prediction is more than two-fold different

# Data without outliers
melted.spikes.filtered = subset(melted.spikes, rem == "no")[c(columns,"variable","value")]

# Full model again, without outliers
mp3 = glm.nb(value ~ variable + left + right + cells + well.location + left:variable + right:variable,
        data = melted.spikes.filtered)

melted.spikes.filtered$prd = predict.glm(mp3, melted.spikes.filtered, type="response")
melted.spikes.filtered$res = abs(residuals(mp3))
```

```
# Flag: completely remove a well if there were more than 2 NAs
Flags = subset(count(subset(melted.spikes, rem=="yes"), c("left","right")), freq > 2)
melted.spikes.2 = subset(melted.spikes, !paste0(left,right) %in%
paste0(Flags$left,Flags$right))[c(columns,"variable","value","rem")]

# Replace the remaining outliers with the predictions of the mp3
melted.spikes.2$prd = predict.glm(mp3, melted.spikes.2, type="response")
melted.spikes.2$value[which(melted.spikes.2$rem == "yes")] =
melted.spikes.2$prd[which(melted.spikes.2$rem == "yes")]

#----------      MODEL & CORRECT FOR COMBINATION EFFECT      ----------

melted.spikes.3 = melted.spikes.2[c(columns,"variable","value")]
melted.spikes.3$value = round(melted.spikes.3$value, 0)  # Since glm.nb requires integers, round the values

# Calculate the correction factor

# Model with interactions
mp1 = glm.nb(value ~ variable + left + right + left:variable + right:variable, data = melted.spikes.3)
# Model without interactions
mp0 = glm.nb(value ~ variable + left + right, data = melted.spikes.3)

melted.spikes.3$prd3 = predict.glm(mp3, melted.spikes.3, type="response")  # Full predictions (filtered
model)
melted.spikes.3$prd1 = predict.glm(mp1, melted.spikes.3, type="response")  # With interactions
melted.spikes.3$prd0 = predict.glm(mp0, melted.spikes.3, type="response")  # Without interactions

# Correction factor: fold-change of the interactions relative to the average (division)
melted.spikes.3$fold = melted.spikes.3$prd1/melted.spikes.3$prd0

# Correct the read counts
melted.spikes.3$corF = melted.spikes.3$value/melted.spikes.3$fold

# Data frame of corrected spike-ins
corrected.spikes = dcast(melted.spikes.3, left + right + well.y + well.x + cell.count + cells + well.location
                  ~ variable, value.var="corF")

#----------      NORMALIZE      ----------

# Replace the spike-ins with corrected versions (only overlapping ones merged):
datclean = merge(dat[ ,-which(colnames(dat) %in% spikes)], corrected.spikes, by = columns)

datnormal = merge(datclean,dat[c("left","right",spikes)], by = c("left","right"), suffixes = c("","nc"))  # Add back
the raw spike-ins

# Non-corrected spikes:
ncspikes = paste0(spikes,"nc")

# Calculate the normalization factor per well
datnormal$RNA_X = 2^rowMeans(log2(datnormal[spikes]+1))-1
min = median(datnormal$RNA_X, na.rm=TRUE)/5
max = median(datnormal$RNA_X, na.rm=TRUE)*5
datnormal = subset(datnormal, (min < RNA_X) & (RNA_X < max))
factor = datnormal$RNA_X / median(datnormal$RNA_X, na.rm=TRUE)

# Normalize the data
datnormal[c(genes,ncspikes,"RNA_X")] = datnormal[c(genes,ncspikes,"RNA_X")] / factor  # Normalize
```

# APPENDIX F Barcode panel used for transcriptomics experiments

Barcode sequences used for transcriptomics experiments (CCA+Barcode+Linker)

**Forward Barcodes (5'>3')**

| | |
|---|---|
| L01 | CCACCTTCTCGATGCGCATTC |
| L02 | CCACCAGTAGCATGCGCATTC |
| L03 | CCACCTAACGCATGCGCATTC |
| L04 | CCAGGCCGATTATGCGCATTC |
| L05 | CCACGTTCGTCATGCGCATTC |
| L06 | CCAGCCTGGTAATGCGCATTC |
| L07 | CCACCAGACGAATGCGCATTC |
| L08 | CCAACGACACCATGCGCATTC |
| L09 | CCAGCCTCTTGATGCGCATTC |
| L10 | CCATACGGCACATGCGCATTC |
| L11 | CCACTCCAGTCATGCGCATTC |
| L12 | CCAAACGTCGGATGCGCATTC |
| L13 | CCAGTTGGAGCATGCGCATTC |
| L14 | CCAACGAGCCTATGCGCATTC |
| L15 | CCAGATCCAGGATGCGCATTC |
| L16 | CCATCCTGGCTATGCGCATTC |
| L17 | CCAGGTCCATGATGCGCATTC |
| L18 | CCACGGTGAGTATGCGCATTC |
| L19 | CCACCAACTCCATGCGCATTC |
| L20 | CCAAGGCCGAAATGCGCATTC |
| L21 | CCAGAGGCCTAATGCGCATTC |
| L22 | CCATTCTCCGCATGCGCATTC |
| L23 | CCAGTGACCACATGCGCATTC |
| L24 | CCACTGTACGGATGCGCATTC |
| L25 | CCACCAGTGGTATGCGCATTC |
| L26 | CCAATAGCCGCATGCGCATTC |
| L27 | CCACTAGCCTGATGCGCATTC |
| L28 | CCACCTTAGGCATGCGCATTC |
| L29 | CCACTACGGTCATGCGCATTC |
| L30 | CCAGAAGTCGCATGCGCATTC |
| L31 | CCATCCACGACATGCGCATTC |
| L32 | CCAGTCTGCCAATGCGCATTC |
| L33 | CCAGGAACTCCATGCGCATTC |
| L34 | CCAACCGGTCTATGCGCATTC |
| L35 | CCAACCGAGGTATGCGCATTC |
| L36 | CCACTGGCTACATGCGCATTC |
| L37 | CCATCCTCCGTATGCGCATTC |
| L38 | CCAGAGCTAGGATGCGCATTC |
| L39 | CCAGCCACGTAATGCGCATTC |
| L40 | CCACGGTCTTCATGCGCATTC |
| L41 | CCAGGTGTACCATGCGCATTC |
| L42 | CCATCGGCTCAATGCGCATTC |
| L43 | CCAAGTAGGCCATGCGCATTC |
| L44 | CCAACCTCCTGATGCGCATTC |
| L45 | CCAAGACGGACATGCGCATTC |
| L46 | CCACGGTCCAAATGCGCATTC |
| L47 | CCACGGTGACAATGCGCATTC |
| L48 | CCAGGCCTTCAATGCGCATTC |
| L49 | CCAACCGGAACATGCGCATTC |
| L50 | CCACTCCTCTGATGCGCATTC |
| L51 | CCACGTCGTTGATGCGCATTC |
| L52 | CCAGAGGACGAATGCGCATTC |
| L53 | CCAACGTTCCGATGCGCATTC |
| L54 | CCAACTGGCCAATGCGCATTC |
| L55 | CCAGTCGTAGCATGCGCATTC |
| L56 | CCATTGTGGCCATGCGCATTC |
| L57 | CCAGCCGGAATATGCGCATTC |
| L58 | CCACCAACCAGATGCGCATTC |
| L59 | CCAGACCTTCCATGCGCATTC |
| L60 | CCACTTGCCTCATGCGCATTC |
| L61 | CCAGTGGCCAAATGCGCATTC |
| L62 | CCATCTAGGCCATGCGCATTC |
| L63 | CCACGATTCGGATGCGCATTC |
| L64 | CCAGTCGTTCCATGCGCATTC |
| L65 | CCAGAGAGGCTATGCGCATTC |
| L66 | CCAGCTCCGATATGCGCATTC |
| L67 | CCAACTCCTCCATGCGCATTC |
| L68 | CCACCTCCAAGATGCGCATTC |
| L69 | CCAGGACTCGAATGCGCATTC |
| L70 | CCAAGACTCCGATGCGCATTC |
| L71 | CCACTCCGGTTATGCGCATTC |
| L72 | CCACCACAGGTATGCGCATTC |
| L73 | CCACACCTACCATGCGCATTC |
| L74 | CCAACTCCGGAATGCGCATTC |
| L75 | CCAAACCTGGCATGCGCATTC |
| L76 | CCAGTCTGTCGATGCGCATTC |
| L77 | CCACCATCACGATGCGCATTC |
| L78 | CCACGTTAGCCATGCGCATTC |
| L79 | CCAGGAGCCTTATGCGCATTC |
| L80 | CCACCGGTATGATGCGCATTC |
| L81 | CCAGGTCGATCATGCGCATTC |
| L82 | CCAACTGAGGCATGCGCATTC |
| L83 | CCACCGGTGTTATGCGCATTC |
| L84 | CCACTACTCCGATGCGCATTC |

**Reverse Barcodes (5'>3')**

| | |
|---|---|
| R01 | CCAGCCACCTTAGCGTAACCT |
| R02 | CCAGAACCACGAGCGTAACCT |
| R03 | CCACTCAGGCAAGCGTAACCT |
| R04 | CCAGTACAGCGAGCGTAACCT |
| R05 | CCACTCGAGCTAGCGTAACCT |
| R06 | CCAGCGACACAAGCGTAACCT |
| R07 | CCAGTCATCGGAGCGTAACCT |
| R08 | CCATGCCTCCAAGCGTAACCT |
| R09 | CCACCGATTCCAGCGTAACCT |
| R10 | CCAGCATAGGCAGCGTAACCT |
| R11 | CCAGACAGAGGAGCGTAACCT |
| R12 | CCACACCGGTAAGCGTAACCT |
| R13 | CCAGTGGAAGGAGCGTAACCT |
| R14 | CCAGGATCCTCAGCGTAACCT |
| R15 | CCAGGACAGTGAGCGTAACCT |
| R16 | CCATGTGGCGTAGCGTAACCT |
| R17 | CCAGCGATCAGAGCGTAACCT |
| R18 | CCAACAAGGCCAGCGTAACCT |
| R19 | CCACTCGGAGAAGCGTAACCT |
| R20 | CCACAAGCTGCAGCGTAACCT |
| R21 | CCAGTCCACCAAGCGTAACCT |
| R22 | CCATCGAGGACAGCGTAACCT |
| R23 | CCATCCGGCAAAGCGTAACCT |
| R24 | CCAGTCTCACGAGCGTAACCT |
| R25 | CCAGATCGGCTAGCGTAACCT |
| R26 | CCAGGCGGATAAGCGTAACCT |
| R27 | CCACAGAGTCGAGCGTAACCT |
| R28 | CCAGCGAACCTAGCGTAACCT |
| R29 | CCACCTGTCGTAGCGTAACCT |
| R30 | CCAGACCACACAGCGTAACCT |
| R31 | CCAAGGAGAGCAGCGTAACCT |
| R32 | CCATCACACCGAGCGTAACCT |
| R33 | CCAAGCACCAGAGCGTAACCT |
| R34 | CCAGCCAGAACAGCGTAACCT |
| R35 | CCAGATTGGCCAGCGTAACCT |
| R36 | CCAAGCGGAGTAGCGTAACCT |
| R37 | CCACCTCACCTAGCGTAACCT |
| R38 | CCAACCTGCAGAGCGTAACCT |
| R39 | CCAGGCAAGCTAGCGTAACCT |
| R40 | CCAGGAGAGGAAGCGTAACCT |
| R41 | CCACCACTCGTAGCGTAACCT |
| R42 | CCATCAGCGAGAGCGTAACCT |
| R43 | CCAACTCGTCGAGCGTAACCT |
| R44 | CCACTTCGGACAGCGTAACCT |
| R45 | CCAGGAGTTCCAGCGTAACCT |
| R46 | CCAGAGCAGTCAGCGTAACCT |
| R47 | CCAGAAGCCTCAGCGTAACCT |
| R48 | CCAGGAGGAACAGCGTAACCT |
| R49 | CCACACCTAGGAGCGTAACCT |
| R50 | CCAGTGAGTCGAGCGTAACCT |
| R51 | CCAGACAGGCAAGCGTAACCT |
| R52 | CCATGCGGACTAGCGTAACCT |
| R53 | CCAGCGTCCTAAGCGTAACCT |
| R54 | CCAACCGTCCAAGCGTAACCT |
| R55 | CCACCAGGACTAGCGTAACCT |
| R56 | CCACTGACCGTAGCGTAACCT |
| R57 | CCACACACCTCAGCGTAACCT |
| R58 | CCACCAGCAAGAGCGTAACCT |
| R59 | CCAGGAGACCAAGCGTAACCT |
| R60 | CCAGCAGGATGAGCGTAACCT |
| R61 | CCATCGGATCGAGCGTAACCT |
| R62 | CCATCACCGTCAGCGTAACCT |
| R63 | CCATGTCGGAGAGCGTAACCT |
| R64 | CCAACAGCTCGAGCGTAACCT |
| R65 | CCACCACAACCAGCGTAACCT |
| R66 | CCATCCTCAGGAGCGTAACCT |
| R67 | CCATGGCGTTCAGCGTAACCT |
| R68 | CCAGGTCAAGGAGCGTAACCT |
| R69 | CCATGGCACTGAGCGTAACCT |
| R70 | CCATCGACCAGAGCGTAACCT |
| R71 | CCAAGCCAGCTAGCGTAACCT |
| R72 | CCATCTCGTGCAGCGTAACCT |
| R73 | CCAAAGGCCTGAGCGTAACCT |
| R74 | CCAATGCCTCGAGCGTAACCT |
| R75 | CCACCATCGCAAGCGTAACCT |
| R76 | CCATTGAGGCGAGCGTAACCT |
| R77 | CCACCAAGCCTAGCGTAACCT |
| R78 | CCAGTGGCTGAAGCGTAACCT |
| R79 | CCACCGAAGAGAGCGTAACCT |
| R80 | CCAGGACTAGCAGCGTAACCT |
| R81 | CCACCGAACACAGCGTAACCT |
| R82 | CCAGAGTCCACAGCGTAACCT |

## APPENDIX G Barcode panels used for genotyping experiments

Barcode squences used for *BRCA* genotyping experiments (CCA+Barcode+Linker)

| Forward Barcodes (5'>3') | |
|---|---|
| L01 | CCATCCTCAGGTAGCGACGAG |
| L02 | CCACCGAACACTAGCGACGAG |
| L03 | CCAACTGAGGCTAGCGACGAG |
| L04 | CCACACCTAGGTAGCGACGAG |
| L05 | CCACACACCTCTAGCGACGAG |
| L06 | CCATACGGCACTAGCGACGAG |
| L07 | CCAGATCCAGGTAGCGACGAG |
| L08 | CCACCAGACGATAGCGACGAG |

| Reverse Barcodes (5'>3') | |
|---|---|
| R01 | CCAACGCGCTACCATACGACG |
| R02 | CCAACGCTAGCCCATACGACG |
| R03 | CCAACTCCTCCCCATACGACG |
| R04 | CCAGCGCATCTCCATACGACG |
| R05 | CCACGGACAAGCCATACGACG |
| R06 | CCACGGTCCAACCATACGACG |
| R07 | CCAGGACGCAACCATACGACG |
| R08 | CCAGACCTTCCCCATACGACG |
| R09 | CCAAGGAGAGCCCATACGACG |
| R10 | CCACCATCACGCCATACGACG |
| R11 | CCACGGCAACACCATACGACG |
| R12 | CCACGGAATGCCCATACGACG |

Barcode sequences used for protection group evaluation (NNN+Barcode+Linker)

| Reverse Barcodes (5'>3') | |
|---|---|
| Barcode 1 | NNNTTGCGCAGCCATACGACG |
| Barcode 2 | NNNCGGACTAGCCATACGACG |
| Barcode 3 | NNNCATTCGCGCCATACGACG |
| Barcode 4 | NNNTGCTCGACCCATACGACG |
| Barcode 5 | NNNCGACTACGCCATACGACG |
| Barcode 6 | NNNGCCGACAACCATACGACG |
| Barcode 7 | NNNCTCTACGCCCATACGACG |
| Barcode 8 | NNNCACGAGTGCCATACGACG |

| Forward Barcodes (5'>3') | |
|---|---|
| Barcode A | NNNGGATAGGCTAGCGACGAG |
| Barcode B | NNNCCGACCTATAGCGACGAG |
| Barcode C | NNNTGGAGACCTAGCGACGAG |
| Barcode D | NNNAGCCGACTTAGCGACGAG |
| Barcode E | NNNCCTAGACCTAGCGACGAG |
| Barcode F | NNNTGGATGGCTAGCGACGAG |
| Barcode G | NNNCCAGATCCTAGCGACGAG |
| Barcode H | NNNTGATCCGCTAGCGACGAG |
| Barcode I | NNNCTTCCGACTAGCGACGAG |
| Barcode J | NNNTCCGATGCTAGCGACGAG |

# APPENDIX H Sample configuration file for the PrimerSelect tool

```
Primer3 File - http://primer3.sourceforge.net
P3_FILE_TYPE=settings

PRIMER_DNA_CONC=10.0
PRIMER_DNTP_CONC=0.6
PRIMER_EXPLAIN_FLAG=1
PRIMER_FIRST_BASE_INDEX=1
PRIMER_GC_CLAMP=0
PRIMER_INSIDE_PENALTY=-1.0
PRIMER_INTERNAL_DNA_CONC=50.0
PRIMER_INTERNAL_DNTP_CONC=0.0
PRIMER_INTERNAL_MAX_GC=80.0
PRIMER_INTERNAL_MAX_HAIRPIN_TH=47.00
PRIMER_INTERNAL_MAX_LIBRARY_MISHYB=12.00
PRIMER_INTERNAL_MAX_NS_ACCEPTED=0
PRIMER_INTERNAL_MAX_POLY_X=5
PRIMER_INTERNAL_MAX_SELF_ANY=12.00
PRIMER_INTERNAL_MAX_SELF_ANY_TH=47.00
PRIMER_INTERNAL_MAX_SELF_END=12.00
PRIMER_INTERNAL_MAX_SELF_END_TH=47.00
PRIMER_INTERNAL_MAX_SIZE=27
PRIMER_INTERNAL_MAX_TM=63.0
PRIMER_INTERNAL_MIN_GC=20.0
PRIMER_INTERNAL_MIN_QUALITY=0
PRIMER_INTERNAL_MIN_SIZE=18
PRIMER_INTERNAL_MIN_TM=57.0
PRIMER_INTERNAL_OPT_GC_PERCENT=50.0
PRIMER_INTERNAL_OPT_SIZE=20
PRIMER_INTERNAL_OPT_TM=60.0
PRIMER_INTERNAL_SALT_DIVALENT=1.5
PRIMER_INTERNAL_SALT_MONOVALENT=50.0
PRIMER_INTERNAL_WT_END_QUAL=0.0
PRIMER_INTERNAL_WT_GC_PERCENT_GT=0.0
PRIMER_INTERNAL_WT_GC_PERCENT_LT=0.0
PRIMER_INTERNAL_WT_HAIRPIN_TH=0.0
PRIMER_INTERNAL_WT_LIBRARY_MISHYB=0.0
PRIMER_INTERNAL_WT_NUM_NS=0.0
PRIMER_INTERNAL_WT_SELF_ANY=0.0
PRIMER_INTERNAL_WT_SELF_ANY_TH=0.0
PRIMER_INTERNAL_WT_SELF_END=0.0
PRIMER_INTERNAL_WT_SELF_END_TH=0.0
PRIMER_INTERNAL_WT_SEQ_QUAL=0.0
PRIMER_INTERNAL_WT_SIZE_GT=1.0
PRIMER_INTERNAL_WT_SIZE_LT=1.0
PRIMER_INTERNAL_WT_TM_GT=1.0
PRIMER_INTERNAL_WT_TM_LT=1.0
PRIMER_LIBERAL_BASE=1
PRIMER_LIB_AMBIGUITY_CODES_CONSENSUS=0
PRIMER_LOWERCASE_MASKING=0
PRIMER_MAX_END_GC=5
PRIMER_MAX_END_STABILITY=9.0
PRIMER_MAX_GC=70.0
PRIMER_MAX_HAIRPIN_TH=24.0
PRIMER_MAX_LIBRARY_MISPRIMING=12.00
PRIMER_MAX_NS_ACCEPTED=0
PRIMER_MAX_POLY_X=4
PRIMER_MAX_SELF_ANY=8.00
PRIMER_MAX_SELF_ANY_TH=45.0
PRIMER_MAX_SELF_END=3.00
PRIMER_MAX_SELF_END_TH=35.0
PRIMER_MAX_SIZE=29
PRIMER_MAX_TEMPLATE_MISPRIMING=12.00
PRIMER_MAX_TEMPLATE_MISPRIMING_TH=40.00
PRIMER_MAX_TM=66.0
PRIMER_MIN_3_PRIME_OVERLAP_OF_JUNCTION=4
PRIMER_MIN_5_PRIME_OVERLAP_OF_JUNCTION=7
PRIMER_MIN_END_QUALITY=0
PRIMER_MIN_GC=25.0
PRIMER_MIN_LEFT_THREE_PRIME_DISTANCE=3
PRIMER_MIN_QUALITY=0
PRIMER_MIN_RIGHT_THREE_PRIME_DISTANCE=3
PRIMER_MIN_SIZE=17
PRIMER_MIN_TM=60
PRIMER_NUM_RETURN=5
PRIMER_OPT_GC_PERCENT=50.0
PRIMER_OPT_SIZE=20
PRIMER_OPT_TM=63.0
PRIMER_OUTSIDE_PENALTY=0
PRIMER_PAIR_MAX_COMPL_ANY=8.00
PRIMER_PAIR_MAX_COMPL_ANY_TH=45.0
PRIMER_PAIR_MAX_COMPL_END=3.00
PRIMER_PAIR_MAX_COMPL_END_TH=35.0
PRIMER_PAIR_MAX_DIFF_TM=5.0
PRIMER_PAIR_MAX_LIBRARY_MISPRIMING=20.00
PRIMER_PAIR_MAX_TEMPLATE_MISPRIMING=24.00
PRIMER_PAIR_MAX_TEMPLATE_MISPRIMING_TH=70.00
PRIMER_PAIR_WT_COMPL_ANY=0.0
PRIMER_PAIR_WT_COMPL_ANY_TH=0.0
PRIMER_PAIR_WT_COMPL_END=0.0
PRIMER_PAIR_WT_COMPL_END_TH=0.0
PRIMER_PAIR_WT_DIFF_TM=0.0
PRIMER_PAIR_WT_IO_PENALTY=0.0
PRIMER_PAIR_WT_LIBRARY_MISPRIMING=0.0
PRIMER_PAIR_WT_PRODUCT_SIZE_GT=0.0
PRIMER_PAIR_WT_PRODUCT_SIZE_LT=0.0
PRIMER_PAIR_WT_PRODUCT_TM_GT=0.0
PRIMER_PAIR_WT_PRODUCT_TM_LT=0.0
PRIMER_PAIR_WT_PR_PENALTY=1.0
PRIMER_PAIR_WT_TEMPLATE_MISPRIMING=0.0
PRIMER_PAIR_WT_TEMPLATE_MISPRIMING_TH=0.0
PRIMER_PICK_ANYWAY=1
PRIMER_PICK_INTERNAL_OLIGO=0
PRIMER_PICK_LEFT_PRIMER=1
PRIMER_PICK_RIGHT_PRIMER=1
PRIMER_PRODUCT_MAX_TM=1000000.0
PRIMER_PRODUCT_MIN_TM=-1000000.0
PRIMER_PRODUCT_OPT_TM=0.0
PRIMER_PRODUCT_SIZE_RANGE=90-200 90-248 85-248 80-
248 75-248
PRIMER_QUALITY_RANGE_MAX=100
PRIMER_QUALITY_RANGE_MIN=0
PRIMER_SALT_CORRECTIONS=1
PRIMER_SALT_DIVALENT=6.0
PRIMER_SALT_MONOVALENT=25.0
PRIMER_SEQUENCING_ACCURACY=20
PRIMER_SEQUENCING_INTERVAL=250
PRIMER_SEQUENCING_LEAD=50
PRIMER_SEQUENCING_SPACING=500
PRIMER_TASK=generic
PRIMER_MUST_MATCH_THREE_PRIME=nnnna
PRIMER_THERMODYNAMIC_OLIGO_ALIGNMENT=1
PRIMER_THERMODYNAMIC_TEMPLATE_ALIGNMENT=0
PRIMER_TM_FORMULA=1
PRIMER_WT_END_QUAL=0.0
PRIMER_WT_END_STABILITY=0.0
PRIMER_WT_GC_PERCENT_GT=0.0
PRIMER_WT_GC_PERCENT_LT=0.0
PRIMER_WT_HAIRPIN_TH=0.0
PRIMER_WT_LIBRARY_MISPRIMING=0.0
PRIMER_WT_NUM_NS=0.0
PRIMER_WT_POS_PENALTY=0.0
PRIMER_WT_SELF_ANY=0.0
PRIMER_WT_SELF_ANY_TH=0.0
PRIMER_WT_SELF_END=0.0
PRIMER_WT_SELF_END_TH=0.0
PRIMER_WT_SEQ_QUAL=0.0
PRIMER_WT_SIZE_GT=1.0
PRIMER_WT_SIZE_LT=1.0
PRIMER_WT_TEMPLATE_MISPRIMING=0.0
PRIMER_WT_TEMPLATE_MISPRIMING_TH=0.0
PRIMER_WT_TM_GT=1.0
PRIMER_WT_TM_LT=1.0
=
```

## APPENDIX I Genotyping primers for protection group evaluation

The *BRCA* genotyping primers used in combination with NNN Barcodes (**APPENDIX G**) for screening 5' trinucleotide protection groups

| # | Gene | Targeted Locus on Chromosome* | Amplicon name | Amplicon Size§ (bp) | Primer | Nested (sequencing) primer sequence (5' > 3') |
|---|---|---|---|---|---|---|
| 1 | *BRCA1* | chr17: 43123944-43124076 | **Amp1'** | 133 | Forward | CGTTGAAGAAGTACAAAATGTCA |
| | | | | | Reverse | AGGTCAATTCTGTTCATTTGCA |
| 2 | *BRCA1* | chr17: 43056890-43057086 | **Amp2'** | 197 | Forward | GTCCAAAGCGAGCAAGAGAA |
| | | | | | Reverse | TGGTTGGGATGGAAGAGTGA |
| 3 | *BRCA1* | chr17: 43092572-43092654 | **Amp3'** | 83 | Forward | AGGCAACGAAACTGGACTCA |
| | | | | | Reverse | TGATGGGAAAAAGTGGTGGTA |
| 4 | *BRCA1* | chr17: 43063766-43063946 | **Amp4'** | 181 | Forward | GAGTTTGTGTGTGAACGGACA |
| | | | | | Reverse | GGTAACTCAGACTCAGCATCA |
| 5 | *BRCA1* | chr17: 43094438-43094574 | **Amp5'** | 137 | Forward | CAGATGGGCTGGAAGTAAGGA |
| | | | | | Reverse | TAGGATTCTCTGAGCATGGCA |
| 6 | *BRCA2* | chr13: 32340412 + 32340606 | **Amp6'** | 195 | Forward | CGAACATTCAGACCAGCTCA |
| | | | | | Reverse | TCAAATTCCTCTAACACTCCCTTA |
| 7 | *BRCA2* | chr13: 32370880 + 32371097 | **Amp7'** | 218 | Forward | ACTGTGCCTGGCCTGATACAA |
| | | | | | Reverse | CATGTTCTTCAAATTCCTCCTGA |
| 8 | *BRCA2* | chr13: 32316504 + 32316691 | **Amp8'** | 188 | Forward | TTAAGACACGCTGCAACAAA |
| | | | | | Reverse | GGTTAACCTGCAAACGATGA |
| 9 | *BRCA2* | chr13: 32346866 + 32346962 | **Amp9'** | 97 | Forward | CTTTAGAGCCGATTACCTGTGTA |
| | | | | | Reverse | TCATTTATAAAAACGAGACTTTTCTCA |
| 10 | *BRCA1* | chr17: 43047589-43047773 | **Amp10'** | 185 | Forward | AATGATGAAGTGACAGTTCCA |
| | | | | | Reverse | ACCAAACCCATGCAAAAGGA |

* Based on the *In-Silico* PCR tool of UCSC Genome Browser

## APPENDIX J Pluripotency primers

Pluripotency and control genes selected for human pluripotent stem cells and sequences of the primers designed to target them

| # | Gene | Type | Targeted Locus on Chromosome§ | Amplicon Size§ (bp) | Primer | Nested (sequencing) primer sequence (5'>3') |
|---|---|---|---|---|---|---|
| 1 | ***B2M*** | Housekeeping | chr15: 44,715,652 - 44,717,620 | 92* | Forward | CTGCCGTGTGAACCATGTGA |
| | | | | | Reverse | CGGCATCTTCAAACCTCCATGA |
| 2 | ***CCND1*** | Cell cycle | chr11: 69,652,219 - 69,652,355 | 137 | Forward | CCAGCTCACGTCCAGGTTCA |
| | | | | | Reverse | CCCTCCCTGCACACACAACA |
| 3 | ***CCNE1*** | Cell cycle | chr19: 29,817,487 - 29,820,778 | 132* | Forward | ACACCCTCTTCTGCAGCCAA |
| | | | | | Reverse | TGTGTCGCCATATACCGGTCA |
| 4 | ***CER1*** | Differentiation | chr9: 14,720,247 - 14,722,182 | 157* | Forward | CAGTGCCCTTCAGCCAGACTA |
| | | | | | Reverse | GTGGTGAACTTGGCAGGCAA |
| 5 | ***DNMT3B*** | Pluripotency | chr20: 32,796,852 - 32,798,474 | 146* | Forward | CTCTGTCAGACATGCCGGGA |
| | | | | | Reverse | TCCACACAGAAACACCGGCA |
| 6 | ***GAPDH*** | Housekeeping | chr12: 6,534,834 - 6,536,753 | 198 / 328* | Forward | TGGGGAAGGTGAAGGTCGGA |
| | | | | | Reverse | GCTTCCCGTTCTCAGCCTTGA |
| 7 | ***LIN28A*** | Pluripotency | chr1: 26,429,287 - 26,429,385 | 99 | Forward | GTCAGGAGGCCAAGAAAGGGA |
| | | | | | Reverse | CAATCTTGTGGCCACTTTGACATAA |
| 8 | ***NANOG*** | Pluripotency | chr12: 7,795,613 - 7,795,789 | 177 | Forward | CGCCCTGCCTAGAAAAGACA |
| | | | | | Reverse | CAAAGCCTCCCAATCCCAAACA |
| 9 | ***POU5F1*** | Pluripotency | chr6: 31,165,590 - 31,165,684 | 95 | Forward | TTGTCAGCTTCCTCCACCCA |
| | | | | | Reverse | ACGACCATCTGCCGCTTTGA |
| 10 | ***SOX2*** | Pluripotency | chr3: 181,713,684 - 181,713,826 | 143 | Forward | ACGGTAGGAGCTTTGCAGGA |
| | | | | | Reverse | ACATTTTGATTGCCATGTTTATCTCGA |
| 11 | ***ZFP42*** | Pluripotency | chr4: 188,004,620 - 188,004,709 | 90 | Forward | CCCCACAACATGTTTAAACTTAGCTA |
| | | | | | Reverse | CTCAAGCTATCCTCCTGCTTCA |
| 12 | **RNA1** | RNA spike-in | - | 95 | Forward | CGCCCCGAGAATATGCTGCA |
| | | | | | Reverse | CCCTCTCTACTTTGGCGCGA |
| 13 | **RNA2** | RNA spike-in | - | 179 | Forward | CCGTAGCCCCTCCGATGATA |
| | | | | | Reverse | CGCGTACCACCATTGCATCA |
| 14 | **RNA6** | RNA spike-in | - | 145 | Forward | CCAGGGGATGATTTCGGCCA |
| | | | | | Reverse | CGCTCTGGTGCCACGATCA |
| 15 | **RNA8** | RNA spike-in | - | 188 | Forward | TCCAGCAGTTTCAGCCAGCA |
| | | | | | Reverse | CAGGCGCTGCAACTGTGTTA |

§ Based on the *In-Silico* PCR tool of UCSC Genome Browser
* Spans more than one exon

## APPENDIX K Mesoderm primers

Differentiation and control genes selected for the Wnt pathway stimulation experiment and sequences of the primers designed to target them

| # | Gene | Type | Targeted Locus on Chromosome§ | Amplicon Size§ (bp) | Primer | Nested (sequencing) primer sequence (5'>3') |
|---|---|---|---|---|---|---|
| 1 | ***CDX2*** | Paraxial/(pre-)somitic mesoderm | chr13:27,963,357-27,968,471 | 165* | Forward | ACCAGATTTTAACCTGCCTCTCA |
| | | | | | Reverse | GCCAAGTGAAAACCAGGACGAAA |
| 2 | ***EOMES*** | Lateral mesoderm | chr3:27,717,015-27,717,199 | 185 | Forward | CACCACCAAGTCCATCTGCAAAA |
| | | | | | Reverse | GCTGTCTCCTAGCAACTCCAGTA |
| 3 | ***EVX1*** | Pan-primitive streak | chr7:27,247,384-27,247,557 | 174 | Forward | CTCTCTCGGTATCTGGCGGTAAA |
| | | | | | Reverse | GAAGGCTCCCACTGGTATCTGAA |
| 4 | ***FOXA2*** | Endoderm | chr20:22,581,785-22,581,882 | 98 | Forward | CTTGCTCTCTCACTTGTCCTCGA |
| | | | | | Reverse | GTGTACTCCCGGCCCATTATGAA |
| 5 | ***GAPDH*** | Housekeeping | chr12:6,537,667-6,537,857 | 191 | Forward | CCAGAACATCATCCCTGCCTCTA |
| | | | | | Reverse | CCGACGCCTGCTTCACCA |
| 6 | ***GSC*** | Pan-primitive streak | chr14:94,768,343-94,768,542 | 200 | Forward | CCTCCCGGCTCTGTACACTA |
| | | | | | Reverse | ACCGGAGAAGAGGGAAGAGGA |
| 7 | ***HAND1*** | Lateral mesoderm | chr5:154,475,054-154,475,164 | 111 | Forward | TTGAGGTAGAAAAGGGTTGGGGA |
| | | | | | Reverse | AATAAAAGCTTTCCCTGTGTTGGA |
| 8 | ***HOXA1*** | Paraxial/(pre-)somitic mesoderm | chr7:27,094,622-27,095,261 | 175 | Forward | AGATCTTCACTTGGGTCTCGTTGA |
| | | | | | Reverse | GGGAAAGTTGGAGAGTACGGCTA |
| 9 | ***MESP1*** | Lateral mesoderm | chr15:89,749,967-89,750,152 | 186 | Forward | ATGGAGGGAGGGGCTGAGAA |
| | | | | | Reverse | CCCAAGTGACAAGGGACAACTGA |
| 10 | ***MIXL1*** | Pan-primitive streak | chr1:226,226,914-226,227,008 | 95 | Forward | CCACCTGCCTTCTGAAGTCTGA |
| | | | | | Reverse | ACAATAACAAGTGCTAAGGGTAATGGA |
| 11 | ***MSGN1*** | Paraxial /(pre-)somitic mesoderm | chr2:17,816,616-17,816,708 | 93 | Forward | CAGGGCCCTTTGAGCTGAATCA |
| | | | | | Reverse | CAGCTGGACAGGGAGAAGAAGAA |
| 12 | ***MSX1*** | Neural crest | chr4:4,863,263-4,863,396 | 134 | Forward | AGTTTCACCTCTTTGCTCCCTGA |
| | | | | | Reverse | TGCCCTCAGTTTCCCCATCTTTA |
| 13 | ***NANOG*** | Pluripotency | chr12:7,795,613-7,795,789 | 177 | Forward | CGCCCTGCCTAGAAAAGACA |
| | | | | | Reverse | CAAAGCCTCCCAATCCCAAACA |
| 14 | ***NODAL*** | Anterior primitive streak | chr10:70,432,830-70,432,955 | 126 | Forward | TTGCCCCTCTCTGTTTCTCCTTA |
| | | | | | Reverse | AAGAATGTGGGTGCCTCTGATGA |
| 15 | ***PAX3*** | Paraxial/(pre-)somitic mesoderm | chr2:222,294,218-222,295,544 | 101* | Forward | CCTCTGCCTCCTTCCTCTCCA |
| | | | | | Reverse | AAACACCGTGCCGTCAGTGA |
| 16 | ***SOX17*** | Endoderm | chr8:54,460,216-54,460,332 | 117 | Forward | AGTTGGATTGTCAAAACCCTATTTCCA |
| | | | | | Reverse | ACACCCAGGACAACATTTCTTTGA |
| 17 | ***T*** | Pan-primitive streak | chr6:166,165,727-166,166,648 | 171* | Forward | CACCGCTATGAACTGGGTCTCA |
| | | | | | Reverse | GCTCCCGTCTCCTTCAGCAAA |
| 18 | ***TBX6*** | Paraxial/(pre-)somitic mesoderm | chr16:30,086,220-30,086,394 | 175 | Forward | GTGGTTCAGTACATGGGTTTGGA |
| | | | | | Reverse | CCTACTCGGCTGCATTTCTGGA |
| 19 | ***ZIC1*** | Neural crest | chr3:147,414,157-147,414,273 | 117 | Forward | TCCACGTCGACCTAACCCAATATTA |
| | | | | | Reverse | CAGGACATGAAACAACATTTTACTGCA |
| 20 | **RNA1** | RNA spike-in | - | 95 | Forward | CGCCCCGAGAATATGCTGCA |
| | | | | | Reverse | CCCTCTCTACTTTGGCGCGA |
| 21 | **RNA2** | RNA spike-in | - | 179 | Forward | CCGTAGCCCCTCCGATGATA |
| | | | | | Reverse | CGCGTACCACCATTGCATCA |
| 22 | **RNA6** | RNA spike-in | - | 181 | Forward | AGAGCTTCGAGATAGTGGGCAAA |
| | | | | | Reverse | AAAGTCTCTCCTCTTGGCCGAAA |

§ Based on the *In-Silico* PCR tool of UCSC Genome Browser
* Spans more than one exon

## APPENDIX L *BRCA* genotyping primers

*BRCA1* and *BRCA2* loci analyzed in this study, and sequences of the primers designed to target them

| # | Gene | Mutation | Amplicon name | Amplicon Size§ (bp) | Primer | Nested (sequencing) primer sequence (5'>3') |
|---|---|---|---|---|---|---|
| 1 | *BRCA1* | **185delAG** | **Amp1** | 127 | Forward | GCGTTGAAGAAGTACAAAATGTC**A** |
| | | | | | Reverse | TTCTGTTCATTTGCATAGGAG**A** |
| 2 | *BRCA1* | **5382insC** | **Amp2** | 200 | Forward | AAGGTCCAAAGCGAGCAAG**A** |
| | | | | | Reverse | TGGTTGGGATGGAAGAGTG**A** |
| 3 | *BRCA1* | **p.Y978*** | **Amp3** | 189 | Forward | CGAAACTGGACTCATTACTCCAAAT**A** |
| | | | | | Reverse | CTCACTGTACTTGGAATGTTCTC**A** |
| 4 | *BRCA1* | **p.A1708E** | **Amp4** | 112 | Forward | GAGTTTGTGTGTGAACGGAC**A** |
| | | | | | Reverse | GTGTTAAAGGGAGGAGGGG**A** |
| 5 | *BRCA1* | **981delAT** | **Amp5** | 135 | Forward | AGATGGGCTGGAAGTAAGG**A** |
| | | | | | Reverse | AGGATTCTCTGAGCATGGC**A** |
| 6 | *BRCA2* | **c.6174delT** | **Amp6** | 96 | Forward | AGTATAGGGAAGCTTCATAAGTC**A** |
| | | | | | Reverse | TGAAGCATCTGATACCTGG**A** |
| 7 | *BRCA2* | **8765delAG** | **Amp7** | 169 | Forward | TCTGGATTATACATATTTCGCAATG**A** |
| | | | | | Reverse | TCATATTAGAAATAACAATGTGTACC**A** |
| 8 | *BRCA2* | **IVS2+1G>A** | **Amp8** | 247 | Forward | CAAGCATTGGAGGAATATCGT**A** |
| | | | | | Reverse | ACGATGATTATGTTGTTAACTGG**A** |
| 9 | *BRCA2* | **p.R2336P** | **Amp9** | 191 | Forward | TAGAGCCGATTACCTGTGT**A** |
| | | | | | Reverse | TGAACAGCACTATAAAATACTACC**A** |
| 10 | *BRCA1* | **p.P1812A** | **Amp10** | 141 | Forward | AGTGACAGTTCCAGTAGTCCT**A** |
| | | | | | Reverse | TACATGCAGGCACCTTACC**A** |

§ Based on the *In-Silico* PCR tool of UCSC Genome Browser

## APPENDIX M Patient samples analyzed with the *BRCA* genotyping assay

Anonymized patient samples analyzed using *BRCA* genotyping assay (**APPENDIX L**), and previously identified mutations by Sanger sequencing

| Sample ID | Mutation* | Genotype | Sample ID | Mutation* | Genotype | Sample ID | Mutation* | Genotype |
|---|---|---|---|---|---|---|---|---|
| **A01** | Mut1 | 185delAG | **C09** | Mut8 | IVS2+1G>A | **F05** | Mut2 | 5382insC |
| **A02** | Mut6 | c.6174delT | **C10** | --- | n | **F06** | Mut2 | insC5382 |
| **A03** | --- | n | **C11** | Mut10 | p.P1812A | **F07** | Mut7 | delAG8765 |
| **A04** | Mut9 | p.R2336P | **C12** | Mut2 | 5382insC | **F08** | Mut2 | 5382insC |
| **A05** | Mut9 | p.R2336P | **D01** | Mut9 | p.R2336P | **F09** | --- | ??? |
| **A06** | Mut8 | A<IVS2+1G | **D02** | Mut3 | p.Y978* | **F10** | Mut6 | c.6174delT |
| **A07** | --- | n | **D03** | Mut5 | delAT981 | **F11** | Mut7 | delAG8765 |
| **A08** | Mut10 | p.P1812A | **D04** | Mut6 | c.6174delT | **F12** | Mut8 | A<IVS2+1G |
| **A09** | Mut7 | delAG8765 | **D05** | Mut2 | 5382insC | **G01** | Mut4 | p.A1708E |
| **A10** | Mut9 | p.R2336P | **D06** | Mut6 | c.6174delT | **G02** | Mut8 | IVS2+1G>A |
| **A11** | --- | n | **D07** | Mut1 | 185delAG | **G03** | Mut6 | c.6174delT |
| **A12** | Mut6 | c.6174delT | **D08** | Mut1 | 185delAG | **G04** | Mut2 | 5382insC |
| **B01** | Mut6 | c.6174delT | **D09** | Mut4 | p.A1708E | **G05** | Mut5 | delAT981 |
| **B02** | Mut4 | p.A1708E | **D10** | Mut1 | 185delAG | **G06** | --- | n |
| **B03** | Mut6 | c.6174delT | **D11** | Mut7 | delAG8765 | **G07** | Mut8 | A<IVS2+1G |
| **B04** | --- | n | **D12** | Mut9 | p.R2336P | **G08** | --- | n |
| **B05** | Mut4 | p.A1708E | **E01** | Mut3 | p.Y978* | **G09** | Mut7 | delAG8765 |
| **B06** | Mut6 | c.6174delT | **E02** | Mut9 | p.R2336P | **G10** | Mut8 | A<IVS2+1G |
| **B07** | Mut1 | 185delAG | **E03** | Mut1 | 185delAG | **G11** | Mut8 | A<IVS2+1G |
| **B08** | Mut8 | A<IVS2+1G | **E04** | --- | n | **G12** | Mut10 | p.P1812A |
| **B09** | Mut6 | c.6174delT | **E05** | --- | n | **H01** | Mut1 | 185delAG |
| **B10** | Mut6 | c.6174delT | **E06** | Mut6 | c.6174delT | **H02** | --- | n |
| **B11** | Mut2 | 5382insC | **E07** | Mut8 | A<IVS2+1G | **H03** | Mut4 | p.A1708E |
| **B12** | --- | n | **E08** | Mut6 | c.6174delT | **H04** | Mut8 | A<IVS2+1G |
| **C01** | --- | n | **E09** | Mut7 | delAG8765 | **H05** | Mut2 | 5382insC |
| **C02** | Mut9 | p.R2336P | **E10** | Mut7 | delAG8765 | **H06** | Mut9 | p.R2336P |
| **C03** | Mut1 | 185delAG | **E11** | Mut10 | p.P1812A | **H07** | Mut7 | delAG8765 |
| **C04** | Mut3 | p.Y978* | **E12** | --- | n | **H08** | --- | n |
| **C05** | Mut3 | p.Y978* | **F01** | Mut7 | delAG8765 | **H09** | Mut10 | p.P1812A |
| **C06** | Mut7 | delAG8765 | **F02** | Mut2 | 5382insC | **H10** | Mut7 | delAG8765 |
| **C07** | Mut2 | 5382insC | **F03** | Mut9 | p.R2336P | **H11** | Mut3 | p.Y978* |
| **C08** | Mut7 | delAG8765 | **F04** | Mut7 | delAG8765 | **H12** | Mut7 | delAG8765 |

* Mutation ID corresponds to the Amplicon name in the primer list (**APPENDIX L**)
n: wild-type

# CURRICULUM VITAE

## Education

2014 – 2020 **Ph.D.:** Technical University of Munich, Experimental Medicine
Helmholtz Center Munich, Institute of Stem Cell Research
2008 – 2010 **M.Sc.:** Sabancı University – Istanbul, Turkey
Biological Sciences and Bioengineering Program
2003 – 2008 **B.Sc.:** Bilkent University – Ankara, Turkey
Department of Molecular Biology and Genetics
2000 – 2003 Meram Science High School – Konya, Turkey

## Research Experience

2013 Project Assistant, Vienna University of Technology, Austria
2009 ERASMUS Intern, Vienna University of Technology, Austria
2008 HHMI Summer Intern, Duke University, North Carolina, USA
2006 Summer Intern, Bilkent University, Turkey

## List of Publications

Uzbas, F.; Opperer, F.; Sönmezer, C.; Shaposhnikov, D.; Sass, S.; Krendl, C.; Angerer, P.; Theis, F. J.; Mueller, N. S.; Drukker, M. **BART-Seq: cost-effective massively parallelized targeted sequencing for genomics, transcriptomics, and single-cell analysis.** *Genome Biology* **2019**, *20* (1), 155. https://doi.org/10.1186/s13059-019-1748-6

Uzbas, F.; May, I. D.; Parisi, A. M.; Thompson, S. K.; Kaya, A.; Perkins, A. D.; Memili, E. **Molecular physiognomies and applications of adipose-derived stem cells.** *Stem Cell Rev and Rep* **2015**, *11* (2), 298–308. https://doi.org/10.1007/s12015-014-9578-0

Uzbas, F.; Sezerman, U.; Hartl, L.; Kubicek, C. P.; Seiboth, B. **A homologous production system for *Trichoderma reesei* secreted proteins in a cellulase-free background.** *Appl Microbiol Biotechnol* **2012**, *93* (4), 1601–1608. https://doi.org/10.1007/s00253-011-3674-8

Uzbaş, F. ***Trichoderma reesei* as an expression system for homologous production of individual cellulases**. Thesis, 2010. Sabancı University, Istanbul, Turkey.